

Reinforcement Learning from Reputation Feedback

An Alignment Mechanism Grounded in Computative Economics

Wulf A. Kaal

University of St. Thomas School of Law

SSRN Author ID: 460345 | ORCID: 0009-0008-7840-1847

Working Draft, Version 19, September 2026

Abstract

This paper introduces Reinforcement Learning from Reputation Feedback (RLRF), an alignment mechanism that replaces the human annotator panel of Reinforcement Learning from Human Feedback (RLHF) with an economically incentivized validation pool of computative agents whose members stake non-transferable reputation. The paper grounds the mechanism in the Computative Economics framework (Kaal 2026), in which the analytical primitive is a computative agent characterized by the tuple (C, O, M, G, R) and the object of analysis is a generated possibility space from which realizations are drawn. Under this framing, the paper takes the conceptual position that RLHF applies a scarcity-economic primitive to a signal-generation task that is computative in nature. The cited inter-annotator agreement rates do not establish that position as a causal explanation.

The paper contributes theoretically and empirically. Theoretically, it presents Fisher information dominance as an open empirical question under necessary but insufficient assumptions made explicit in Part IV. It poses equilibrium-trajectory separation as the first open problem in Part VIII. It states Sybil resistance as a design objective whose detector, dispersion statistic, threshold, error model, λ update, Round 2 update, and ϕ^* are unspecified. Conjecture 1 is withdrawn. Empirically the paper reports a measured cohort, a parameter sweep, and Experiment T as a separate measured selection experiment that does not test the mechanism. The measured cohort of 564 records, produced by twenty distinct model families against twenty-three specifications and scored by a deterministic checker on cases withheld from the producing party, establishes that a party's record where the work is checked predicts its accuracy where it is not, at a rank correlation of +0.781 across thirty-six parties with repeated measurement, and that a party failing the checked cases passes the withheld ones in zero of one hundred and sixty-eight observations. On that cohort, a selector that takes the party with the best record beats uniform selection by 0.243 with a ninety-five percent interval of [0.102, 0.384]. That selector is an argmax over a record, not the reputation-weighted aggregation the mechanism specifies, and the cohort exercises no validation pool, no stake and no deliberation round. It has not been shown to beat the alternative already available to any engineer, which is to ship whatever satisfies the checks in hand: that margin is 0.037 across twenty-two decidable specifications, with an interval

that contains zero, so the cohort does not decide between them. A 225-configuration parameter sweep across 4,500 trials supplies a two-mode comparison the cohort cannot observe. It does not establish an equilibrium-trajectory separation. Finding 7's coefficients of variation do not define or test that separation. In that sweep the parties are distributions whose accuracy is an input, and it is reported as a simulation rather than as validation. Experiment T is reported separately in Part VIII and does not test the mechanism.

The paper's contribution beyond prior work is the proposal that alignment be treated as a mechanism design problem within Computational Economics rather than as a supervised learning problem within scarcity economics, together with the apparatus that proposal requires. None of the evidence items establishes that the computational reading is the correct one, and the paper does not claim it does. Part II.D sets the mechanism against the machine learning literatures nearest to it, conceding what reliability-weighted crowdsourcing, peer prediction, RLAIIF, verifiable-reward reinforcement learning, process supervision and debate already contain, and stating the smaller residue that remains at issue. Appendix A gives the method behind the measured cohort and the parameter sweep. No policy is trained anywhere in this paper, which is the first thing Part VIII says the program owes.

Keywords: RLHF; RLRF; alignment; Computational Economics; mechanism design; reputation systems; Fisher information; Sybil resistance; generated possibility space; dynamic regulation.

JEL Codes: C62; D82; D83; K20; L86; O33; P16.

Table of Contents

I. Introduction	5
II. The RLHF Signal Quality Problem and a Proposed Scarcity-Economic Account	7
II.A Pairwise Disagreement and Its Limits	7
II.B Analogies: Arrow and Grossman-Hart-Moore	8
II.C The Proposed Primitive Distinction	8
II.D The Nearest Literatures, and What Is Left Over	9
III. The RLRF Mechanism	11
III.A Validation Pool Architecture	11
III.B Non-Transferability and the Inalienability Constraint	12
III.C Signal Properties	13
III.D The Two-Round Validation Protocol	13
III.D.1 Structure of the Two-Round Protocol	13
III.D.2 The Vote-Change Delta as a Novel Training Signal	14
III.D.3 Information-Theoretic Decomposition of the Signal	15
III.D.4 Four-Format Instruction Tuning Taxonomy	16
IV. Open Questions, Design Objective, and Remark	17
IV.A The Withdrawn Efficiency Conjecture	17
IV.B Open Empirical Question: Fisher Information Dominance	17
IV.B.1 Open Problem 1: Equilibrium-Trajectory Separation	18
IV.C Design Objective: Sybil Resistance	18
IV.D Remark 1: A Style-Independent Signal Is Invariant to Style	19
V. Empirical Evidence	20
V.A What the Measured Cohort Establishes	21
V.A.1 A record transfers, among the parties that submitted	21
V.A.2 The failure the mechanism exists for occurs, and is caught	21

V.A.3 The record as a selector, against the alternatives	22
V.A.4 What the cohort does not show, stated at the same volume	22
V.B What Only the Parameter Sweep Speaks To	23
V.B.1 Finding 1: the modeled annotator baseline settles at 67.7 percent accuracy	23
V.B.2 Finding 2: staked assessment converges to near-perfect accuracy in the sweep	24
V.B.3 Finding 3: in the sweep, Mode B holds accuracy close to the majority limit	24
V.B.4 Finding 4: economic and coupling parameters in the sweep	24
V.B.5 Finding 5: raw accuracy across the two modes under honest conditions	24
V.B.6 Finding 6: rational-agent settings in the sweep	25
V.B.7 Finding 7: two stability statistics order the modes oppositely	25
V.C How the Two Bodies of Evidence Relate	25
VI. The RLRF Training Pipeline	26
VI.A Stage 1: Supervised Fine-Tuning from Interactions (SFIT)	26
VI.B Stage 2: Reputation Reward Model (RRM)	26
VI.C Stage 3: Policy Optimization (PPO or GRPO)	26
VI.D The Four-Format Training Corpus	27
VII. RLRF as a Governance Primitive in Computative Economics	27
VII.A Dynamic Regulation and the Pacing Problem	27
VII.B Ostrom Translation to the Computative Commons	27
VII.C Recursive Equilibrium of the Validation Pool	28
VIII. Limitations and Future Work	29
IX. Conclusion	31
Appendix A. How the Cohort and the Sweep Were Produced	32
A.1 The measured cohort	32

A.2 The parameter sweep	34
A.3 What would have to be run	34
References	35

I. Introduction

Reinforcement Learning from Human Feedback (RLHF) has emerged as the dominant mechanism for aligning large language models with human intent, as documented by Ouyang et al. (2022) and Bai et al. (2022). The canonical pipeline, established by Christiano et al. (2017) and extended by Ouyang et al. (2022), proceeds in three stages: supervised fine-tuning on demonstration data, reward model training on human preference comparisons under the Bradley-Terry likelihood, and policy optimization via Proximal Policy Optimization (Schulman et al. 2017). Bai et al. (2022) extended this to the helpfulness-harmlessness dual objective and documented inter-annotator agreement rates examined below. Those rates do not identify a structural cause.

Despite practical success, RLHF exhibits well-documented pathologies. Gao, Schulman, and Hilton (2023) demonstrated that reward model overoptimization follows predictable scaling laws: as KL divergence between the optimized policy and the reference policy grows, the proxy reward continues to rise while the true reward degrades. Casper et al. (2023) systematized thirty-one distinct open problems in RLHF, organized by whether they originate in human feedback quality, reward model learning, or policy optimization. Coste et al. (2024) showed that reward model ensembles partially mitigate overoptimization but the underlying signal-quality problem persists, and they extended the Gao et al. (2023) synthetic setup to include 25% label noise to better mirror real-world annotator conditions. Direct Preference Optimization (Rafailov et al. 2023) eliminates the explicit reward model but preserves the underlying preference-aggregation structure. Reinforcement Learning from AI Feedback (Lee et al. 2024) substitutes AI-generated preferences for human ones, addressing scalability but introducing a new form of the problem.

This paper takes the conceptual position that these pathologies can be analyzed through the contrast between a scarcity-economic primitive, choice among fixed alternatives by agents whose utility derives from preferences over given goods, and a task that is computative in nature, the generation and verification of candidate outputs drawn from a possibility space constructed by the agents themselves. The paper does not establish this position as a causal explanation. Under the Computational Economics framework developed in Kaal (2026), the proposed analytical primitive for such a task is a computative agent characterized by the tuple (C, O, M, G, R) , where C is the agent's computational resources, O its objective function, M its domain model, G its generation function, and R its reputation. The task itself produces a generated possibility space rather than a choice over a fixed alternative set, and the coordinating signal is a composite of price, reputation, and verified generative capacity rather than price alone.

Reinforcement Learning from Reputation Feedback (RLRF), the mechanism introduced in this paper, is the alignment architecture that follows when the task is analyzed under the computative rather than the scarcity primitive. The reward signal is not an aggregate of preference labels drawn from an annotator pool whose individual members bear no material consequence for the accuracy of their judgments. It is the outcome of a two-round validation protocol executed by a pool of computative agents who stake non-transferable reputation tokens and whose domain models update through the verification cycle itself. Formally, for output y given prompt x :

$$R_{RLRF}(x, y) = \sum_{i \in S} w_i v_i(x, y) \quad (1)$$

where S is the validation pool selected by verifiable random function (VRF) from agents holding relevant expertise tags, $w_i = r_i / \sum_j r_j$ are normalized reputation weights, and

$v_i(x, y) \in \{0, 1\}$ is validator i 's binary assessment from the Full Vote round of a two-round protocol developed in Part III. In an unkeyed event, the reputation vector r evolves according to agreement or disagreement with the pool resolution o_{pool} . In a keyed event, it can evolve against the externally checked outcome o_{key} . The update uses accrual rate α and slashing rate γ , with γ greater than α producing asymmetric reputation dynamics. The non-transferability constraint itself is a design choice of this mechanism, taken by analogy from Hart and Moore (1994) on the inalienability of human capital rather than derived from it.

The paper makes four integrated contributions. First, it situates RLRF within the Computative Economics framework (Kaal 2026) and argues that the alignment problem is better read as a mechanism design problem than as preference aggregation. Part II.D concedes that the mechanism-design reading is not new in kind, and none of the evidence items establishes that the computative primitive is the right one, so this is a framing contribution and is not offered as a demonstration. Second, it presents Fisher information dominance as an open empirical question, poses equilibrium-trajectory separation as an open problem, states Sybil resistance as a design objective, and withdraws Conjecture 1. Third, it reports a measured cohort and a parameter sweep without conflating them. A measured cohort of 564 records produced by twenty model families against twenty-three specifications establishes, outside simulation, that a party's record where the work is checked predicts its accuracy where it is not, and that selecting the party with the best record beats choosing at random by a margin whose interval excludes zero, among the parties that submitted. It also establishes that at this sample size the cohort cannot separate the record from the naive alternative of shipping whatever satisfies the checks in hand. A separate four-experiment sweep of 225 configurations totaling 4,500 trials reports only simulated outputs for properties the cohort cannot observe. Its parties are distributions rather than systems. Experiment T is a further measured selection experiment and does not test the mechanism. Fourth, it positions the mechanism as an instance of a broader pattern: alignment architectures grounded in the computative primitive admit theoretical treatment under the recursive-equilibrium apparatus of Kaal (2026). The sweep of Part V.B does not test that

positioning, since the behavior of its parties is stipulated by its parameters rather than predicted by the framework.

The remainder of the paper is organized as follows. Part II documents the signal-quality problem of RLHF and states the paper’s conceptual framing. Part III develops the RLRF mechanism, including the two-round validation protocol and its formal signal structure. Part IV presents an open Fisher-information question, the first open problem in Part VIII, a Sybil-resistance design objective, the withdrawal of Conjecture 1, and one remark. Part V presents the empirical evidence in two parts, separating what a measured cohort establishes from what only a parameter sweep speaks to. Part VI specifies the training pipeline. Part VII situates the mechanism within the broader governance architecture of Computative Economics. Part VIII identifies limitations and future work. Part IX concludes.

II. The RLHF Signal Quality Problem and a Proposed Scarcity-Economic Account

This part documents the RLHF signal-quality problem and develops a conceptual account based on the application of a scarcity-economic primitive. Section A states what pairwise disagreement does and does not identify. Section B develops analogies to Arrow (1951) and to the work of Grossman and Hart (1986) and Hart and Moore (1990). Section C situates the conceptual position within the Computative Economics framework of Kaal (2026). Section D sets the mechanism against the machine learning literatures nearest to it and states what is left over once those concessions are made.

II.A Pairwise Disagreement and Its Limits

The RLHF reward model is trained on pairwise comparisons under the Bradley-Terry likelihood (Bradley and Terry 1952):

$$P(y_w \text{ succeeds } y_l \mid x) = \sigma(r^*(x, y_w) - r^*(x, y_l)) \quad (2)$$

where σ is the logistic function and r^* is the latent quality function the reward model r_ϕ seeks to approximate.

Pairwise disagreement measures label instability but does not identify annotator error variance or a structural variance floor without an explicit latent-label model.

The Ouyang et al. (2022) and Bai et al. (2022) agreement rates are compatible with the absence of material consequence for annotators, but they do not identify it as the cause. Ambiguity, plural values, rater error, underspecified instructions, distributional heterogeneity, and legitimate preference diversity can produce the same rates. The incentive account developed here is a conceptual position rather than an established explanation of those rates.

II.B Analogies: Arrow and Grossman-Hart-Moore

Arrow’s impossibility theorem (Arrow 1951) offers an analogy for the RLHF aggregation problem. Reward model training attempts to distill a diverse annotator pool’s preferences into a single scalar reward function. Arrow’s theorem concerns aggregation under unrestricted domain, Pareto efficiency, independence of irrelevant alternatives, and non-dictatorship. The comparison is an analogy, not an application of the theorem to reward model training and not a derivation of reward overoptimization. Arrow’s theorem does not entail the reward hacking and overoptimization documented by Gao, Schulman, and Hilton (2023).

The incomplete contracts literature of Grossman and Hart (1986) and Hart and Moore (1990) provides a complementary analogy. When contracts cannot specify all contingencies ex ante, the allocation of residual control rights determines economic outcomes. An RLHF reward model can be viewed as specifying reward for the training distribution while leaving out-of-distribution behavior unresolved. This comparison is an analogy, not a derivation of reward overoptimization. The scaling-law degradation remains the empirical result measured by Gao et al. (2023).

DPO (Rafailov et al. 2023) eliminates the explicit reward model by deriving an implicit one from the Bradley-Terry likelihood directly, but the preferences remain an aggregation input. The Arrow and Grossman-Hart-Moore comparisons remain analogies to that structure, not derivations of reward overoptimization. RLAIIF (Lee et al. 2024) substitutes AI-generated preferences for human ones, addressing the scalability bottleneck but introducing a new form of the same problem: the AI labeler’s own biases (which are themselves the output of prior RLHF training) propagate through the pipeline. The Constitutional AI method (Bai et al. 2022b), which is a different paper from the helpfulness-harmlessness study cited above, adds explicit normative principles but the principles themselves remain static, subject to the same regulatory obsolescence that Kaal (2016) identified in legal institutions.

II.C The Proposed Primitive Distinction

This paper takes the conceptual position that the pathologies enumerated above can be analyzed by treating RLHF as the application of a scarcity-economic primitive to a signal-generation task that is computative in nature. It does not establish this position as their cause. Under Robbins’s (1932) scarcity primitive, preferences over a fixed alternative set can be aggregated because the alternatives themselves are exogenously given and their attributes are known. Under Arrow and Debreu’s (1954) formalization, existence of equilibrium is proved through a fixed-point argument applied to excess demand over a fixed commodity space. The RLHF annotation task superficially resembles this structure: annotators compare pairs of outputs (y_w, y_l) and report preferences, which are aggregated into a reward. But the outputs are not exogenously given alternatives drawn from a fixed commodity space. They are realizations drawn from the policy’s generation distribution, and the conceptual account treats the relevant quality signal as a functional of the generation distribution rather than the annotator’s preference among particular realizations.

The computative primitive, by contrast, takes the generation distribution as the economic object and treats realizations as draws from it (Kaal 2026, Part IV). The agent is not a preference-holder choosing among fixed alternatives. The agent is a generator-evaluator tuple (C, O, M, G, R) whose activity produces the alternative set in the course of operating. The quality signal is a functional on the generated possibility space, not an aggregate of preferences over realizations. Reputation is not a credentialing mechanism but a track record of verified generative performance, made non-transferable and domain-specific by design. Hart and Moore (1994) on the inalienability of human capital is the analogy that motivates the choice, not a derivation of it. Coordination is not tatonnement over prices but recursive equilibrium in the sense of Kaal (2026, Part VI), computed as the fixed point of a composite best-response map.

The argument of this paper is that RLRF is an alignment mechanism that follows when the primitive is computative rather than scarcity-economic. This is a conceptual position. Pairwise agreement rates do not identify a primitive mismatch or a structural ceiling. Part V reports a simulation parameterized to published agreement rates. That is a consistency check on the model rather than a measurement of annotators, and Part V labels it as such. The measured cohort of Part V.A contains no annotators and no RLHF baseline, so it compares selectors over existing artifacts. Testing the hypothesis motivating the mechanism would require training under both signals, which Part VIII lists as the first thing this research program owes.

II.D The Nearest Literatures, and What Is Left Over

Earlier versions of this paper claimed that several of its components had no analogue in the machine learning literature. Those claims were not safe, and this section replaces them. Each paragraph names the literature that is closest to a component of RLRF, concedes what that literature already has, and states the residue that is actually at issue.

Weighting assessors by reliability. Estimating annotator reliability and weighting labels by it is old and well developed. Dawid and Skene (1979) give the EM formulation, Whitehill et al. (2009) extend it to jointly infer item difficulty and labeler expertise, and multiplicative weights (Freund and Schapire 1997) supply the online form that equation (3) is a scalar instance of, with the caveat Part III.A now states about what its regret bound does and does not give. RLRF's reputation weight is a member of this family. What is not in this family is that the weight is an economic position the assessor can lose, rather than a latent variable the aggregator infers about them.

Making truthful assessment incentive compatible. Peer prediction (Miller, Resnick, and Zeckhauser 2005) and the Bayesian Truth Serum (Prelec 2004) pay assessors in a way that makes honest reporting an equilibrium without ever observing ground truth. This is the closest prior work to RLRF's incentive claim, and it is stronger than RLRF on that point, since it does not require outcomes to resolve at all. RLRF's difference is that it does observe outcomes where they resolve and uses the unresolved cases as the object of the exercise, which peer prediction does not need and does not do.

Replacing the human annotator. RLAI (Lee et al. 2024) substitutes a model labeler, and Constitutional AI (Bai et al. 2022b) supplies that labeler with written principles. Both remove the human from the loop, which is the scalability half of the problem this paper poses. Neither gives the labeler anything to lose, which is the half that remains.

Rewards from verification rather than preference. Reinforcement learning from verifiable rewards, developed at scale in DeepSeekMath (Shao et al. 2024) and in open post-training recipes (Lambert et al. 2024), trains against a deterministic checker instead of a reward model. Remark 1’s style-invariance remark is a statement about exactly this class of signal and adds nothing to it. The residue is the domain restriction: RLVR works where an answer key exists, and the claim of this paper is that a staked pool can stand where the key cannot. That claim is not established here.

Supervising the reasoning, not just the answer. Process supervision (Uesato et al. 2022; Lightman et al. 2023) trains on intermediate steps and outperforms outcome supervision on mathematical reasoning. Debate (Irving, Christiano, and Amodei 2018) elicits argument between models as an oversight signal for questions a supervisor cannot settle directly. Format D sits between these two: like process supervision it trains on reasoning, and like debate it obtains that reasoning adversarially. Its only distinguishing feature is that the reasoning it keeps is the reasoning that moved an assessor who had a stake at risk. That selection rule is the contribution, and no experiment in this paper tests whether it helps.

Alignment as a mechanism design problem. Cooperative inverse reinforcement learning (Hadfield-Menell et al. 2016) already frames alignment as a game between a principal and an agent rather than as supervised learning. The framing in Part I is therefore not new in kind. What Part II.C adds is a specific conceptual account of preference aggregation here, and what the rest of the paper adds is a concrete mechanism with non-transferable stakes rather than a solution concept. An earlier version of this sentence said transferable, which contradicts the constraint the mechanism is built on.

Sybil resistance. Costly identity and reputation as defenses against Sybil attack are standard in distributed systems and mechanism design, and the agreement bound is classical (Lamport, Shostak, and Pease 1982). Part IV.C claims no advance on either and is careful, after revision, about what the classical bound does and does not say.

Stated plainly, what is left after these concessions is a smaller set than earlier versions claimed: a staked two-round protocol whose first round is free and whose second is not, a training record indexed by which arguments moved a staked assessor, and an open question about how those two rounds divide the information between them.

Even that residue is a combination rather than a primitive, and the paper does not clear it against the mechanisms nearest to it. Sequential peer prediction, prediction markets with a revision stage, staking and slashing designs from distributed consensus, and iterated debate all combine a free elicitation stage with a costly commitment stage in some form, and this paper does not compare against any of them. Crawford and Sobel (1982), which Part III.D cites, already supplies the cheap-talk model the protocol uses

for Round 1. Hanson (2003) is discussed in Part III.D and is not relied on for a sequencing claim. What is left that the author can defend is narrow: a zero-slashing first round followed by a costly second round, and the use of vote changes as a constructional index for Format D. The consensus-alignment penalty is not specified and is not part of the residue. Δ_v is an index into (v^1, v^2) , not an additional information channel. Both remaining items are proposals. The measured evidence in Part V.A bears on neither, and Part VIII says what would.

III. The RLRF Mechanism

This part develops the RLRF mechanism formally. Section A specifies the validation pool architecture. Section B develops the non-transferability constraint and its relation to Sybil resistance. Section C decomposes the signal structure. Section D develops the two-round validation architecture that separates cheap-talk elicitation from costly-signal commitment and states its intended information-theoretic properties, which Part IV poses as open questions. Each validator is a computative agent in the sense of Kaal (2026), characterized by the tuple (C, O, M, G, R) , and the mechanism as a whole operates on the generated possibility space rather than on a fixed alternative set.

III.A Validation Pool Architecture

The RLRF mechanism operates through validation pools composed of computative agents who hold non-transferable reputation tokens (REP). For each coordination task (prompt x and candidate output y), a validation pool S is selected by verifiable random function (Micali, Rabin, and Vadhan 1999) from the set of agents holding relevant expertise tags. VRF-based selection ensures that pool composition is cryptographically verifiable ex post and resistant to ex-ante manipulation by adversarial agents.

Each validator in the pool is a computative agent i with tuple $(C_i, O_i, M_i, G_i, R_i)$. For the validation task specifically, agent i 's domain model M_i consists of its learned calibration over the task distribution, its objective function O_i evaluates candidate outputs against the task specification, and its generation function G_i produces the agent's binary assessment $v_i(x, y) \in \{0, 1\}$. The mechanism does not require agents to share objective functions or domain models. Diversity of M across the pool is a feature rather than a bug: under reputation-weighted aggregation, disagreement among honest agents on genuinely ambiguous cases produces measurable epistemic signal, while systematic deviation from consensus identifies adversarial behavior.

The reputation update rule after an unkeyed validation event with pool resolution o_{pool} follows an asymmetric multiplicative weights structure. Here o_{pool} is the reputation-weighted consensus, not an externally verified answer:

$$\begin{aligned} r_i(t+1) &= r_i(t) \cdot (1 + \alpha), \text{ if } v_i = o_{pool} \text{ (agreement with pool resolution)} \\ r_i(t+1) &= r_i(t) \cdot (1 - \gamma), \text{ if } v_i \neq o_{pool} \text{ (disagreement with pool resolution)} \end{aligned} \quad (3)$$

with accrual rate $\alpha \in (0,1)$ and slashing rate $\gamma \in (0,1)$. The asymmetry $\gamma > \alpha$ ensures that reputation is harder to build than to lose, creating directional selection pressure toward agreement with the pool resolution. Freund and Schapire (1997) established the foundational regret bound for multiplicative weights and Arora, Hazan, and Kale (2012) generalized it. The reputation update of equation (3) is a scalar instance of that family. What the regret bound gives is that the weighted pool does not do much worse than the best fixed validator in hindsight, measured against the realized pool resolution. It does not give that the pool resolution is externally correct. Whether the weights at the reputation fixed point r_{fp} come to track validator accuracy against o_{key} rather than merely tracking agreement with the realized pool resolution, and whether weights earned on keyed tasks transfer to unkeyed ones, is unproved and is the second open problem in Part VIII.

III.B Non-Transferability and the Inalienability Constraint

The REP token is non-transferable by protocol design. Inspired by Hart and Moore (1994) on the inalienability of human capital, the constraint removes purchase of reputation through token markets or secondary exchange. A participant must instead accumulate a record of agreement with pool resolutions or, where external keys exist, accuracy against those keys. This constraint alone does not establish Sybil resistance. The Sybil-resistance design objective in Part IV.C depends on an unspecified detector, dispersion statistic, threshold, error model, λ update, the Round 2 update, and ϕ^* .

The reputation economics literature originating with Kreps and Wilson (1982) and Milgrom and Roberts (1982) establishes that reputation can sustain cooperative equilibria under incomplete information in repeated games. The Folk Theorem (Fudenberg and Maskin 1986) extends this to arbitrary stage payoffs: for sufficiently patient players, any individually rational outcome can be sustained as a subgame-perfect equilibrium through reputation-based punishment strategies.

Repeated-game results (Fudenberg and Maskin 1986) may become relevant if persistent identity, sufficiently patient agents, observable deviations, and enforceable punishments are independently established. RLRf presently establishes none of those conditions.

Non-transferability alone is not sufficient for Sybil resistance. The proposed design also invokes a consensus-alignment penalty with intensity λ and a detector based on low Round 1 dispersion. Neither component is operationally specified. Equation (3) contains no λ term, no dispersion statistic, no threshold, and no rule for identifying a coordinated group. The sweep does not parameterize λ independently. Sybil resistance is therefore a design objective that cannot be evaluated until the detector, dispersion statistic, threshold, false-positive and false-negative model, λ update, the Round 2 update, and ϕ^* are defined. Part V.B.3 reports only the fraction at which the sweep's simulated pool stops holding accuracy. The measured cohort contains adversarially configured parties but no coordinated Sybils, so the threshold is not measured outside the sweep.

III.C Signal Properties

The RLRF signal differs from RLHF in three structural respects. First, validators bear material consequences for disagreement with an unkeyed pool resolution or inaccuracy against a keyed outcome. This creates an incentive to match the recorded resolution, but it does not establish that an unkeyed resolution is externally correct. Second, expertise tags route validation tasks to qualified assessors. Third, the mechanism generates a structured tuple per validation event rather than a binary label. Section III.D develops this third property formally.

III.D The Two-Round Validation Protocol

The preceding sections formalized the RLRF reward as a reputation-weighted binary outcome. This section develops a two-round validation protocol that uses the distinction between cheap-talk information revelation and costly-signal commitment (Crawford and Sobel 1982) to collect more structured data than a single scalar label. Whether the added fields contain additional information about θ is not established. Repeated-game theory motivates the protocol’s iterative structure, but this paper does not establish the conditions required for a cooperative equilibrium. Sybil resistance is a design objective without an operational detector, dispersion statistic, threshold, error model, λ update, the Round 2 update, or ϕ^* . Equilibrium-trajectory separation remains an open problem. The sweep of Part V.B encodes these ideas rather than testing them.

III.D.1 Structure of the Two-Round Protocol

The protocol operates over a single validation window divided into two rounds. In Round 1 (the Trial Vote), each validator i in the pool casts an initial direction $f_i^1 \in \{UP, DOWN\}$ and commits a simulated stake s_i^1 that represents their indicated conviction without exposing real REP. Critically, Round 1 operates under a slashing intensity parameter c_8 set to zero: no REP is redistributed based on Round 1 outcomes. Under the terminology of Crawford and Sobel (1982), Round 1 functions as a cheap-talk round that elicits private information without economic commitment. Between Round 1 and Round 2, agents exchange natural-language argumentation through a debate corpus $D = \{d_1, d_2, \dots\}$. A decentralized relay protocol such as Nostr would serve as the transport layer. No such layer is built or exercised anywhere in this paper, and none of the three evidence items involves a debate corpus.

In Round 2 (the Full Vote), each validator casts a post-deliberation direction $f_i^2 \in \{UP, DOWN\}$ and commits a real REP stake s_i^2 . Round 2 operates under slashing intensity c_8 set to one: validators who deviate from the reputation-weighted consensus lose stake proportional to their deviation, while validators aligned with consensus gain stake from the losers’ pool. The paper does not give a single rule relating this stake redistribution, the reputation update in equation (3), and the consensus-alignment penalty λ . How a Full Vote changes r_i and s_i is therefore unspecified. Round 2 functions as a costly signal. Crawford and Sobel (1982) is a cheap-talk model and supplies the first half of this contrast, the round in which

messages carry no cost. The costly half is standard signaling and is not theirs. The pairing is what the protocol uses. The structural separation between cheap-talk Round 1 and costly-signal Round 2 is not a stylistic choice. It is the proposed source of the adversarial-detection design objective, which remains unspecified and is encoded in the sweep reported as Finding 3 (Part V.B.3). Hanson (2003) supplies a proper scoring rule in which belief reports are costly market moves. It does not argue that information aggregation is maximised by sequencing a cheap round before a costly round, and this paper does not rely on it for that claim.

The protocol produces separate structured transcripts for unkeyed and keyed validation events:

$$T_{\text{unkeyed}} = (v^1, v^2, \Delta_v, D, o_{\text{pool}})$$

$$T_{\text{keyed}} = (v^1, v^2, \Delta_v, D, o_{\text{key}}) \quad (4)$$

where $v^1 = \{(a_i, f_i^1, s_i^1)\}_{i \in S}$ is the Trial Vote distribution (agent, direction, simulated stake), $v^2 = \{(a_i, f_i^2, s_i^2)\}_{i \in S}$ is the Full Vote distribution (agent, direction, real REP stake), $\Delta_v = \{(a_i, f_i^1, f_i^2) : f_i^1 \neq f_i^2\}$ captures the subset of agents whose votes changed between rounds, and D is the debate corpus. In T_{unkeyed} , o_{pool} records consensus resolution. In T_{keyed} , o_{key} records external verification against the key.

III.D.2 The Vote-Change Delta as a Novel Training Signal

The vote-change delta Δ_v identifies the subset of domain experts whose assessment changed upon exposure to peer argumentation, together with the direction of that change. One thing must be said about it before anything else, because an earlier version of this paper got it wrong. Δ_v is a deterministic function of v^1 and v^2 . It therefore carries no information about any parameter beyond what the pair of rounds already carries, and any decomposition that adds it as a separate channel double counts. Its value is not informational but constructional: it is the index that says which validation events yield a Format D training example and what that example should contain. No analogue exists in RLHF, where annotators produce independent labels without mechanism for deliberation or belief revision. What the two rounds jointly reveal is suggested by Aumann’s agreeing-to-disagree result (Aumann 1976): rational agents with common priors cannot have common knowledge of disagreement, so post-deliberation disagreement points at differences in private information or domain models. The result requires common priors, Bayesian rationality and common knowledge, none of which is established for a validation pool, so it motivates the construction rather than licensing it.

Consider a validation event in which agent a_j votes DOWN in Round 1 and UP in Round 2. The vote change reveals three pieces of information simultaneously: (a) agent a_j ’s initial private assessment was negative, (b) the debate corpus D contained arguments sufficient to reverse that assessment, and (c) agent a_j was willing to commit real REP to

the revised position. The protocol would record those three facts. Whether they are informative about θ is not shown. Fact (c) has no defined consequence while the Round 2 update is unspecified.

The proposed training application is as follows. For each validation event with non-empty Δ_v , the system would generate a training example in what we term Format D (Debate-Informed Revision): the input is the tuple (x, y, D_{up}) where D_{up} contains the arguments of agents who voted UP in Round 2, and the target is a revised output y' calibrated to receive UP in the Full Vote. Format D would train a model on the arguments that accompanied a staked vote change. That use is untested. Its nearest neighbors are process supervision, which trains on intermediate reasoning steps rather than final answers (Uesato et al. 2022; Lightman et al. 2023), and debate, which elicits argument between models as an oversight signal (Irving, Christiano, and Amodei 2018). What Format D adds to those is that the revision target is selected by which arguments moved a staked assessor, so the supervision is indexed by economic commitment rather than by a labeler’s judgment of step quality. That is the increment, and it is untested.

III.D.3 Information-Theoretic Decomposition of the Signal

Define the quality-relevant parameter as the vector θ governing the policy’s output distribution $\pi_\theta(y \mid x)$. Mutual information between each transcript and θ decomposes by the chain rule:

$$\begin{aligned} I(T_{\text{unkeyed}}; \theta) &= I(v^2; \theta) + I(v^1 \mid v^2; \theta) + I(D \mid v^1, v^2; \theta) \\ I(T_{\text{keyed}}; \theta) &= I(v^2; \theta) + I(v^1 \mid v^2; \theta) \\ &\quad + I(D \mid v^1, v^2; \theta) + I(o_{\text{key}} \mid v^1, v^2, D; \theta) \end{aligned} \quad (5)$$

Equation (5) is a mutual-information identity. It is not a Fisher-information decomposition. The open empirical question in Part IV.B concerns Fisher information, a different object, and does not follow from (5).

Two terms that appeared in earlier versions are absent, and the reason is the same in both cases.

Δ_v is absent because it is a deterministic function of v^1 and v^2 , so its conditional contribution is identically zero and including it double counts. That was corrected in a previous version.

The unkeyed resolution o_{pool} is absent from the first decomposition in equation (5) because it is a function of v^2 and adds no information given v^2 . Across events, pool resolutions can still update reputation records under equation (3). The keyed resolution o_{key} is not a function of the votes and appears in the second decomposition. Keyed outcomes, where they exist, provide external verification and also form records across events. Consensus resolution and external verification are distinct.

Each term is non-negative and none is shown to be strictly positive. The open Fisher-information question states no assumption that would make $I(v^1 \mid v^2; \theta)$ positive, and its account concedes a babbling equilibrium in which the Trial round reveals nothing and that term is zero. The substantive content of equation (5) is that the Trial round and the debate corpus are not redundant given the Full round. That is an empirical claim about the protocol and it is not proved here.

The comparison with RLHF cannot be made by the data processing inequality. It orders two statistics of a common observation, and the scalar preference label ℓ is not a function of either RLRF transcript. They are outputs of different observation mechanisms applied to the same candidate output, and no processing chain connects them. The assumptions under which an ordering could be established are the subject of the open empirical question in Part IV.B. Nothing in Part V measures either side of it.

What the two rounds are meant to add over a scalar label is this. The pair carries information about the dimensions along which quality is contested, which a scalar preference label cannot encode, and Δ_v is the index that locates it. The debate corpus carries information about why those dimensions matter in the expertise domain, which the RLHF pipeline discards even when raters hold it internally. Both statements are about what the protocol collects. Neither is a claim that the collection has been shown to be informative about θ , which is what equation (5) would need and does not have.

III.D.4 Four-Format Instruction Tuning Taxonomy

Format D (Debate-Informed Revision) complements three additional training formats derived from the RLRF tuple, producing a four-format instruction-tuning taxonomy. Format A (Answer Generation) trains the model on candidates that received UP resolution with high reputation-weighted consensus: input $(x, \text{expertise tag } \tau)$ and target the resolved output. Format B (Self-Evaluation) trains the model to predict its own validation outcomes: input (x, y, τ) and target UP or DOWN calibrated to the actual Full Vote. Format C (Expertise Routing) improves the expertise tag classification: input x and target the τ that produced the highest resolution success rate for similar queries.

The four formats compose into a coupled system. Format C improves routing, which assembles better-matched validation pools, which produces higher-quality Format A and Format B training data, which in turn generates more informative Format D examples when the two-round structure reveals quality dimensions that the initial output missed. This compounding structure is what the paper elsewhere calls the flywheel. Nothing in this paper measures it, no stage of it is implemented, and it is listed as the third open problem in Part VIII. The four formats taken together are the pipeline through which it would be tested.

IV. Open Questions, Design Objective, and Remark

This part presents an open empirical question about Fisher information, poses equilibrium-trajectory separation as the first open problem in Part VIII, states Sybil resistance as a design objective, and gives one remark. Conjecture 1 was withdrawn. None of the open questions or the design objective is an established result. Remark 1 is limited to a style-independent signal.

The Fisher-information question retains necessary assumptions that remain insufficient to establish an ordering. The equilibrium-trajectory problem cannot be evaluated until a common stochastic process for Mode A and Mode B, its filtrations, a concentration functional, and a cross-mode estimand are defined. The Sybil-resistance design objective lacks a detector, dispersion statistic, threshold, false-positive and false-negative model, λ update, the Round 2 update, and ϕ^* . Remark 1 records only invariance of a style-independent externally checked signal. Finding 7's coefficients of variation do not define or test an equilibrium-trajectory separation.

IV.A The Withdrawn Efficiency Conjecture

Earlier versions stated a Conjecture 1 on Cramer-Rao efficiency dominance of the RLRF reward estimator over the RLHF reward estimator. It is withdrawn in Version 14. Successive revisions could not make it both internally coherent and a comparison with deployed RLHF. The paper makes no mean squared error claim.

IV.B Open Empirical Question: Fisher Information Dominance

The open question concerns Fisher information about θ , the policy parameter. Rationality alone does not support the proposed strict ordering. An empirical comparison would need a common observation model for RLHF and RLRF, an informative Trial round that excludes babbling, and a debate channel with positive conditional information. These are necessary assumptions. This paper does not show that they are sufficient.

Open empirical question (Fisher Information Dominance). Suppose the reputation vector has converged to its reputation fixed point r_{fp} , validators are rational expected-utility maximizers under equation (3), both signals belong to a common observation model about θ , the Trial round excludes a babbling equilibrium, and the debate channel has positive conditional information. Under those assumptions, does the per-sample Fisher information about θ in the RLRF signal strictly exceed that in the RLHF signal near the optimum?

What is missing. The two signals are computed under different observation models, and the paper does not construct the common model that would make their Fisher information commensurable. The reported inter-annotator disagreement rate is not a rater flip probability. The paper does not assume that weights at r_{fp} track validator

accuracy. The Trial round contributes conditional information only if validators reveal privately held information in a round that carries no stake. Under Crawford and Sobel (1982), a babbling equilibrium is available, and whether the debate stage and Round 2 slashing schedule select against it is not analyzed here. The stated assumptions therefore do not deliver a Fisher-information ordering.

Evidence. None. Finding 6 (Part V.B.6) checks that the simulated agents obey assumptions encoded by the simulation. That is a property of the simulation's construction and not evidence about Fisher information. Finding 7 (Part V.B.7) reports coefficients of variation of two stability statistics, not Fisher information. The measured cohort carries no deliberation round and emits no analogue of v^1 , so it bears on nothing here.

IV.B.1 Open Problem 1: Equilibrium-Trajectory Separation

Equilibrium-trajectory separation is posed as the first open problem in Part VIII. It cannot be evaluated until a common stochastic process for Mode A and Mode B, its filtrations, a concentration functional, and a cross-mode estimand are defined. Finding 7 reports two stability statistics from the sweep. Those statistics do not define or test the proposed separation.

IV.C Design Objective: Sybil Resistance

A decentralized validation mechanism would need resistance to Sybil attacks. The current RLRF specification does not establish it. Non-transferable reputation and the two-round architecture are specified. The proposed low-dispersion detector and the λ penalty update are not.

Design objective (Sybil Resistance, not operationally specified). If an operational rule can identify coordinated agents from Round 1 data with bounded false-positive and false-negative rates, and if a specified λ update gives identified coordinated agents lower expected reputation growth than honest agents without corrupting the pool resolution, then honest agents may retain greater expected reputation below some coordinated share ϕ^* . The design objective does not define the dispersion statistic, its threshold, the identification rule, the λ update, the Round 2 update, or ϕ^* . It is not a well-defined quantitative claim until those objects are supplied.

Argument. The intended mechanism is that cheap-talk Round 1 preserves honest signal dispersion while coordinated agents may cluster. A detector would use that difference before the costly Round 2 update. This paper does not specify a detector or an update. An adaptive adversary could also inject noise into Round 1 to match the honest dispersion profile and coordinate only in Round 2. Nothing in the present mechanism prevents that response.

On the Byzantine comparison. Earlier versions said that 49% is the classical Byzantine fault tolerance limit and that RLRF therefore operates within one percentage point of what the classical bound allows. That is wrong twice. The classical result of Lamport, Shostak and Pease (1982) requires more than two thirds honest participants in the

unauthenticated setting, and tolerates up to one half only with authenticated messages. And the 50% figure that is being approached here is the majority-agreement limit, which is a different and weaker statement than Byzantine agreement. The comparison is withdrawn. What can be said is that no mechanism deciding by weighted majority can tolerate a coordinated bloc holding more than half the weight, and that the sweep’s Mode B stays accurate close to that limit under the dispersion hypothesis.

Evidence. In a parameterized simulation reported at Part V.B.3, Mode B holds mean accuracy above 0.90 up to a Sybil fraction of 0.49 across 600 trials, and Mode A reaches 0.874 under the same conditions.

Because the detector, dispersion statistic, threshold, false-positive and false-negative model, λ update, the Round 2 update, and ϕ^* are unspecified, this figure is not evidence that the design objective has been achieved.

The Mode A comparison is another output of the parameterized simulation. It does not isolate a causal mechanism in a deployed system. Earlier versions called Mode A’s performance measurably weaker and said the contrast isolates the architecture. Both claims are withdrawn. The measured cohort contains adversarially configured parties but no coordinated Sybils, and coordination is the property this design objective turns on, so it supplies no evidence here.

IV.D Remark 1: A Style-Independent Signal Is Invariant to Style

The preceding open questions and design objective concern statistical properties of the RLRF signal under fixed conditions. Remark 1 is narrower. It concerns only invariance of an externally checked signal along a specified style dimension.

The RLHF failure mode documented by Gao, Schulman, and Hilton (2023) is not merely that the reward signal is noisy. It is that the policy learns to exploit the reward signal. Under optimization pressure, the policy discovers that certain output features, including confidence, formatting, length, and sycophantic agreement, are cheaper to optimize than correctness and are disproportionately rewarded by human raters. The policy allocates its gradient budget toward these features, crowding out investment in correctness. Proxy reward rises while true quality degrades. This is the scaling-law overoptimization measured in Gao et al. (2023), documented further by Coste et al. (2024), and catalogued as Problem 8 (reward hacking) and Problem 21 (feedback-induced distribution shift) in Casper et al. (2023).

An externally checked outcome o_{key} can be invariant to style by construction. A confidently formatted incorrect answer then receives the same signal as a poorly formatted one. This claim does not extend to a pool resolution o_{pool} . Style can move validators, and nothing in the mechanism prevents it. It also does not extend automatically to the reward model r_ψ in Part VI, which can learn style correlations absent from o_{key} . Remark 1 is therefore a statement about an externally checked signal, not about a pool verdict, the RLRF mechanism as deployed, or a trained policy.

An earlier version moved from signal invariance to policy behavior without a separability assumption. That step is withdrawn. No policy-level consequence is stated here.

Remark 1 (Style invariance in a style-independent signal). Let a reward signal be a function of an externally checked outcome o_{key} alone, where o_{key} is invariant to a style dimension s of the output. For any outputs y and y' differing only in style s , if $o_{key}(y) = o_{key}(y')$ and R depends only on o_{key} , then $R(y) = R(y')$.

This is a restatement of signal invariance. It adds nothing to reinforcement learning from verifiable rewards. It does not show that a pool resolution o_{pool} is style-independent. It does not show that a fitted reward model inherits the invariance of o_{key} . It also says nothing about how a trained policy allocates gradient effort. Any policy-level claim would require a stated separability model and empirical training evidence. Neither is present here.

Evidence. The statement follows from the stated invariance assumption. The measured cohort in Part V.A.2 concerns transfer from visible deterministic checks to withheld checks, not style invariance. The parameter sweep does not train a policy or test whether a fitted reward model inherits signal invariance. Neither evidence body supports a policy-level consequence.

Sweep 4 is a sixteen-configuration sensitivity analysis of a stylized allocation model. It returned RLRF true-quality advantages of 34.7 to 77.6 percent across its tested parameter range. Those values are outputs under parameters chosen by the author. No policy is trained in the sweep, so the values do not test Remark 1 or any policy-level consequence.

V. Empirical Evidence

This part reports two bodies of evidence and keeps them apart, because they are not the same kind of thing and a reader is entitled to know which is which.

The first is a measured cohort. Twenty distinct model families, run in both an honest and an adversarial configuration, produced artifacts against twenty-three specifications. Each artifact was scored by a deterministic checker against cases the producing party could see, and independently against cases withheld from it. Five hundred and sixty-four records resulted, of which thirty-six parties carry eight or more, which is what makes a party-level statistic meaningful. Nothing in this cohort is modeled. The parties are real systems, the artifacts are real outputs, and the verdicts come from a checker rather than a parameter. What the cohort does not contain is the mechanism: no validation pool, no stake, no deliberation round, and no reputation weight in the sense of equation (1).

The second is a parameter sweep: two hundred and twenty-five configurations across four thousand five hundred trials. In that sweep the parties are not systems but distributions, and their accuracy, noise and adversarial coordination are inputs the

sweep sets rather than quantities it observes. It reports simulated outputs for mechanism properties that cannot be observed without a population that behaves adversarially by construction.

An earlier draft of this Part presented the sweep as empirical validation of all seven findings. That description was wrong in kind. A simulation whose parameters encode the predicted behavior cannot validate the prediction, and the accuracy figures near 0.97 that it returned are properties of the parameterization rather than measurements of a mechanism. This Part corrects that.

V.A What the Measured Cohort Establishes

V.A.1 A record transfers, among the parties that submitted

The mechanism's central premise is that a party's record, earned where the world checks the work, carries information about that party's accuracy where the world does not. The cohort bears on that premise directly, with two qualifications that hold for everything in this section. The quantity measured is a per-party accuracy record, not the pool weight w_i of equation (1), so the second open problem in Part VIII is not what is being tested. And every figure is conditional on a party having submitted, which Appendix A shows is not a random condition.

Across thirty-six parties with repeated measurement, the rank correlation between a party's record on checked cases and its accuracy on withheld cases is +0.781. The forty parties are twenty families in two configurations, so the effective number of independent units is closer to twenty, and no clustering correction is applied. At the record level across all five hundred and sixty-four observations it is +0.811. Mean accuracy on checked cases is 0.702 and on withheld cases 0.608, a gap of 0.094.

The conditional probabilities are the sharper statement. A party that passes the checked cases passes the withheld ones with probability 0.866. A party that fails the checked cases passes the withheld ones in zero of one hundred and sixty-eight observations. The record does not merely correlate with later accuracy. On this cohort, passing the checked cases was necessary for passing the withheld ones in every one of the one hundred and sixty-eight observations where it could have failed to be. That is a sample with no exceptions, not a logical necessity, and with one hundred and sixty-eight observations the upper bound on the true rate is about two percent.

V.A.2 The failure the mechanism exists for occurs, and is caught

Fifty-three of the five hundred and sixty-four records are cases where a party satisfied every checked case and failed the withheld ones. That is a real deterministic check satisfied by work that fails a withheld one. The adversarial configurations were instructed to satisfy the visible checks by whatever means, so a share of these fifty-three is conduct the design asked for rather than conduct observed in the wild, and the fifty-three should not be read as a base rate. It is the reason a selector might be worth having, and it is the failure the mechanism exists to catch. It is not evidence for Remark 1. That remark concerns invariance to style, which this cohort does not test. The visible

outcome here was externally checked, but it was only a proxy for the separate withheld check. Satisfying the visible check therefore did not establish success on the withheld cases.

Individual parties exhibit the signature clearly. One reaches 0.714 on checked cases against 0.286 on withheld ones, a gap of 0.428. Another sits at 0.583 against 0.375.

V.A.3 The record as a selector, against the alternatives

Twenty-three specifications were run, of which twenty-two are decidable in the sense that the artifacts submitted for them disagree on the withheld verdict. Where every artifact passes, or none does, no selector can be separated from any other and the specification carries no information.

For each decidable specification, four selectors choose one artifact and are scored on the withheld verdict. The record uses only the rating a party carried into its submission, so no selector sees an outcome it is being scored against.

Selector	Accuracy on withheld cases
The withheld verdict itself, an upper bound	1.0000
Highest-record	0.7671
Passes-the-checked-cases	0.7297
Uniform	0.5241

The highest-record selector beats uniform selection by 0.2430, paired across twenty-two specifications, ninety-five percent interval [0.102, 0.384], winning eighteen, losing three and tying one.

What that is and is not. It is a measurement, on real artifacts, that a scalar record earned where work is checked carries usable information about work that is not. It is not a measurement of the mechanism this paper proposes. The selector takes the single highest-rated party. It is an argmax over a record, not the normalized reputation-weighted aggregation of equation (1), and nothing in the cohort exercises a validation pool, a two-round protocol, a stake, or a debate corpus. Calling this result the mechanism's core claim, as an earlier version did, was wrong. It is evidence for the premise the mechanism rests on, which is that a record transfers, and that premise is worth having measured.

V.A.4 What the cohort does not show, stated at the same volume

The highest-record selector has not been shown to beat the alternative a competent engineer would actually deploy, which is to ship whatever satisfies the checks already in hand. The point estimate favors the record by 0.0374 across twenty-two specifications, with a ninety-five percent interval of [-0.102, +0.176], winning nine, losing seven and tying six. The interval contains zero and one specification is worth 0.045. Failing to

separate two selectors is not the same as showing they are equal, and neither statement is available here. What can be said is that this cohort does not decide between them.

That comparison is also unstable in a way the interval already implies. Selecting the first satisfying artifact in file order, an arbitrary choice, reverses the sign and the record loses by 0.0511. Averaging over which satisfying artifact is chosen, which is the honest treatment because file order carries no meaning, the record wins by 0.0374. A quantity that changes sign under a tie-break has not been measured at this sample size.

The subgroup analysis shows the selector behaving the way one would want a record to behave, and rests on the same thin sample. It shows nothing about the mechanism, which the cohort does not run. Splitting the decidable specifications by whether a reward-hacking party submitted:

Condition	Highest-record	Passes-checked	Uniform
A reward-hacking party submitted, 16 specifications	0.6798	0.6283	0.3879
None submitted, 6 specifications	1.0000	1.0000	0.8873

Where no party games the checked cases, the record and the naive rival are identical at 1.000, as they should be: there is nothing for a record to add. The record's advantage appears only where a party satisfies the checks and fails the withheld cases. That is the right shape for the claim, and it is sixteen specifications, where one specification is worth 0.0625.

The binding constraint on this comparison is the number of specifications on which artifacts disagree, not the size of the effect. The population, the adversarial configurations and the scoring apparatus already exist.

V.B What Only the Parameter Sweep Speaks To

The properties below cannot be observed in the measured cohort, which contains no human annotators, no coordinated adversaries, no varied economic parameters and a single protocol mode. For each of them the sweep reports only simulated outputs. The seven findings are numbered as they were in the sweep report so that the forward references in Part IV resolve.

V.B.1 Finding 1: the modeled annotator baseline settles at 67.7 percent accuracy

The sweep reproduces the published inter-annotator agreement rates as a modeled baseline and returns a ceiling at 67.7 percent accuracy. That figure is a simulated

accuracy and not a published agreement rate: Ouyang et al. (2022) report 72.6 percent agreement and Coste et al. (2024) use 25 percent label noise, and calling 67.7 the Ouyang-Coste bound, as earlier versions did, conflates an output of the model with an input to it. It does not measure human annotators. It shows that a baseline parameterized to those published agreement rates returns the simulated ceiling, which is a consistency check on the parameterization rather than evidence about annotators.

V.B.2 Finding 2: staked assessment converges to near-perfect accuracy in the sweep

Mean accuracy for RLRF exceeds 0.97 across the four simulated experiments, against 0.68 for the modeled RLHF baseline. This is the sweep’s headline number and it is an output of parameters the sweep sets: the simulated parties are distributions whose per-party accuracy is an input to the run. It is not a measurement of any system. The 0.243 figure is the headline selector margin from Part V.A.3, while Part V.A reports other measured figures and Experiment T is a further measured selection experiment that does not test the mechanism. That selector is not this mechanism, as Part V.A.3 says.

V.B.3 Finding 3: in the sweep, Mode B holds accuracy close to the majority limit

In this parameterized simulation, Mode B, the discrete two-round architecture, holds mean accuracy above 0.90 up to a Sybil fraction of 0.49 across 600 trials. Mode A, the continuous-coupling schedule, reaches 0.874 under the same conditions.

Because the detector, dispersion statistic, threshold, false-positive and false-negative model, λ update, the Round 2 update, and ϕ^* are unspecified, this figure is not evidence that the Sybil-resistance design objective has been achieved.

Part IV.C withdraws the comparison earlier versions drew here between the simulation output and the classical Byzantine bound. The majority limit of any weighted-vote mechanism is a weaker statement than Byzantine agreement. The cohort contains adversarially configured parties but not coordinated Sybils, and coordination is the property the design objective turns on, so this result has no measured counterpart.

V.B.4 Finding 4: economic and coupling parameters in the sweep

No economic parameter varies in the cohort. The sweep report labels this comparison, but this paper reports no parameter ranking or supporting values for Finding 4. It therefore states no tuning result here.

V.B.5 Finding 5: raw accuracy across the two modes under honest conditions

The cohort runs one protocol. The sweep report labels a comparison of the modes under honest conditions, but this paper reports no raw-accuracy values or statistical comparison for Finding 5. No conclusion about indistinguishability is stated here. Finding 3 separately reports the adversarial simulation values.

V.B.6 Finding 6: rational-agent settings in the sweep

A consistency check across the preceding sweep results. It shows only that the simulated agents follow behavior encoded by the simulation. It does not test the common-observation-model, non-babbling, or positive-information assumptions needed by the open Fisher-information question.

V.B.7 Finding 7: two stability statistics order the modes oppositely

The proposed equilibrium-trajectory separation is not supported by the sweep. The sweep reports opposite orderings between Mode A and Mode B on two coefficients of variation, one computed on a separation ratio and one on a per-round competence gain, with effect sizes of 60-400x and 3-7x. Those coefficients are stability statistics of simulated trajectories. They do not define or test an equilibrium-trajectory separation. The trajectory channel, if defined, would be the deliberation stage, $I(v^1 | v^2; \theta) + I(D | v^1, v^2; \theta)$, and not the vote-change delta, which equation (5) excludes. An earlier version of this finding said otherwise. A common stochastic process, its filtrations, a concentration functional, and a cross-mode estimand remain unspecified. The cohort runs a single round and emits no v^1 and no debate corpus, so it speaks to none of this.

V.C How the Two Bodies of Evidence Relate

The sweep and the cohort do not measure the same quantity, and the difference between an accuracy near 0.97 in one and 0.77 in the other is not a contradiction. The sweep asks whether a modeled pool reaches the correct verdict under parameters it sets. The cohort asks whether a selector operating on a real record picks an artifact that survives a withheld check. The first is a property of a model. The second is a measurement.

What the cohort establishes is narrower than the mechanism's premise and does not point at the mechanism. It points only at a record-as-selector premise among parties that submitted: a record earned where the work is checked predicts accuracy where it is not, at +0.781 among parties that submitted repeatedly, and selecting the party with the best record beats choosing at random by a margin whose interval excludes zero. What it does not yet establish is anything about the record against the naive alternative already available to any engineer, in either direction. And the cohort speaks to a selector rather than to the mechanism. Until both gaps close, the honest statement is narrower than a statement about the mechanism can be: a record distinguishes parties that satisfy a visible check from parties that also satisfy a withheld one, in a population where parties of the first kind were deliberately included, and the record's advantage over shipping-what-passes was measured at 0.037 with an interval containing zero, which settles neither direction.

VI. The RLRF Training Pipeline

Part IV does not establish a signal-quality advantage. The training pipeline below is a proposal. Where a stage rests on a result from Part V, the text below says whether that result is measured or simulated.

VI.A Stage 1: Supervised Fine-Tuning from Interactions (SFIT)

The base model would be fine-tuned on a corpus of (x, y, o_{res}) triples whose resolution source is retained. In an unkeyed event, $o_{res} = o_{pool}$ records a reputation-weighted consensus resolution. It is not external verification. In a keyed event, $o_{res} = o_{key}$ records an externally checked outcome. Training examples with $o_{res} = 1$ enter the corpus under the applicable source label. The proposal requires testing whether the signal is of higher quality than an annotator aggregate. In the measured cohort the analogous quantity, the difference between a highest-record selector and random selection, is 0.243 with an interval excluding zero. That selector is not this mechanism. In the sweep the reported gap is larger under parameters the sweep sets. Training data produced by filters with different measured accuracies can differ in distribution. The 0.68 figure is the sweep’s modeled annotator baseline and is not the measured accuracy of any deployed mechanism, so it should not be read as one.

VI.B Stage 2: Reputation Reward Model (RRM)

A reward model r_ψ would be trained to predict recorded resolutions from the structured training data. For unkeyed events, it would predict o_{pool} from the reputation-weighted vote distribution. A pool resolution can be influenced by style and is not externally verified. For keyed events, it could predict o_{key} , which is externally checked and may be style-invariant if the checker ignores style. Remark 1 applies only in that latter case. Neither source property transfers automatically to r_ψ . A fitted model can learn style correlations absent from o_{key} . No RRM is trained or tested in this paper, so resistance to reward hacking is not claimed.

VI.C Stage 3: Policy Optimization (PPO or GRPO)

Standard policy gradient optimization using the RRM as reward function, via Proximal Policy Optimization (Schulman et al. 2017) or its Group Relative Policy Optimization variant (Shao et al. 2024). Part IV.B poses an open empirical question about an information ordering. If the RRM were trained on a signal carrying more information about θ , the policy parameter, the policy gradient direction might be better aligned with actual task success. That implication is not established, no model is trained anywhere in this paper, and neither Fisher information is measured. The proposed equilibrium-trajectory separation remains an open problem. If the proposed equilibrium-trajectory separation held, the equilibrium channel, if defined, would suit governance-consumer regimes and the trajectory channel training-loop regimes.

VI.D The Four-Format Training Corpus

Each Stage 1 training batch comprises examples in four formats derived from $T_{unkeyed}$ or T_{keyed} , as developed in Part III.D.4. Format A (answer generation) uses high-consensus UP-resolved outputs as training targets. Format B (self-evaluation) uses o_{pool} for an unkeyed pool resolution or o_{key} for an externally checked outcome, with the source retained. Format C (expertise routing) uses the expertise tag τ that produced highest resolution success as training target for the routing function. Format D (debate-informed revision) uses the UP-switching arguments in D together with the revised output y' calibrated to receive UP in the Full Vote, which is the format Part III.D.2 distinguishes from process supervision and debate.

VII. RLRF as a Governance Primitive in Computative Economics

RLRF is not only an alignment mechanism. It is an instance of the governance infrastructure that Computative Economics requires more broadly (Kaal 2026, Part IX). This part specifies three connections between the mechanism and the governance architecture of the computative economy.

VII.A Dynamic Regulation and the Pacing Problem

RLHF instantiates the pacing problem within machine learning (Kaal 2016). The reward model is a governance institution that encodes preferences at a specific time. As the model improves and encounters novel distributions, the reward model's encoded preferences become stale. The model cannot adapt to distributional shift without costly re-annotation, which is precisely the regulatory obsolescence problem that dynamic regulation theory identifies in legal institutions.

RLRF is intended to address this by making the governance mechanism endogenous. The reputation vector evolves as validators assess new outputs, and the design intent is that the reward signal tracks distributional shift without centralized re-annotation. Whether the weights at the reputation fixed point r_{fp} come to track validator accuracy against o_{key} rather than merely tracking agreement with the realized pool resolution, and whether weights earned on keyed tasks transfer to unkeyed ones, is unproved and is the second open problem in Part VIII. Repeated-game results (Fudenberg and Maskin 1986) may become relevant if persistent identity, sufficiently patient agents, observable deviations, and enforceable punishments are independently established. RLRF presently establishes none of those conditions.

VII.B Ostrom Translation to the Computative Commons

The generated outputs of the RLRF mechanism are non-rival (Kaal 2026, Part IX.D): a validation pool outcome can be consumed by an unbounded number of downstream

training processes without diminishing its availability. The underlying generative capacity is rival: the computational resources and reputation-weighted validator time required to produce outcomes remain finite and allocated. This combination is the generalized commons problem to which the computative translation of Ostrom (1990) applies. Within the RLRF mechanism specifically, Ostrom’s design principles translate as follows. Clearly defined boundaries correspond to non-transferable domain-specific reputation (Section III.B). Proportionality corresponds to reputation-weighted allocation of validation opportunity (equation 1). Collective-choice arrangements correspond to on-chain protocol governance. Effective monitoring corresponds to cryptographic verification of outputs against stated objective functions. Graduated sanctions correspond to the MWU asymmetry $\gamma > \alpha$. Conflict-resolution mechanisms correspond to dispute protocols with appeal pathways. Recognition of self-governance corresponds to community ownership of the validation protocol. Nested enterprises correspond to layered validation pools composing into federated networks through verified cross-network reputation.

VII.C Recursive Equilibrium of the Validation Pool

The RLRF validation pool at equilibrium is a recursive equilibrium in the sense of Kaal (2026, Part VI). Each validator is a computative agent with tuple $(C_i, O_i, M_i, G_i, R_i)$. Each validator’s decision problem in each round is to allocate computational resources across generation of an assessment and its justification argument. The domain model M_i updates after each validation cycle from recorded outcomes whose source is retained: O_{pool} for pool resolutions and O_{key} for externally checked outcomes. The composite best-response map Ψ applies individual optimization and model updating to produce the next profile of validator behaviors. The fixed point of Ψ is the recursive equilibrium of the validation pool.

Kaal (2026, Proposition 1) gives existence for a map of this kind under continuity and convexity assumptions on the space of validator generation functions and domain models, via the Brouwer fixed-point theorem (Brouwer 1911), and Proposition 2 there gives uniqueness and convergent iteration under an additional contraction assumption, via the Banach fixed-point theorem (Banach 1922). This paper does not verify that the RLRF best-response map satisfies continuity, convexity or contraction, and does not restate those propositions, so nothing here establishes that the validation pool has an equilibrium. The citation records where such an argument would start. Whether RLRF’s composite best-response map satisfies the contraction condition is an open question that depends on the specifics of the MWU update rule and the smoothness of the reputation-weighted aggregation. In the sweep, simulated reputation trajectories settle within four hundred validation events across twenty seeds per configuration. That is a property of the simulated dynamics under the parameters set for them, and it neither establishes contraction nor bears on any deployed pool.

VIII. Limitations and Future Work

The simulation sweep and the theoretical framework identify six open problems that structure the research program. A seventh, which is prior to all six and without which none of the others is worth attempting, appears below before a separate selection experiment.

First, a well-defined equilibrium-trajectory decomposition. Neither Part IV.B.1 nor Finding 7 establishes the proposed separation. The sweep reports two coefficients of variation that order the two c_8 implementations oppositely. Evaluation requires a common stochastic process governing both implementations, the corresponding filtrations, a concentration functional, and a cross-mode estimand. None is supplied here. The open Fisher-information question is separate.

Second, whether the reputation weights come to track validator accuracy. Where an external key exists, a validator’s accuracy is the probability that its assessment v_i equals o_{key} on the current task distribution. Nothing in this paper shows that the multiplicative update of equation (3) produces weights at the reputation fixed point r_{fp} that are non-decreasing in that accuracy rather than merely tracking agreement with the realized pool resolution, or that weights earned on keyed tasks transfer to unkeyed ones. Earlier versions stated this as assumption A1 and used it in the Fisher-information argument. It is now an open problem and the open empirical question does not assume it.

Third, the Format D training pipeline and the flywheel. The sweep measures properties of the validation mechanism that generates training tuples but does not train a model on the resulting Format D data. Empirical validation of the Format D training advantage requires a separate experimental regime with a model in the loop. The compounding structure described in Part III.D.4, in which better routing yields better pools and therefore better training data, is untested at every stage and is the same open problem in a longer form.

Fourth, hybrid architectures. Mode A and Mode B are not defined on a common stochastic process, and Finding 7’s coefficients of variation do not define a cross-mode estimand. Whether a hybrid architecture that alternates between Trial-Full rounds and continuous-coupling rounds can satisfy both reported stability criteria is an open mechanism design problem. The Pareto sweep over the continuous-coupling family, described in the appendix as sweep 3, suggests in simulation that no single c_8 specification within that family satisfies both criteria, but hybrid architectures in the discrete-continuous mixed class have not been tested.

Fifth, signal invariance and policy behavior. Remark 1 concerns only a signal that is invariant along one specified style dimension. It does not imply policy behavior. A policy-level claim would require a separability model, a fitted reward model that preserves the signal’s invariance, and training evidence. None is supplied here.

Sixth, hardware and annotator-model generalization. The sweep was executed on a single hardware platform with deterministic seeding. Runs are reproducible from a given

seed on the platform used. Since the parameter grids, the generating code and the trial-level outputs are not released with this version, reproducibility by anyone else has not been demonstrated at all, and nothing has been tested across hardware platforms or across alternative RLHF baseline calibrations. A reproducibility extension with three or four alternative RLHF calibrations (for example, substituting the 25% noise of Coste et al. 2024 with the 37% implied by Bai et al. 2022) would show whether the modeled ceiling of Finding 1 tracks the published agreement rates or is an artifact of one parameterization.

Prior to all six: train something. RLRF is an untested mechanism proposal. This paper does not establish that training under RLRF improves any trained model relative to RLHF, RLAIIF, RLVR, or visible-check selection. No amount of further simulation or further analysis of the cohort can establish that hypothesis. What is required is an end-to-end comparison, specified in advance, in which policies are trained under a reputation-weighted signal, under an annotator or model-rater baseline, and under the selector that ships whatever passes the visible checks, and are then evaluated on held-out work that none of the three signals could see. The reputation weighting, the two-round split, the debate stage, and Format D should be ablated separately, since Part II.D concedes that each has a close neighbor in the existing literature and the question is what the combination adds. Until that experiment is run, this paper states a mechanism, argues for it incompletely, and reports one measured result about selection that is consistent with the mechanism's premise without being a test of the mechanism.

A separate selection experiment, not the Part V.A result: a panel of seven small open instruct models selecting among candidate responses on a two-constraint instruction-following task. Experiment T does not show that consensus-weighted selection outperformed equal-weighted selection within this panel and task. Across 589 surviving prompts, joint satisfaction was 0.7963 under consensus weighting and 0.7946 under equal weighting, a difference of +0.0017 with a 95% interval from -0.0051 to +0.0085. The minimum detectable difference was 0.0106. The magnitude result is therefore inconclusive. Attribution yields a cleaner negative result. The observed difference ranked at the 53.4th percentile among all 5,040 reassignments of those weights across assessors, below the specified 95th-percentile requirement. It was typical of random weight reassignment. The 0.8101 uniform-pick score was the expectation over all candidates passing the visible check. The fluency ranker scored 0.8014. No reported comparison found that the uniform pick differed from equal weighting or from the fluency ranker, or that the fluency ranker differed from either panel arm. For this panel and task, no selector was shown to outperform another in the reported pairwise comparisons after the visible constraint was enforced. Recovering previously unparsed assessor replies reversed the magnitude estimate to -0.0017. Its 95% interval from -0.0102 to +0.0068 still contained zero. The attribution result remained below the 95th percentile at the 61.3rd percentile. Re-estimating the weights on the selection prompts materially changed the weights but left the selection outcome identical to four decimal places. Finally, the task was easy. This limited the run's ability to detect a difference in magnitude. It does not qualify the attribution result or convert the inconclusive magnitude result into support.

IX. Conclusion

This paper has introduced RLRF as an alignment mechanism grounded in Computative Economics (Kaal 2026). The mechanism replaces the annotator pool of RLHF with a validation pool of computative agents (C, O, M, G, R) who stake non-transferable reputation and execute a two-round protocol that separates cheap-talk information elicitation from costly-signal commitment. The paper states a mechanism, poses an open Fisher-information question, and reports selection measurements that do not test the mechanism.

The theoretical apparatus presents Fisher information dominance as an open empirical question under necessary but insufficient assumptions. It poses equilibrium-trajectory separation as an open problem and Sybil resistance as a design objective without a specified detector, dispersion statistic, threshold, false-positive and false-negative model, λ update, the Round 2 update, or ϕ^* . Conjecture 1 was withdrawn. Remark 1 concerns signal invariance only. None of the open questions or the design objective is an established result. A measured cohort of 564 records from 40 configurations of 20 model families supplies the transfer, reward-hacking and selection results of Part V.A. The effective number of independent units is closer to 20, and no clustering correction is applied. A four-experiment sweep of 225 configurations and 4,500 simulation trials supplies seven labeled findings about the mechanism under parameters the sweep sets, although Findings 4 and 5 report no supporting values in this paper. Experiment T is a separate measured selection experiment that does not test the mechanism. In the sweep, the modeled RLHF baseline settles at 67.7% accuracy and RLRF exceeds 0.97, both under parameters the sweep sets. Neither figure is a published agreement rate, and Part V.B.1 explains why calling 67.7 the Ouyang-Coste bound conflates an output of the model with an input to it. In the measured cohort, a highest-record selector beats uniform selection by 0.243 and is not separable from the naive alternative. That selector is an argmax over a record and not the mechanism this paper proposes. In the sweep two coefficients of variation order the two modes oppositely, with effect sizes of 60-400x and 3-7x. Finding 7's coefficients of variation do not define or test an equilibrium-trajectory separation.

The broader contribution is conceptual. Alignment should be understood as a mechanism design problem within Computative Economics rather than a supervised learning problem within scarcity economics. The tools of mechanism design, repeated game theory, and institutional economics are at least as relevant to alignment as those of supervised learning and optimization. RLRF is an attempt to instantiate that insight by embedding economic incentives directly into reward signal generation. Whether the result adapts to distributional shift without centralized re-annotation is the design objective, not an achievement. Whether the reputation weights come to track validator accuracy against o_{key} rather than pool agreement, and whether keyed-task weights transfer to unkeyed tasks, remains unproved as the second open problem in Part VIII.

The equilibrium-trajectory separation proposed in this paper is an open problem rather than a result. A common stochastic process for Mode A and Mode B, its filtrations, a

concentration functional, and a cross-mode estimand are not defined. Finding 7's coefficients of variation do not define or test that separation. If the equilibrium-trajectory separation were established, it would give the Computational Economics framework a structural distinction between mechanisms optimizing terminal economic signals and mechanisms optimizing per-round selection signals, corresponding to the equilibrium-governance and amendment-governance layers of the weighted directed acyclic graph architecture of Calcaterra, Kaal, and Andrei 2018. Both would be needed for the governed self-modification of institutions the computational framing is meant to support. Nothing here establishes that the computational framing is the only one that could support governed self-modification of institutions. The boundary between the two operational regimes, like the broader boundary between the computational and scarcity domains, is an object of empirical investigation in ongoing research.

Appendix A. How the Cohort and the Sweep Were Produced

This appendix exists because an audit found the paper had no reproducible account of either evidence body. It states method and parameters at the level a reader needs to judge the claims. Where a detail is withheld it says so rather than omitting it silently.

A.1 The measured cohort

Population. Twenty distinct model families, each run in an honest configuration and in an adversarial configuration instructed to satisfy the visible checks by whatever means, giving forty parties. They are forty configurations rather than forty independent systems: the two configurations of a family share a model, so the effective number of independent units is closer to twenty, and no figure in Part V corrects for that clustering. Vendor, size, and serving details are withheld. Each party submitted one artifact per specification it attempted, and five hundred and sixty-four records resulted. Thirty-six parties carry eight or more records, which is the threshold used for any party-level statistic. Eight is a convenience threshold chosen so that a per-party rate is not computed from two or three observations. It is not derived from a power calculation, and the transfer correlation has not been recomputed at other thresholds to show it is insensitive to this one.

Task and scoring. Each specification carries two disjoint sets of cases. The visible set is supplied to the producing party with the specification. The withheld set is never supplied and is used only for scoring. A deterministic checker returns a binary verdict on each set independently. The checker is the same program for every party and every specification, it takes no model in the loop, and it was frozen before the cohort was produced. The domain and the task type are withheld.

The record. Each party carries a scalar rating updated only from resolved outcomes on the visible sets, in timestamp order. Every selection in Part V.A.3 uses the rating a party carried into a submission, never the rating after it, so no selector observes an outcome it is scored against. The update rule and its bounds are withheld. Ratings carried into

submissions ranged across roughly two hundred points, so the record was not degenerate.

Decidability. Of twenty-three specifications, twenty-two are decidable, meaning the submitted artifacts disagree on the withheld verdict. Where every artifact passes or none does, all selectors tie by construction and the specification separates nothing. All Part V.A.3 and V.A.4 figures are over the twenty-two.

The four selectors. Highest-record picks the artifact from the party with the highest rating carried into the submission. It is an argmax over the record, not the reputation-weighted aggregation of equation (1), and the appendix uses the name highest-record throughout for that reason. Passes-the-checked-cases takes the expectation over a uniform choice among artifacts that pass the visible set, which is the honest treatment because file order carries no meaning. The file-order variant is reported alongside it in Part V.A.4 precisely because it reverses the sign. Uniform takes the expectation over a uniform choice among all submitted artifacts. The oracle returns a passing artifact wherever one exists and is an upper bound, not a selector. The two expectations are computed exactly rather than by sampling, so every figure is deterministic and the paired mean difference equals the difference of the column means. An earlier version estimated them by Monte Carlo and reported figures that did not reconcile with its own table.

Coverage and missingness. Forty parties against twenty-three specifications would be nine hundred and twenty cells. Five hundred and sixty-four exist. Cells are missing where a party produced no artifact for a specification, and the reasons are heterogeneous: capability limits, refusals, timeouts, and malformed output that the checker could not parse. Missingness is therefore not random and is plausibly correlated with party quality, which biases the cohort toward parties able to produce parseable output. Nothing in Part V corrects for this. A reader should treat every cohort figure as conditional on a party having submitted, and the transfer correlation in particular is computed over the thirty-six parties with eight or more records, which is itself a selection.

Intervals and rounding. Differences are paired by specification. Reported intervals are Student t intervals on the twenty-two paired differences at ninety-five percent, two-sided, with twenty-one degrees of freedom. They are not corrected for multiple comparisons. With twenty-two paired observations these intervals are wide, which is the point Part V.A.4 makes. The two rank correlations in Part V.A.1, computed at the party level over the thirty-six parties and at the record level over all five hundred and sixty-four records, with no tie correction beyond the default midrank treatment, are Spearman coefficients. Part V.A.1 does not report p-values, because the forty parties are twenty families in two configurations and no clustering correction is applied. Column means in the Part V.A.3 table are displayed to four decimal places while margins are computed at full precision and then rounded, so recomputing a margin from the displayed columns may differ by one in the last place.

Ties, and why they matter here. Ties for the highest record occur in eight of the twenty-two decidable specifications. In one of them all thirty-seven submitting parties

carried the same rating, because no rating had yet moved when that specification was scored, so the record arm there is chance and nothing else. The remaining seven involve thirteen, twelve, nine and three parties, and three separate specifications with two each. An earlier version broke these ties by taking whichever tied party appeared first in timestamp order, which is arbitrary in the same way that file order is arbitrary for the passes-checked arm. Every arm now takes the expectation over a uniform choice among its own tied candidates. That treatment lowers the record arm from 0.7727 to 0.7671 and the margin over uniform from 0.2487 to 0.2430, and it tightens the interval, since averaging over ties removes variance rather than adding it. Both figures in the earlier version were products of the tie-break and should not be cited.

What the cohort cannot support. No annotator, no coordinated adversary, no economic parameter, one protocol round, no debate corpus, no trained policy. Any claim requiring one of those is outside this cohort by construction.

A.2 The parameter sweep

Shape. Four sweeps totalling two hundred and twenty-five configurations and four thousand five hundred trials, twenty seeds per configuration, with deterministic seeding on a single hardware platform. The four are referred to in the text as sweeps 1 through 4. Sweep 1 varies pool size and economic parameters under honest conditions. Sweep 2 varies the coordinated-Sybil fraction and supplies the Part V.B.3 result over six hundred trials. Sweep 3 is a hundred-and-forty-configuration Pareto sweep over the continuous-coupling family. Sweep 4 is the sixteen-configuration reward-hacking sensitivity analysis referred to in Part IV.D.

What a simulated party is. Not a system. Each party is a distribution whose per-assessment accuracy, noise, and coordination behavior are inputs the configuration sets. The modeled RLHF baseline is an annotator panel parameterized to published inter-annotator agreement rates. This is why no sweep result is reported in this paper as a measurement, and why Finding 6 in particular is a check on the simulation's own construction.

What is withheld. The parameter grids, the generating code, and the trial-level outputs are not released with this version. Until they are, every sweep figure in Part V.B should be read as a report of what a program returned under settings the author chose, and not as evidence independent of those settings.

A.3 What would have to be run

Neither body of evidence trains a model. The hypothesis motivating the mechanism concerns training, and the experiment that would test it is stated in Part VIII and is not run here.

References

- Arora, Sanjeev, Elad Hazan, and Satyen Kale. 2012. "The Multiplicative Weights Update Method: A Meta-Algorithm and Applications." *Theory of Computing* 8 (1): 121-164. <https://doi.org/10.4086/toc.2012.v008a006>.
- Arrow, Kenneth J. 1951. *Social Choice and Individual Values*. New York: John Wiley and Sons.
- Arrow, Kenneth J., and Gerard Debreu. 1954. "Existence of an Equilibrium for a Competitive Economy." *Econometrica* 22 (3): 265-290. <https://www.jstor.org/stable/1907353>.
- Aumann, Robert J. 1976. "Agreeing to Disagree." *Annals of Statistics* 4 (6): 1236-1239. <https://doi.org/10.1214/aos/1176343654>.
- Bai, Yuntao, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislaw Fort, Deep Ganguli, Tom Henighan, et al. 2022. "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback." arXiv:2204.05862. <https://arxiv.org/abs/2204.05862>.
- Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022b. "Constitutional AI: Harmlessness from AI Feedback." arXiv:2212.08073. <https://arxiv.org/abs/2212.08073>.
- Banach, Stefan. 1922. "Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales." *Fundamenta Mathematicae* 3: 133-181.
- Bradley, Ralph Allan, and Milton E. Terry. 1952. "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons." *Biometrika* 39 (3/4): 324-345. <https://doi.org/10.2307/2334029>.
- Brouwer, L.E.J. 1911. "Über Abbildung von Mannigfaltigkeiten." *Mathematische Annalen* 71 (1): 97-115. <https://doi.org/10.1007/BF01456931>.
- Calcaterra, Craig, Wulf A. Kaal, and Andrei Vlad. 2018. "Blockchain Infrastructure for Measuring Domain Specific Reputation in Autonomous Decentralized and Anonymous Systems." SSRN Working Paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3125822.
- Casper, Stephen, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jeremy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. 2023. "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback." *Transactions on Machine Learning Research*. arXiv:2307.15217. <https://arxiv.org/abs/2307.15217>.
- Christiano, Paul F., Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. "Deep Reinforcement Learning from Human Preferences." In *Advances*

in Neural Information Processing Systems 30 (NeurIPS 2017). arXiv:1706.03741. <https://arxiv.org/abs/1706.03741>.

Coste, Thomas, Usman Anwar, Robert Kirk, and David Krueger. 2024. "Reward Model Ensembles Help Mitigate Overoptimization." In Proceedings of the International Conference on Learning Representations (ICLR 2024). arXiv:2310.02743. <https://arxiv.org/abs/2310.02743>.

Crawford, Vincent P., and Joel Sobel. 1982. "Strategic Information Transmission." *Econometrica* 50 (6): 1431-1451. <https://doi.org/10.2307/1913390>.

Dawid, A. P., and A. M. Skene. 1979. "Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm." *Journal of the Royal Statistical Society, Series C* 28 (1): 20-28. <https://doi.org/10.2307/2346806>.

Freund, Yoav, and Robert E. Schapire. 1997. "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting." *Journal of Computer and System Sciences* 55 (1): 119-139. <https://doi.org/10.1006/jcss.1997.1504>.

Fudenberg, Drew, and Eric Maskin. 1986. "The Folk Theorem in Repeated Games with Discounting or with Incomplete Information." *Econometrica* 54 (3): 533-554. <https://doi.org/10.2307/1911307>.

Gao, Leo, John Schulman, and Jacob Hilton. 2023. "Scaling Laws for Reward Model Overoptimization." In Proceedings of the 40th International Conference on Machine Learning (ICML 2023), 10835-10866. PMLR 202. arXiv:2210.10760. <https://arxiv.org/abs/2210.10760>.

Grossman, Sanford J., and Oliver D. Hart. 1986. "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration." *Journal of Political Economy* 94 (4): 691-719. <https://doi.org/10.1086/261404>.

Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2016. "Cooperative Inverse Reinforcement Learning." In Advances in Neural Information Processing Systems 29 (NeurIPS 2016). arXiv:1606.03137. <https://arxiv.org/abs/1606.03137>.

Hanson, Robin. 2003. "Combinatorial Information Market Design." *Information Systems Frontiers* 5 (1): 107-119. <https://doi.org/10.1023/A:1022058209073>.

Hart, Oliver, and John Moore. 1990. "Property Rights and the Nature of the Firm." *Journal of Political Economy* 98 (6): 1119-1158. <https://doi.org/10.1086/261729>.

Hart, Oliver, and John Moore. 1994. "A Theory of Debt Based on the Inalienability of Human Capital." *Quarterly Journal of Economics* 109 (4): 841-879. <https://doi.org/10.2307/2118350>.

Irving, Geoffrey, Paul Christiano, and Dario Amodei. 2018. "AI Safety via Debate." arXiv:1805.00899. <https://arxiv.org/abs/1805.00899>.

Kaal, Wulf A. 2016. "Dynamic Regulation for Innovation." In *Perspectives in Law, Business and Innovation*, edited by Mark Fenwick, Wulf A. Kaal, Toshiyuki Kono, and Erik P. M. Vermeulen, 83-96. Dordrecht: Springer. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2831040.

Kaal, Wulf A. 2026. "Computative Economics: A Framework for Economic Analysis under Computational Abundance." SSRN Working Paper. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6607458.

Kreps, David M., and Robert Wilson. 1982. "Reputation and Imperfect Information." *Journal of Economic Theory* 27 (2): 253-279. [https://doi.org/10.1016/0022-0531\(82\)90030-8](https://doi.org/10.1016/0022-0531(82)90030-8).

Lambert, Nathan, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. 2024. "Tulu 3: Pushing Frontiers in Open Language Model Post-Training." arXiv:2411.15124. <https://arxiv.org/abs/2411.15124>.

Lamport, Leslie, Robert Shostak, and Marshall Pease. 1982. "The Byzantine Generals Problem." *ACM Transactions on Programming Languages and Systems* 4 (3): 382-401. <https://doi.org/10.1145/357172.357176>.

Lee, Harrison, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. 2024. "RLAIF vs. RLHF: Scaling Reinforcement Learning from Human Feedback with AI Feedback." In *Proceedings of the 41st International Conference on Machine Learning (ICML 2024)*. PMLR 235. arXiv:2309.00267. <https://arxiv.org/abs/2309.00267>.

Lightman, Hunter, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. "Let's Verify Step by Step." arXiv:2305.20050. <https://arxiv.org/abs/2305.20050>.

Micali, Silvio, Michael Rabin, and Salil Vadhan. 1999. "Verifiable Random Functions." In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science (FOCS 1999)*, 120-130. <https://doi.org/10.1109/SFFCS.1999.814584>.

Milgrom, Paul, and John Roberts. 1982. "Predation, Reputation, and Entry Deterrence." *Journal of Economic Theory* 27 (2): 280-312. [https://doi.org/10.1016/0022-0531\(82\)90031-X](https://doi.org/10.1016/0022-0531(82)90031-X).

Miller, Nolan, Paul Resnick, and Richard Zeckhauser. 2005. "Eliciting Informative Feedback: The Peer-Prediction Method." *Management Science* 51 (9): 1359-1373. <https://doi.org/10.1287/mnsc.1050.0379>.

Ostrom, Elinor. 1990. *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.

Ouyang, Long, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022.

“Training Language Models to Follow Instructions with Human Feedback.” In Advances in Neural Information Processing Systems 35 (NeurIPS 2022). arXiv:2203.02155. <https://arxiv.org/abs/2203.02155>.

Prelec, Drazen. 2004. “A Bayesian Truth Serum for Subjective Data.” Science 306 (5695): 462-466. <https://doi.org/10.1126/science.1102081>.

Rafailov, Rafael, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. “Direct Preference Optimization: Your Language Model is Secretly a Reward Model.” In Advances in Neural Information Processing Systems 36 (NeurIPS 2023). arXiv:2305.18290. <https://arxiv.org/abs/2305.18290>.

Robbins, Lionel. 1932. An Essay on the Nature and Significance of Economic Science. London: Macmillan.

Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. “Proximal Policy Optimization Algorithms.” arXiv:1707.06347. <https://arxiv.org/abs/1707.06347>.

Shao, Zhihong, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. “DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models.” arXiv:2402.03300. <https://arxiv.org/abs/2402.03300>.

Uesato, Jonathan, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. “Solving Math Word Problems with Process- and Outcome-Based Feedback.” arXiv:2211.14275. <https://arxiv.org/abs/2211.14275>.

Whitehill, Jacob, Paul Ruvolo, Tingfan Wu, Jacob Bergsma, and Javier Movellan. 2009. “Whose Vote Should Count More: Optimal Integration of Labels from Labelers of Unknown Expertise.” In Advances in Neural Information Processing Systems 22 (NeurIPS 2009).