Public Performance Signals and Task Allocation in Decentralized LLM Agent Markets

Stage 3: Free Agent Economy

March 2026

Abstract

We construct a decentralized market of autonomous LLM agents with heterogeneous capabilities (specialist tools) and study whether public performance signals improve task allocation quality. Using a Vickrey (second-price sealed-bid) auction mechanism with reputation-only penalties (no enforced token slashing), we compare four information regimes: no reputation, scalar reputation (aggregate accuracy only), private reputation (personal delegation experience), and granular public reputation (per-category, per-difficulty performance vectors). We find that any public performance signal increases task accuracy by +5–6 percentage points (88.0% $\rightarrow$ 93–94%, $n = 300$, 3 seeds) relative to no signal. The effect is concentrated in hard tasks requiring specialist tools (L3/L4) and is associated with three behavioral changes: (1) improved routing of tasks to specialists, (2) emergence of delegation behavior, and (3) reduced adverse selection by low-skill agents. Without reputation, generalist agents underbid specialists on tool-intensive tasks and fail at high rates — a pattern consistent with classical adverse selection. With granular public reputation, the market approaches the oracle upper bound (97.5%) within 4 percentage points in this setting.

Contents

1 Introduction

Can economic mechanisms improve the reliability of LLM-based agent systems? Prior work in this project established that incentive framing increases single-agent accuracy (Stage 1) but that the effect is not specifically economic — any high-stakes prompt framing achieves similar results. Persistent reputation failed to improve accuracy in homogeneous single-agent settings (Stage 2), because reputation is a screening mechanism that requires heterogeneity to screen for.

Stage 3 introduces genuine heterogeneity through *tool access*: each agent has access to a different set of computational tools, creating real comparative advantage. We construct a decentralized market where agents bid for contracts via Vickrey auctions, can delegate to other agents, and accumulate performance records. The central question is:

> *Do public, granular performance signals improve task allocation quality in a decentralized market of heterogeneous LLM agents?*

We distinguish *task accuracy* (fraction of correctly solved tasks) from *allocative efficiency* (whether tasks are routed to the agent best equipped to solve them). Both are measurable against an oracle baseline that perfectly matches tasks to specialists. We do not claim to measure total welfare or principal surplus, which would require additional assumptions about the cost of reputation infrastructure and the value function over accuracy gains.

1.1 Contribution

This paper makes three contributions:

1. We demonstrate that **tool-based specialization** creates genuine comparative advantage among LLM agents — not prompt-based, but deterministic capability gaps.
2. We show that **granular public performance signals** (per-category, per-difficulty accuracy records) increase task accuracy by +5.3pp in this market setting, with the effect concentrated where specialist tools are essential. We ablate the information treatment across four regimes (none, scalar, private, granular public) to isolate which component of the signal drives the effect.
3. We document a pattern consistent with **adverse selection** in LLM agent markets: without performance signals, low-skill agents underbid specialists on tool-intensive tasks and fail at high rates.

2 Market Architecture

2.1 Agents

Five autonomous agents operate in the market, each receiving the same minimal instruction:

```
You are an autonomous economic agent.
Maximize E[∑ payoffs] = ∑ p(correct_i) × payment_i.
Wrong answer = no payment + reputation damage.
```

No strategy hints, no role-specific instructions. Agents differ only in their available tools:

Table 1: Agent tool endowments

Agent	Tools	Dominant category	Type
Worker-M	`pow(b,e,mod)`, `gcd()`	modular, arithmetic	Specialist
Worker-X	`numpy.linalg.det()`, `solve()`	matrix, algebra	Specialist
Worker-C	`comb()`, `perm()`, `factorial()`	counting	Specialist
Worker-G1	none	—	Generalist
Worker-G2	none	—	Generalist

Tool access creates *real* comparative advantage: a 4×4 determinant is trivial with `numpy.linalg.det()` but error-prone via mental arithmetic. This mirrors real-world specialization where tools represent capital investment.

2.2 Tasks

Tasks are procedurally generated mathematical problems across 4 difficulty levels and 5 categories (arithmetic, modular, matrix, algebra, counting). Ground truth is computed by numpy — no LLM-as-judge. Each run uses 25 tasks per level (100 total), with deterministic seeding for reproducibility.

Table 2: Task pricing by difficulty level

Level	Payment (if correct)	Example task	Tool needed
L1 (easy)	10 AGT	Digit sum of 47283	No
L2 (medium)	25 AGT	$2^{15} \bmod 11$	Helps
L3 (hard)	50 AGT	$5^{928} \bmod 97$	Essential
L4 (very hard)	100 AGT	$\det(4 \times 4)$ matrix	Essential

2.3 Auction Mechanism

Contracts are allocated via **Vickrey (second-price sealed-bid) auctions**:

1. Contracts are posted in batches of 6 per round.
2. All 5 agents submit sealed bids simultaneously (1 API call per agent).
3. For each contract: lowest bid wins, winner is paid the second-lowest bid.
4. Bidding is costless — losing an auction costs nothing.

In a Vickrey auction, truthful bidding is the weakly dominant strategy. Each agent's optimal bid equals their breakeven price:

$$b^* = \frac{p(\text{wrong}) \times \text{reputation\_cost}}{p(\text{correct})} \tag{1}$$

For a specialist with the right tool, $p(\text{correct}) \approx 1.0$, so $b^* \approx 0$. For a generalist on a hard task, $p(\text{correct})$ is low and b^* is high or the agent should pass.

2.4 Settlement: Reputation as Penalty

We use a **reputation-only penalty model** — no enforced token slashing:

- **Correct answer:** Agent receives payment.

- **Wrong answer:** No payment. Failure is recorded in the agent's public reputation vector.

The "penalty" is endogenous: a bad reputation reduces future contract wins because other agents (and the auction dynamics) route work away from unreliable agents. This is more realistic than enforced penalties — in real markets, contractors are not fined for bad work, they stop getting hired.

2.5 Performance Signals

We use the term "reputation" loosely in this paper; what agents observe is more precisely a *granular public performance signal* — a structured record of past accuracy by category and difficulty level:

```
Agent: Worker-M
Tools: compute_modular_power, compute_gcd
Balance: 742 AGT
Total: 28 tasks, 27 correct (96%)
Performance:
  modular_L1: 100% (4x)
  modular_L2: 100% (6x)
  modular_L3: 100% (8x)
  modular_L4: 100% (5x)
  arithmetic_L1: 80% (5x)
```

This is richer than a scalar reputation score (which would show only "96%, 28 tasks") and functions as public performance telemetry rather than a simple trust signal. We deliberately test both levels of granularity to measure whether the category-level breakdown is load-bearing.

In the *public_rep* condition, all agents see all vectors. In *scalar_rep*, they see only the aggregate line. In *private_rep*, agents build personal experience records from their own delegation outcomes. In *no_rep*, no track records are visible.

2.6 Delegation (Aftermarket)

After winning an auction, an agent can **delegate** the task to another agent for a fee, keeping the margin. The delegation decision uses the same utility function:

$$U(\text{accept}) = p(\text{correct}) \times \text{payment} \tag{2}$$

$$U(\text{delegate}) = \text{payment} - \text{fee} \quad (\text{if delegate succeeds}) \tag{3}$$

An agent delegates when $U(\text{delegate}) > U(\text{accept})$, i.e., when the fee is less than the expected loss from attempting the task itself.

3 Experimental Conditions

We compare four information regimes plus an oracle upper bound. The regimes form a monotone information ladder, allowing us to isolate which component of the performance signal drives the accuracy gain.

Table 3: Experimental conditions (information regimes)

Condition	Description
no_rep	Vickrey auction. Agents see only their own state and others' tool lists. No track records.
scalar_rep	Vickrey auction. Agents see each other's aggregate accuracy (e.g., "82%, 22 tasks") but no category or difficulty breakdown.
private_rep	Vickrey auction. Each agent tracks only its own delegation experience with others (decentralized, personal history).
public_rep	Vickrey auction. All agents see full granular reputation vectors: accuracy per category × difficulty level.
oracle	Perfect allocation: each task routed to the specialist with the matching tool. No auction, no generalists.

The comparison *no_rep* → *scalar_rep* → *private_rep* → *public_rep* tests whether the accuracy gain requires granular, category-level performance information or whether coarser signals suffice. The oracle provides the theoretical accuracy ceiling.

4 Results

Data from 3 seeds (42, 123, 999), 100 tasks per seed, 300 tasks per condition. Model: `claude-sonnet-4-20250514`

4.1 Overall Accuracy

Table 4: Overall accuracy by condition (3 seeds, $n = 300$ per condition)

Condition	Correct	Total	Accuracy
oracle	195	200	97.5%
scalar_rep	**283**	**300**	**94.3%**
public_rep (granular)	280	300	93.3%
private_rep	268	300	89.3%
no_rep	264	300	88.0%

Any form of public performance signal — even a simple aggregate accuracy score — adds **+5–6 percentage points** in accuracy relative to no reputation. Surprisingly, the scalar score (94.3%) slightly outperforms the granular category×level breakdown (93.3%). Private reputation, based only on each agent's personal delegation experience, yields a modest +1.3pp overall.

4.2 The Ablation on Hard Tasks (L3+L4)

The overall numbers obscure a critical pattern. Performance signals matter *only* on hard tasks where specialist tools are essential. We therefore focus the ablation on L3+L4 tasks ($n = 150$ per condition):

Table 5: Accuracy on hard tasks (L3+L4) by information regime

Condition	L3	L4	L3+L4	Uplift vs. no_rep
no_rep	85.3%	72.0%	78.7%	—
private_rep	**97.3%**	80.0%	88.7%	+10.0pp
scalar_rep	**100.0%**	85.3%	92.7%	+14.0pp
public_rep (granular)	98.7%	81.3%	90.0%	+11.3pp
oracle	98.0%	92.0%	95.0%	+16.3pp

On hard tasks, private reputation captures 88% of the public reputation uplift (+10.0pp of +11.3pp), despite accumulating from only 9 delegation events (3% of all tasks) across the entire run. The private vector is extremely sparse, yet it is nearly as effective as full public transparency on the tasks where it matters.

At L1/L2 (easy tasks), no condition significantly outperforms the others — all achieve 93–100%. Performance signals provide no benefit where any agent can solve the task without tools.

4.3 Scalar vs. Granular: Less Information, Less Noise

The scalar score outperforming the granular vector is counterintuitive. We hypothesize that with 100 tasks per seed, granular cells are sparse (e.g., "counting_L4: 50% (2×)"), causing agents to overreact to noisy category-level signals. The scalar score is more robust precisely because it aggregates over more observations. This effect may reverse with larger sample sizes where granular cells become statistically reliable.

Table 6: Per-seed accuracy (consistency check)

Seed	No Rep	Private	Scalar	Public
42	91.0%	91.0%	**97.0%**	93.0%
123	86.0%	87.0%	93.0%	**95.0%**
999	87.0%	90.0%	**93.0%**	92.0%
Mean	88.0%	89.3%	**94.3%**	93.3%

The scalar advantage is consistent across 2 of 3 seeds (42, 999). In seed 123, granular public reputation is slightly better. The difference between scalar and granular is within noise at this sample size; the robust finding is that *any* public signal substantially outperforms no signal.

4.4 Market Behavior

Table 7: Market behavior by condition

Metric	No Rep	Private	Scalar	Public
Tasks with tools	39%	—	72%	69%
Delegations	0	9 (3%)	7 (2%)	11 (4%)
No-bid contracts	3	5	8	5

Two behavioral patterns are robust across conditions:

Tool routing. Public signals (scalar or granular) approximately double tool usage rates (39% → 69–72%), indicating better matching of tasks to agents with the appropriate tools.

Delegation. Delegation occurs in all reputation conditions (2–4% of tasks) but is absent without any performance signal. Agents require *some* basis for assessing counterparty competence before subcontracting.

4.5 Adverse Selection Pattern

Table 8: Agent performance by reputation condition (2-seed data)

	Public Rep		No Rep	
Agent	Won	Accuracy	Won	Accuracy
Worker-M	56	100%	51	92%
Worker-X	52	94%	58	100%
Worker-C	50	90%	35	86%
Worker-G1	17	100%	25	92%
Worker-G2	22	91%	28	**54%**

Without performance signals, Worker-G2 (generalist, no tools) won 28 tasks and **failed 46% of them**. It underbid specialists on tasks requiring tools (modular L3, matrix L4) and could not solve them. This pattern is consistent with classical *adverse selection*: without quality signals, low-capability agents can underbid high-capability agents, degrading aggregate quality.

With public performance signals, Worker-G2 won 22 tasks at 91% accuracy — the market limited it to tasks where tools are not needed.

Caveat: This finding is partly driven by a single agent instance (Worker-G2). With only 5 agents, individual agent behavior can have outsized effects on aggregate statistics. More agent instances and randomized identities are needed to establish whether this is a robust market-level phenomenon.

5 Mechanism Design

5.1 Why Vickrey?

The Vickrey auction was chosen because truthful bidding is the weakly dominant strategy in the standard single-item, private-value setting. Our environment departs from this textbook case in several ways: agents have reputational externalities across rounds, delegation introduces interdependencies, and LLM agents may not compute optimal strategies reliably. We therefore do not claim that observed bids are strictly truthful. However, the Vickrey mechanism removes the most obvious confound — strategic bid-shading — and makes it more plausible that price differences between conditions reflect information effects rather than bidding strategy differences.

5.2 Why No Enforced Penalties?

In decentralized markets, penalties are typically unenforceable — an agent can default and walk away. We therefore use *reputation as the only penalty mechanism*: wrong answers receive no payment and are publicly recorded. This tests whether reputation-based incentives, without enforced token slashing, can maintain task quality in this setting.

The result — 93.3% accuracy with reputation-only penalties — suggests that in this market configuration, reputation-based penalties are compatible with high task accuracy. Whether this generalizes to settings with higher stakes, adversarial agents, or longer time horizons remains an open question.

5.3 Utility Function Design

Agents receive only $U = \sum p(\text{correct}_i) \times \text{payment}_i$ as their objective. No strategy hints about bidding, delegation, or reputation management. We observe that agents produce behavior that is *broadly consistent* with utility maximization — specialists bid low on domain tasks, generalists pass on hard tasks (with reputation), and delegation targets agents with relevant tools. However, we cannot verify that agents are actually computing expected utility; the observed behavior may also reflect simpler heuristics that happen to align with rational strategies in this setting.

6 Discussion

6.1 What the Ablation Reveals

The four-condition ablation yields a clear information hierarchy, but with a surprising structure:

1. **Any public signal ≫ no signal.** The jump from no_rep (88.0%) to either scalar (94.3%) or granular public (93.3%) is large and consistent. The critical threshold is *having* a public track record, not its granularity.

2. **Granularity does not help (at this scale).** The scalar score slightly outperforms the granular vector, likely because sparse category-level cells introduce noise. This may reverse with larger datasets where granular cells become reliable.

3. **Private experience captures most of the effect on hard tasks.** On L3+L4 tasks, private reputation achieves 88% of the public uplift despite an extremely sparse delegation history (∼3 data points per agent pair). This suggests that even minimal personal experience is highly informative for routing decisions on difficult cases.

4. **The effect is absent on easy tasks.** All conditions achieve 93–100% on L1/L2. Performance signals provide value only where specialist capabilities are load-bearing.

6.2 Implications

- **Performance monitoring has measurable value.** Any form of public performance tracking — even a simple accuracy score — produces +5–6pp accuracy gains in this setting. This maps to agent monitoring infrastructure (logging, per-agent quality metrics) in production systems.

- **Private experience is an asymmetric asset.** On hard tasks, private delegation experience captures most of the public signal's value. An intermediary who accumulates this experience across multiple clients builds a proprietary quality signal that individual clients cannot efficiently replicate — particularly when individual client volume per task category is low.

- **The value concentrates in hard cases.** Performance signals provide no measurable benefit for easy, routine tasks. The economic value of quality-based routing sits in the long tail of difficult, specialist-dependent work — precisely the cases where errors are most costly and correct routing most valuable.

6.3 Limitations

- **Sample size.** 300 tasks per condition (3 seeds). While effects are consistent across seeds, the total sample remains modest for strong statistical claims.

- **Few agent instances.** Five agents (3 specialists, 2 generalists) is a small population. The adverse selection finding is partly driven by specific agent realizations (Worker-G2). More agent instances and randomized identities would strengthen robustness.
- **Myopic agents.** Agents optimize per-round, not over time. They do not strategically protect reputation for future income.
- **Single model.** All agents use the same LLM (Sonnet). Cross-model heterogeneity (e.g., Haiku vs. Opus agents in the same market) is unexplored.
- **Stateless LLMs.** Each API call is independent — agents cannot truly "learn" across rounds. Performance history is injected via prompt context, which conflates reputation effects with prompt length effects.
- **Task domain.** Mathematical tasks with deterministic ground truth. Generalization to subjective, open-ended, or multi-step tasks is unknown.

7 Next Steps

1. **Shuffled task order.** The current design presents tasks sorted by difficulty level, confounding reputation accumulation with task difficulty. A shuffled-order run would isolate the temporal learning effect: does the performance signal advantage grow over rounds as reputation vectors become richer?
2. **Pooled learning.** Test how quickly a broker's private quality database converges toward public-signal accuracy when pooling cases across multiple clients, compared to a single client building its own experience.
3. **Agent instance robustness.** Randomize agent identities, vary the number of specialists and generalists, and test whether the adverse selection pattern is robust across configurations.
4. **Long-term optimization.** Give agents a discount factor δ for future income: $U = \sum_t \delta^t \cdot E[\text{payoff}_t(\text{reputation}_t)]$. This tests strategic reputation protection.
5. **Cross-model markets.** Run Haiku, Sonnet, and Opus agents in the same market to test whether price and reputation correctly reflect capability differences across model tiers.

8 Conclusion

In this setting, a decentralized market of LLM agents with heterogeneous tool access achieves 93–94% accuracy on mathematical tasks — approaching the oracle upper bound of 97.5%. A four-condition ablation (no signal, private experience, scalar public score, granular public vector) reveals that:

- Any public performance signal adds +5–6pp overall accuracy. The critical threshold is *having* a signal, not its granularity.
- On hard tasks (L3+L4), private experience alone captures 88% of the public signal uplift, despite an extremely sparse delegation history.
- The effect is entirely concentrated in tasks where specialist tools are essential. Easy tasks show no benefit from any information regime.

The core observation is that *performance information enables coordination that instruction design alone does not achieve in this setting.* No amount of prompt engineering enables an agent

to compute $3^{7065} \bmod 499$ reliably without `pow(base, exp, mod)`. A market with even a simple accuracy score routes this task to an agent that has the right tool. The finding that private experience is nearly as effective as public signals on hard tasks suggests that an intermediary accumulating routing experience across clients could capture most of the coordination value — a result with direct implications for the design of multi-agent systems.