RESEARCH POSITION PAPER • NOVEMBER 2026

Measuring Agents Through Markets:

An Economic Framework for Evaluating AI and Robotic Systems

Raghav Rmadya

Reppo Labs

Abstract

Autonomous AI agents and robotic systems increasingly operate in settings where binary success metrics are insufficient and reference judgments are costly or contested. A study of production agents reports continued dependence on human evaluation [1], while JudgeBench documents failures of powerful language-model judges on demanding correctness comparisons [2]. Diverse model juries [3] and distributed real-robot comparisons [4] suggest that heterogeneous evidence can strengthen assessment, but leave a coordination problem unresolved: who should evaluate a trajectory, what information should be purchased, and how should evaluators be compensated?

We develop a research position for market-mediated evaluation that treats evaluation as procurement of task-relevant evidence rather than a vote on truth. Our framework routes review effort according to estimated information value, distinguishes verifiable, adjudicated, and open-textured outcomes, and sketches settlement-dependent incentives. We formulate falsifiable comparisons with single judges, fixed panels, and non-incentivized human/model juries for software agents and robotic policies. We also identify failure conditions, including correlated errors, collusion, and the absence of independent settlement. No original experimental results or incentive-compatibility guarantees are claimed; the contribution is a testable agenda for assessing whether economic coordination improves evaluation reliability per unit of cost.

Keywords: AI evaluation; autonomous agents; robotics; prediction markets; mechanism design; LLM-as-a-judge; reward modeling; distributed evaluation; reinforcement learning.

1. Introduction

Artificial intelligence is moving from generating answers to executing actions.

Language-model-based agents increasingly perform tasks involving external tools, software environments, and multi-step decision processes. Robotic foundation models are extending similar capabilities into physical environments, where systems must perceive, plan, and act under uncertainty.

These developments create a fundamental problem: how should the performance of an autonomous system be measured?

Conventional software testing generally assumes that expected behavior can be specified in advance. Although agents can sometimes be evaluated through deterministic tests, many important tasks resist such specifications.

Consider an agent responsible for resolving a customer dispute. It may correctly identify a customer's order, initiate the appropriate refund, and produce a coherent explanation. Yet the action could still be economically unreasonable, inconsistent with company policy, or inappropriate given the customer's circumstances.

A robotic system presents a similar challenge. Successfully moving an object between two locations does not necessarily imply competent behavior. The robot may apply excessive force, take an unsafe path, damage surrounding objects, or perform poorly when environmental conditions change.

In both cases, task completion is only one dimension of quality.

The evaluation problem becomes more difficult as systems acquire greater autonomy. Longer action sequences create more opportunities for error, while the number of potentially relevant environmental conditions increases.

Pan et al. [1], in *Measuring Agents in Production*, provide evidence of these practical difficulties. Their study includes 20 case studies and a survey of 86 practitioners working with deployed systems across 26 domains. The authors report that 74% depend primarily on human evaluation, while reliability remains the leading development challenge.

This finding suggests that the evaluation infrastructure surrounding autonomous systems has not advanced at the same pace as the systems being deployed.

An increasingly common response is to use one AI model to evaluate another. Such approaches offer substantial speed and cost advantages over human review, but they introduce questions about accuracy, consistency, and independence.

An alternative is to distribute evaluation across multiple judges. However, distributing judgments does not resolve how evaluators should be selected, compensated, held accountable, or aggregated.

We argue that these problems are partly questions of economic coordination.

The central question of this paper is:

> *Can market mechanisms improve the reliability and scalability of AI evaluation by coordinating diverse evaluators and rewarding informative judgments?*

We investigate this question across software agents and embodied robotic systems. Our contribution is a conceptual architecture and experimental agenda rather than a claim of demonstrated superiority.

2. The Limits of Existing Evaluation Approaches

2.1 Human evaluation remains difficult to scale

Human evaluation is valuable because people can interpret context, assess trade-offs, and apply judgment when objective measures are incomplete.

However, human review is expensive and operationally demanding.

Evaluators require domain knowledge, consistent instructions, appropriate compensation, and mechanisms for resolving disagreement. Complex tasks may require multiple specialists, and evaluation costs can rise with task duration or environmental complexity.

Pan et al. [1] show that human involvement remains central to deployed agent evaluation.

This is important because organizations increasingly require evaluation not only during model development but throughout deployment.

A continuously operating agent may generate thousands of decisions, tool interactions, and outputs. Reviewing every event manually may be impractical.

The challenge is therefore not simply replacing human judgment. It is developing systems that allocate human judgment where it contributes the greatest value.

2.2 AI judges introduce a second reliability problem

Language-model-based judges are increasingly used to assess generated outputs, rank model responses, and produce reward signals for training.

They are attractive because they can evaluate outputs rapidly and consistently apply structured instructions.

However, the validity of these judgments cannot be assumed.

Tan et al. [2] introduce JudgeBench, a benchmark designed to evaluate AI judges on difficult response comparisons involving knowledge, reasoning, mathematics, and coding.

Unlike benchmarks that primarily measure agreement with human preferences, JudgeBench emphasizes objective correctness.

Their results show that several powerful judges, including GPT-4o, perform only slightly better than random guessing on particularly challenging comparisons.

These findings reveal a potential mismatch between the apparent sophistication of an AI judge and its ability to identify errors in another model.

Haldar and Hockenmaier [5] identify a related problem: language-model judges can produce inconsistent ratings across repeated evaluations.

Together, these results suggest that evaluation systems must themselves be evaluated.

A judgment is not reliable merely because it was generated by a capable model, and the production of a numerical score does not establish the validity of the underlying assessment.

2.3 Multiple judges can improve evaluation

One response to individual judge limitations is to distribute evaluation across multiple models.

Verga et al. [3] propose the Panel of LLM Evaluators (PoLL), which combines judgments from smaller models belonging to different model families.

Across the tasks and datasets examined, their approach outperforms a single large judge while reducing evaluation costs by more than a factor of seven.

The significance extends beyond cost.

The study suggests that evaluator diversity can be useful, particularly when judgments from different models provide complementary information.

However, model diversity alone does not guarantee independent or accurate judgments. Evaluators may share training data, architectural assumptions, or systematic biases. They may also agree on an incorrect conclusion.

This distinction is fundamental.

Agreement is a property of a group. Correctness is a property of its relationship to

the task being evaluated.

A reliable evaluation system must distinguish between the two.

3. The Evaluation Problem in Robotics

The challenges of evaluating autonomous agents become more pronounced when actions occur in physical environments.

A software agent may be evaluated through recorded tool calls, outputs, and environmental states. A robotic policy must additionally contend with physical dynamics, sensor uncertainty, and potentially irreversible consequences.

A robot's performance depends on factors including object geometry, surface properties, lighting, obstacles, hardware condition, and the configuration of its surroundings.

A policy that performs reliably in one laboratory may fail in another.

Consequently, a fixed benchmark may not adequately represent the distribution of conditions encountered during deployment.

3.1 Evidence from RoboArena

Atreya et al. [4] address this challenge through RoboArena, a distributed evaluation framework for generalist robotic policies.

Rather than requiring every institution to evaluate robotic systems in identical environments, RoboArena allows participating evaluators to select tasks and environments while conducting double-blind pairwise comparisons.

The authors report more than 600 real-robot evaluation episodes involving seven generalist policies across seven academic institutions.

Their findings indicate that distributed evaluation can produce more informative policy rankings than conventional centralized evaluation approaches in the studied setting.

This provides an important precedent.

It demonstrates that the evaluation of embodied intelligence can benefit from distributed participation and the aggregation of preference information across heterogeneous physical environments.

However, RoboArena does not establish that economic incentives improve these outcomes.

It demonstrates the potential value of distributed evaluation, leaving open questions about evaluator selection, compensation, accountability, and the allocation of resources as networks scale.

3.2 Evaluating outcomes versus behavior

A robust robotics evaluation system should distinguish between whether a task succeeded and how the robot performed it.

Consider a robotic arm instructed to move a fragile object into a container.

A binary success metric might assign the same score to two policies that both complete the task.

Yet one policy may move slowly and safely, while another subjects the object to damaging

forces.

A richer evaluation process could assess:

- Task completion and final-state accuracy
- Collision avoidance and safety constraints
- Force and motion efficiency
- Robustness to environmental variation
- Recovery from unexpected events
- Generalization across tasks and environments

Some properties can be measured directly using sensors or predetermined thresholds. Others require expert interpretation.

The evaluation system must therefore support multiple kinds of evidence rather than treating every judgment as an interchangeable preference vote.

4. Evaluation as an Economic Coordination Problem

The limitations described above motivate a broader interpretation of evaluation.

Evaluation is not exclusively a prediction or classification problem. It also involves deciding whose judgment to request, what information to collect, how much to spend, and how to distinguish informative contributions from uninformative ones.

These questions have an established foundation in economics.

Avery, Resnick, and Zeckhauser [6], in *The Market for Evaluations*, examine mechanisms for encouraging the efficient provision of evaluations.

Their work identifies several challenges: evaluations may be underprovided, evaluators may participate in an inefficient order, and the useful quantity of evaluation depends on information already collected.

The same problems arise in autonomous AI.

Some agent decisions may be straightforward enough for automated verification. Others may require specialized knowledge or independent human review.

A system that applies the same evaluation procedure to every task may either overspend on simple decisions or underinvest in difficult ones.

An economically coordinated framework could allocate evaluation resources dynamically.

4.1 Three functions of evaluation markets

We distinguish three potentially useful economic functions.

- **Information allocation.** Evaluation budgets can be directed toward tasks where additional judgments are expected to reduce uncertainty or improve decisions.
- **Information elicitation.** Evaluators can report judgments and confidence, with incentives intended to encourage informative and appropriately calibrated reports.
- **Accountability.** Evaluator performance can be measured over repeated tasks, with compensation and future participation influenced by independently assessed quality.

These functions do not require a cryptocurrency or blockchain.

They represent a general mechanism-design problem. Digital assets or public settlement infrastructure may offer implementation benefits, but those choices should be evaluated separately from the underlying measurement framework.

4.2 Incentives do not automatically produce truth

A critical limitation is that markets require an appropriate resolution mechanism.

In conventional prediction markets, a forecast can often be settled against an observable future event.

AI evaluation is more complicated.

When evaluating factual correctness, an independent test or verifiable result may provide a basis for settlement.

When evaluating writing quality, decision appropriateness, or subjective user satisfaction, an unambiguous ground truth may not exist.

Rewarding evaluators solely for agreeing with a majority could encourage conformity rather than accuracy.

Peer-prediction mechanisms, including the approach developed by Miller, Resnick, and Zeckhauser [7], offer one theoretical avenue for eliciting informative reports without directly observing ground truth.

Prelec [8] likewise investigates mechanisms for eliciting truthful responses to subjective questions.

However, such mechanisms rely on assumptions about evaluator information and incentives. They can be vulnerable to coordinated reporting, shared biases, and other strategic behavior.

Accordingly, economically backed evaluation should not be equated with objectively correct evaluation. Its effectiveness must be established empirically.

5. A Framework for Market-Mediated Agent Evaluation

We propose a general architecture in which evaluation requests are routed to a heterogeneous population of evaluators, judgments are collected under specified incentive rules, and the resulting evidence is aggregated into an auditable assessment.

The framework contains five stages.

5.1 Stage I: Specify the evaluation task

Each evaluation begins with a structured request containing:

- The output, action sequence, or trajectory being evaluated
- Relevant task context and available evidence
- An explicit rubric defining desirable behavior
- Applicable safety and policy constraints
- An evaluation budget and required response time
- A description of how judgments can be verified or adjudicated

This final element is particularly important.

An evaluation task should explicitly identify whether its outcome is objectively verifiable,

expert-adjudicated, or fundamentally subjective.

Different categories may require different incentives and aggregation mechanisms.

5.2 Stage II: Select evaluators

Evaluators may include language models, specialized verification tools, domain experts, or human reviewers.

Selection should account for domain expertise, historical reliability, evaluation cost, and independence from other participants.

For example, a mathematical proof might be evaluated using a combination of symbolic verification, specialized reasoning models, and human review.

A robotic manipulation trajectory might involve sensor-based checks, computer-vision models, and robotics specialists.

The objective is not to maximize the number of judgments. It is to acquire an informative set of judgments within the available budget.

5.3 An information-allocation objective

A central hypothesis is that a coordinator can improve the value of evaluation by selecting which evidence to purchase rather than applying the same panel to every task. Let S_t denote the evaluators already assigned to task t; let U_t denote the utility of decisions supported by the eventual assessment; and let c_{it} denote the marginal cost of assigning evaluator i. One conceptual rule is

$$i^* \in \arg\max_{i \notin S_t} \frac{\mathbb{E}[U_t \mid S_t \cup \{i\}] - \mathbb{E}[U_t \mid S_t]}{c_{it}}. \tag{1}$$

This expresses an expected-value-of-information objective, not a claim that the optimal policy is known or easily computable. Estimating the marginal value of review is itself a substantive research problem: evaluator quality can depend on the task, review outcomes are selectively observed, and reviewers may respond strategically to compensation. Randomized assignment and held-out audits are therefore necessary to prevent an adaptive selector from repeatedly favoring reviewers whose apparent success merely reflects easier assignments.

5.4 A minimally specified evaluation marketplace

To clarify what would actually be traded, consider an illustrative *procurement* mechanism rather than a speculative market in tokens. A requester posts a task, evaluation budget, evidence requirements, and resolution regime. Qualified evaluators submit private offers specifying the price of a review and their evidence modality. The coordinator purchases a subset of reviews based on estimated marginal information value per unit cost, while reserving a randomized audit sample to counter selection bias. Reviewers submit time-stamped reports before seeing one another's assessments. For verifiable outcomes, compensation may combine a base fee with a bounded proper-scoring component after independent settlement. For tasks without reliable settlement, compensation should not be described as truth-incentivizing: a fixed review fee, a blinded external audit when feasible, and a transparent disagreement record are more defensible defaults. This is a proposed design sketch, not an equilibrium result, and neither truthful bidding nor collusion resis-

tance is established. The added value over ordinary panel evaluation must be measured, not assumed.

5.5 Stage III: Collect judgments and confidence

Evaluators assess the submitted artifact against the defined rubric.

A judgment may contain:

1. A categorical or numerical assessment
2. A confidence estimate, where meaningful
3. Evidence supporting the assessment
4. A description of detected failures
5. An explicit indication of uncertainty

Where possible, evaluations should initially be submitted independently to reduce the risk of herding.

Judgments may then be aggregated and, if necessary, subjected to additional review.

A market mechanism can provide incentives for evaluators to express their confidence, but financial commitment should not automatically determine the evidentiary weight of a judgment.

Otherwise, participants with more capital could exert disproportionate influence regardless of their expertise.

5.6 Stage IV: Aggregate evidence and resolve disagreement

Let the reported probability from evaluator i be denoted by p_i, with an aggregation weight w_i. A simple illustrative estimator is:

$$\hat{p} = \frac{\sum_{i=1}^{n} w_i p_i}{\sum_{i=1}^{n} w_i}. \tag{2}$$

The weights may incorporate evaluator reliability, task-specific expertise, and independence.

However, their estimation must avoid circularity. An evaluator should not be considered reliable merely because their judgments previously agreed with the same group used to establish consensus.

For objectively verifiable tasks, some judgments can be assessed against independent outcomes.

When an observable binary outcome is available, proper scoring rules provide a basis for encouraging calibrated probabilistic reports [9]. For a binary outcome $y \in \{0, 1\}$ and reported probability $p \in [0, 1]$, a normalized Brier-style payment component can take the form

$$s(p, y) = 1 - (p - y)^2. \tag{3}$$

Properness concerns incentives for truthful probabilistic reporting under its stated assumptions; it does not guarantee immunity to collusion, leaked outcomes, common-mode errors, or heterogeneous risk preferences. Payment design and independent settlement should be evaluated together.

For subjective tasks, the framework must retain disagreement and uncertainty rather than automatically compressing them into a single authoritative score.

The output should therefore include not only an aggregate assessment but also information about evaluator disagreement, supporting evidence, and the resolution procedure.

5.7 Stage V: Feed results back into learning systems

The resulting evaluation signals can serve multiple purposes.

At inference time, an agent might use feedback to reconsider a proposed action, request additional information, or escalate a decision to a human operator.

During development, evaluation outcomes can identify recurring failure modes and support comparisons between model or policy versions.

For post-training, sufficiently validated judgments can be converted into preference pairs, reward signals, or curated examples.

The suitability of these signals for training must be tested independently.

A reward signal optimized for agreement among evaluators may not correspond to genuine improvements in task performance.

The proposed loop is therefore:

$$\textbf{Agent action} \rightarrow \textbf{Evaluation request} \rightarrow \textbf{Independent judgments} \rightarrow \textbf{Economic coordination} \rightarrow \textbf{Adjudicated feedback} \rightarrow \textbf{Model or policy improvement}$$

A central research question is whether this loop provides better information per unit of evaluation cost than conventional alternatives.

6. Applications Across Digital and Physical Agents

6.1 Software agents

Consider an autonomous coding agent asked to modify a production application.

An evaluation system could combine executable tests, security analysis, code-quality review, and an assessment of whether the change satisfies the user's original requirements.

Automated tests may provide strong evidence of functional correctness, but they may miss architectural weaknesses, security concerns, or unintended consequences.

Economically coordinated evaluation could direct specialized reviewers toward changes that are uncertain, high impact, or poorly covered by existing tests.

A successful outcome would be a measurable reduction in undetected failures relative to the cost of evaluation.

6.2 Decision-making agents

An AI agent responsible for recommending commercial or operational decisions may have several plausible actions available.

Evaluating such an agent requires more than checking whether its answer is grammatically correct.

Reviewers may need to assess whether the agent considered relevant evidence, applied appropriate constraints, acknowledged uncertainty, and selected a defensible course of action.

In cases where multiple decisions are reasonable, the evaluation system should not manufacture certainty.

Market mechanisms may be useful for eliciting differing assessments, but a credible system must preserve justified disagreement.

6.3 Robotic systems

In robotics, evaluation requests could include synchronized video, sensor telemetry, action sequences, task descriptions, and environmental conditions.

Automated systems could assess measurable constraints such as collision events, force thresholds, and task completion.

Specialized reviewers could assess more ambiguous properties, including manipulation quality, robustness, and the appropriateness of recovery behavior.

The framework could also allocate evaluation resources toward environments where policy performance is uncertain.

However, market-generated feedback should not override deterministic safety constraints or become the sole basis for safety-critical deployment decisions.

Robotic policies require independent physical testing and appropriate operational oversight.

6.4 Three regimes for resolving evaluation claims

The epistemic status of the task should determine how judgments are resolved. We distinguish the following regimes.

- **Verified outcomes.** Independent tests or delayed observations provide a reference outcome, such as a hidden correctness test or a measured robotic force-limit violation. Proper scoring may be feasible, subject to information-leakage controls.
- **Adjudicated outcomes.** A costly but meaningful reference decision can be obtained through blinded, domain-qualified review, for example when assessing compliance with an underspecified instruction or the quality of a robot's recovery behavior. Disagreement among adjudicators should be reported rather than erased.
- **Open-textured judgments.** Legitimate preferences or value trade-offs prevent a stable single label. The goal is to elicit the distribution and reasons for judgments, not to certify a majority opinion as objective truth.

This taxonomy is not a proposal to place all decisions into a market. Where an inexpensive deterministic check is decisive, an economic coordinator offers little additional benefit; where no credible settlement mechanism exists, financial stakes may reward conformity rather than truth.

7. Research Hypotheses

The proposed framework yields several testable hypotheses.

H1: Evaluator diversity improves judgment quality. Heterogeneous evaluator groups will outperform individual judges on at least some tasks when assessed against independently established reference judgments or outcomes.

H2: Economic incentives improve calibration under suitable resolution conditions. Properly designed incentive mechanisms will improve the calibration or informativeness of evaluator reports relative to equivalent non-incentivized evaluation procedures.

H3: Adaptive allocation improves evaluation efficiency. Directing additional evaluations toward uncertain or high-impact cases will achieve a better reliability-cost trade-off than assigning a fixed number of judges to every task.

H4: Market-mediated feedback improves downstream agent performance. When used for agent selection, revision, or training, appropriately validated market-mediated feedback will improve held-out task performance relative to matched conventional evaluation pipelines.

These are hypotheses, not established findings.

In particular, the proposed framework may fail to outperform simpler aggregation methods once evaluation cost, latency, manipulation risk, and implementation complexity are considered.

8. Experimental Design

A credible empirical assessment should isolate the contribution of economic coordination from the benefits of evaluator diversity.

We propose comparing five evaluation conditions:

Table 1: Proposed evaluation conditions.

Condition	Evaluation approach
A	Single language-model judge
B	Conventional panel of diverse model judges
C	Mixed human and model panel without economic incentives
D	Mixed panel with incentive-based elicitation
E	Full market-mediated system with adaptive allocation and adjudication

All conditions should operate on comparable task sets, with appropriate controls for evaluation budgets and access to evidence.

The experiments should cover three task categories.

Objectively verifiable tasks: coding, mathematical reasoning, factual verification, and agent workflows with independently observable outcomes.

Subjective tasks: ambiguous recommendations, writing quality, policy interpretation, and context-dependent decisions evaluated by independent expert panels.

Robotic tasks: recorded manipulations, navigation attempts, and real-world policy comparisons with synchronized telemetry and verified outcomes.

Evaluation metrics should include agreement with independently established reference outcomes, repeated-rating consistency, probability calibration, robustness to manipulated prompts or inputs, evaluation cost, end-to-end latency, and downstream performance.

For robotics, additional metrics should include task completion, safety violations, performance variation across environments, and failure-detection sensitivity.

Crucially, the reference judgments used to evaluate the system must not be generated by the same mechanism being tested.

Studies should also include deliberate attacks, such as coordinated evaluators, misleading evidence, artificial consensus, and attempts to manipulate evaluator rewards.

The strongest evidence would come from demonstrating that economic coordination contributes measurable improvements beyond those attributable to judge diversity or increased expenditure. In particular, the same potential reviewer population should be used in conditions C–E, tasks should be stratified by difficulty and domain, and incentive/routing assignments should be randomized where practical. Independent auditors should assess a prespecified held-out sample rather than using the market's own consensus as the reference label. A negative result, in which additional payments and routing fail to improve accuracy, calibration, or cost-adjusted failure detection, would count against the central hypothesis.

9. Limitations and Open Questions

Several limitations constrain the proposed approach.

First, **economic incentives do not guarantee truthful judgments**. Participants may coordinate, misreport, or optimize for the incentive mechanism rather than the underlying task.

Second, ground truth is often incomplete. Some tasks admit objective verification, while others involve irreducible uncertainty or legitimate disagreement.

Third, evaluator independence is difficult to establish. Two apparently independent models may share training data, failure modes, or incentives.

Fourth, market participation introduces selection effects. Financial requirements may discourage qualified evaluators or disproportionately favor participants with greater capital.

Fifth, the cost of coordination may exceed its benefits. In many settings, a simple automated test or conventional model ensemble may be more efficient.

Sixth, evaluation mechanisms can create feedback loops. If systems are repeatedly trained to optimize market-generated scores, they may learn to exploit weaknesses in the evaluation procedure.

Finally, physical systems introduce safety constraints that cannot be resolved through economic consensus alone.

These limitations suggest that economically coordinated evaluation should complement, rather than categorically replace, established verification and evaluation methods.

10. Discussion

The research considered in this paper points toward a common challenge.

Production agents require evaluation at a scale difficult to sustain through human review alone. Automated judges can reduce that burden, but their reliability remains uneven. Diverse panels can improve some evaluations, while distributed robotics benchmarks demonstrate the value of collecting evidence across multiple environments and institutions.

Together, these findings motivate a broader question about the organization of evalua-

tion.

Should evaluating autonomous systems remain a collection of isolated testing procedures, or can it become a persistent infrastructure layer that coordinates expertise, evidence, and incentives?

A market-mediated architecture offers one possible answer.

Its most promising contribution may not be the replacement of any individual judge. Instead, it may lie in deciding which evaluations are worth performing, identifying the most informative evaluators, measuring uncertainty, and maintaining accountability over time.

This perspective also shifts the role of evaluation in autonomous learning.

Rather than treating evaluation as a final benchmark applied after development, it becomes an ongoing process through which agents receive feedback about their actions and their consequences.

The economic mechanism is useful only insofar as it improves the quality and availability of that feedback.

11. Conclusion

Autonomous AI systems are becoming capable of operating in increasingly complex digital and physical environments. Yet the methods used to evaluate these systems remain constrained by human review costs, unreliable automated judgments, and incomplete coverage of real-world conditions.

Existing research establishes important components of a possible solution.

Studies of production agents reveal persistent evaluation challenges. Research on language-model judges identifies weaknesses in automated assessment. Diverse evaluator panels demonstrate the potential value of aggregating judgments. Distributed robotic evaluation shows that performance measurement can extend across heterogeneous physical environments.

Market mechanisms introduce an additional dimension: the coordinated allocation of evaluation resources and incentives.

We have proposed a conceptual framework for integrating these ideas and identified empirical tests needed to assess its usefulness.

The central proposition is not that markets inherently produce truth, but that carefully designed economic mechanisms may improve how evaluation information is collected, verified, and used.

The next challenge for autonomous AI may therefore be as much about developing better ways to measure intelligence as developing more capable intelligence itself.

> *If autonomous systems are to improve continuously, the systems evaluating them must become more reliable, scalable, and accountable.*

AI assistance and author responsibility

OpenAI ChatGPT provided substantive assistance in organizing the literature, developing and refining the proposed conceptual framework and research questions, drafting and editing prose, and preparing the LaTeX source. The named author is responsible for verifying

the references and claims, reviewing the scientific argument, and approving any submitted version. No original experiments or measured results are claimed in this manuscript.

References

[1] M. Pan, N. Arabzadeh, R. Cogo, et al. “Measuring Agents in Production”. In: *Proceedings of the 43rd International Conference on Machine Learning*. Vol. 306. Proceedings of Machine Learning Research. PMLR, 2026, pp. 95649–95695. URL: https://proceedings.mlr.press/v306/pan26a.html.

[2] S. Tan, S. Zhuang, K. Montgomery, et al. “JudgeBench: A Benchmark for Evaluating LLM-based Judges”. In: *International Conference on Learning Representations*. 2025. URL: https://arxiv.org/abs/2410.12784.

[3] P. Verga, S. Hofstatter, S. Althammer, et al. *Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models*. 2024. arXiv: 2404.18796. URL: https://arxiv.org/abs/2404.18796.

[4] P. Atreya, K. Pertsch, T. Lee, et al. “RoboArena: Distributed Real-World Evaluation of Generalist Robot Policies”. In: *Proceedings of the 9th Conference on Robot Learning*. Vol. 305. Proceedings of Machine Learning Research. PMLR, 2025, pp. 336–364. URL: https://proceedings.mlr.press/v305/atreya25a.html.

[5] R. Haldar and J. Hockenmaier. “Rating Roulette: Self-Inconsistency in LLM-As-A-Judge Frameworks”. In: *Findings of the Association for Computational Linguistics: EMNLP 2025*. Association for Computational Linguistics, 2025, pp. 24986–25004. DOI: 10.18653/v1/2025.findings-emnlp.1361. URL: https://aclanthology.org/2025.findings-emnlp.1361/.

[6] C. Avery, P. Resnick, and R. Zeckhauser. “The Market for Evaluations”. In: *American Economic Review* 89.3 (1999), pp. 564–584. DOI: 10.1257/aer.89.3.564.

[7] N. Miller, P. Resnick, and R. Zeckhauser. “Eliciting Informative Feedback: The Peer-Prediction Method”. In: *Management Science* 51.9 (2005), pp. 1359–1373. DOI: 10.1287/mnsc.1050.0379.

[8] D. Prelec. “A Bayesian Truth Serum for Subjective Data”. In: *Science* 306.5695 (2004), pp. 462–466. DOI: 10.1126/science.1102081.

[9] T. Gneiting and A. E. Raftery. “Strictly Proper Scoring Rules, Prediction, and Estimation”. In: *Journal of the American Statistical Association* 102.477 (2007), pp. 359–378. DOI: 10.1198/016214506000001437.

Manuscript status: Research position paper. No original experiments, benchmark results, or formal mechanism-design guarantees are claimed. The proposed architecture and hypotheses require empirical validation.