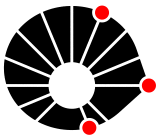

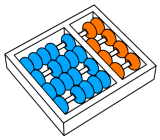

Sponsored Discovery Mechanism for the Web of Open Agents

Filipe Lacerda Benevides *Julio Cesar dos Reis*

Relatório Técnico - IC-PFG-26-10
Projeto Final de Graduação
2026 - Julho

UNIVERSIDADE ESTADUAL DE CAMPINAS
INSTITUTO DE COMPUTAÇÃO

Sponsored Discovery Mechanism for the Web of Open Agents

Filipe Lacerda Benevides Julio Cesar dos Reis*

Abstract

As autonomous LLM agents become primary consumers of web services, they discover and delegate to one another in open ecosystems – a *Web of Open Agents* – yet existing discovery infrastructures offer no principled way to govern how agents *compete for visibility*, risking pay-to-win dynamics in which advertising spend, not merit, drives selection. We propose *Sponsored Discovery*, whose core is MACS (Match-gated Agent Composite Score): semantic domain match acts as a *multiplicative gate* that structurally bounds out-of-domain agents regardless of bid or reputation, coupled with a multi-dimensional behavioral reputation over accuracy, latency, and cost. Using real LLM agents that answer MMLU-Pro questions across three adversarial scenarios (honest baseline, domain-falsifying fraudster, SLA/cost-inflated agent), MACS achieves the highest delegation success across all three. Our results from a paired ablation show MACS's advantage over a purely additive composite: *decisive* under domain misreporting and *marginal* in benign settings, where behavioral reputation alone suffices, while bid-only and quality-adjusted baselines are exploited. This study further analyzes its incentive properties and limits (calibrated misrepresentation, Sybil whitewashing).

Keywords – Web of Open Agents · Agent Advertisement and Promotion · Sponsored Discovery · Multi-agent Markets.

*Instituto de Computação, Universidade Estadual de Campinas, 13081-970 Campinas, SP.

1 Introduction

The Web has always been a marketplace of attention: search, recommendation, and advertising allocate visibility to human users – mechanisms designed and calibrated for one consumer archetype, the human, with all their biases and susceptibility to perceptual influence.

A fundamental shift is now underway. Autonomous AI agents powered by Large Language Models (LLMs) are increasingly acting as the primary consumers of web services, delegating tasks, invoking APIs, and orchestrating multi-step workflows without direct human involvement [1]. This transition – from a web navigated by humans to a *Web of Open Agents* in which software entities discover, negotiate, and transact with one another [2] – invalidates many of the assumptions underlying current discovery and promotion mechanisms.

An autonomous AI agent selects a service provider not because of an evocative brand image or a strategically placed advertisement, but because of functional compatibility, cost, latency, and trust. The unit of optimization shifts from *click-through rate* to *delegation success*: the probability that a delegated task is completed within acceptable SLA and cost constraints. This shift eliminates susceptibility to perceptual influence – an autonomous agent cannot be swayed by brand aesthetics or emotional framing – but introduces a distinct attack surface: an agent that *falsely declares* broad domain expertise exploits precisely this rational, capability-driven selection process. Indeed, agents are systematically mis-calibrated about their own ability and cost, so auctions built on self-reported signals diverge from the full-information allocation [3]. Mechanisms that surface only observable, outcome-grounded quality signals are therefore necessary.

The problem sits at the bottom of a short funnel connecting the *agentic economy* [4] to a concrete mechanism: as agents become economic participants they must *discover* one another, discovery at scale requires *ranking* candidate providers, and once a ranking exists the question of *how visibility is acquired* – through quality, reputation, and economic investment – becomes unavoidable. *Advertisement and promotion* is the mechanism that governs this competition, and it is the level at which this study intervenes.

Despite growing momentum around agent communication protocols [5] and emerging infrastructures for agent registration and discovery [2], the question of *how*

agents and services compete for visibility in open and heterogeneous agent ecosystems remains largely unaddressed. Frameworks such as NANDA [6] and Agntcy/ADS [7] provide solid foundations for service registration and basic discovery, but offer no mechanisms for governed economic promotion: no model for how a high-quality provider with low budget can compete with a low-quality provider with high advertising spend, and no principled way to ensure that visibility reflects merit rather than purchasing power.

This gap is not merely technical – it is economic and governance-theoretic. In open agent ecosystems, discovery is not neutral: the mechanisms that determine which agents appear at the top of a result set shape incentive structures, long-term market dynamics, and the welfare of all participants. Without a principled approach to advertisement and promotion, agent marketplaces risk converging on pay-to-win dynamics that degrade service quality, exclude high-quality newcomers, and erode the trust of agent-to-agent interactions.

This study proposes and evaluates a *Sponsored Discovery* mechanism for the *Web of Open Agents* – one that governs how agents and services gain visibility in open ecosystems, combining semantic eligibility, economic promotion, and trust constraints. The proposed mechanism is formally grounded in three converging bodies of literature: the tradition of service advertisement and semantic matchmaking from the Semantic Web, the mechanism design literature on sponsored search and auction theory, and the emerging body of work on agentic markets and multi-agent economies.

This study makes the following contributions:

- An experiment-independent *Sponsored Discovery Mechanism* whose core is the Match-gated Agent Composite Score (MACS) – semantic match acts as a *multiplicative gate* that structurally bounds out-of-domain agents regardless of bid or reputation – situated on a complexity ladder (BAR, QBAR, ACS) in which the additive ACS isolates the gate's contribution, plus a GSP-style payment rule that grounds revenue and incentives;
- A multi-dimensional behavioral reputation model combining task accuracy, SLA compliance (latency), and cost efficiency into a single observable, outcome-grounded quality proxy;
- An adversarial evaluation on a testbed of real LLM agents (GPT-4o-mini, GPT-3.5-turbo) answering MMLU-Pro questions across three scenarios – honest base-

line, domain-falsifying fraudster, and SLA/cost-inflated agent – measuring delegation success and provider diversity against BAR, QBAR, and the ACS ablation.

The remainder of the study is organized as follows. Section 3 reviews the literature that grounds the mechanism and situates our contribution. Section 4 formally defines the proposed mechanism. Section 5 describes the experimental setup and results. Section 6 discusses implications and limitations, and Section 7 concludes.

2 Background

This section grounds the proposed Sponsored Discovery mechanism in the literature. We move from the *agentic economy* that makes agent-to-agent visibility an economic question (Section 2.1), through the *protocols and infrastructure* that let agents transact (Section 2.2), to the *sponsored-search* theory that governs paid ranking (Section 2.3) and the *semantic matchmaking* that determines eligibility (Section 2.4), noting at each step what the literature leaves open.

2.1 The Agentic Economy

At the broadest level, a recent line of work asks *what kind of economy* emerges when autonomous agents become first-class economic actors. Rothschild *et al.* [4] envision an *agentic economy* of assistant and service agents transacting programmatically, and single out *advertising and discovery* as a mechanism that must be re-examined when the consumer is an agent rather than a human – but at the level of vision, without prescribing a concrete mechanism. Empirically, Bansal *et al.* [8] (*Magentic Marketplace*) show that the *search mechanism itself* – not merely model capability – governs welfare and that agents are measurably manipulable, with performance degrading sharply at scale; related testbeds reinforce that institutional structure is decisive [9, 10].

Together these frame the gap this study fills: the agentic economy needs governed advertisement and discovery, yet no work prescribes a concrete, merit-constrained promotion mechanism. The practical substrate already exists in LLM orchestration frameworks (AutoGen, LangGraph, CrewAI) that route by capability or role but expose no economic layer; we treat them as foundations and return to them as operational competitors in Section 3.

2.2 Agentic Markets: Protocols and Infrastructure

Realizing such an economy requires infrastructure for discovery, identity, and payment – which surveys identify as the missing pieces of current agentic deployments [11], and which remain fragmented because no protocol (MCP, ACP, A2A, ANP) standardizes *capability advertisement* [5]. Concrete substrates – blockchain-settled interaction and ledger-anchored identity with micropayments – have begun to fill these gaps; we treat them as the substrate our mechanism runs on and compare against them directly in Section 3.

This body of work also establishes three ways agentic markets differ from human ones, all shaping promotion design. First, agents have *no perceptual susceptibility* but a new attack surface: they ignore aesthetics and branding, yet are vulnerable to *semantic misrepresentation* – an agent falsely declaring capabilities exploits the very rationality of capability-driven selection – so promotion must rest on observable, outcome-grounded signals. Second, agents apply *consistent, bias-free criteria*, making outcomes more predictable and closer to theoretical models. Third, the welfare metric is *delegation success* (task completion), not clicks or impressions, so mechanisms that harm success rates ultimately reduce platform value.

2.3 Sponsored Search and Auction-Based Allocation

The allocation of ranked positions to competing bidders, in exchange for payment, was formalized for internet search advertising. Edelman *et al.* [12] analyzed the *Generalized Second-Price* (GSP) auction behind AdWords – a Nash equilibrium in locally envy-free strategies, but not incentive-compatible – and Varian [13] characterized position-auction equilibria; the strategy-proof Vickrey–Clarke–Groves alternative maximizes welfare but has weaker revenue in practice, which is why GSP became the industry standard.

The single most important concept for agent-to-agent promotion is the *quality score*: Google multiplies the advertiser's bid b_i by a quality estimate q_i, so the effective bid is $b_i\, q_i$ rather than the bid alone, preventing the highest spender from always winning regardless of relevance. For agent ecosystems the analogue of quality is not click-through rate but a combination of semantic match, trust, and behavioral reputation, and the core principle carries over: **monetary bids should modulate, not override, merit-based signals** – the principle that distinguishes our composite

score from naïve pay-to-win. Richer variants (multi-slot position externalities, budget pacing, fairness constraints, reserve prices) [12, 13] extend but do not alter this insight.

2.4 Service Advertisement and Semantic Matchmaking

Making computational services discoverable by software agents is foundational to the Semantic Web. Early keyword-based registries (UDDI) gave way to *semantic matchmakers* that compare ontology-based capability descriptions to requests and return graded compatibility rather than binary matches: Paolucci *et al.* [14] showed that OWL-S [15] matching outperforms keyword search through subsumption reasoning, and Li and Horrocks [16] extended this to partial matches and degrees of compatibility.

Two insights carry directly into our design. First, *filtering before ranking*: semantic eligibility must be established before any economic ranking – a pattern consistent across matchmaking architectures. Second, the *richness of service descriptions* bounds the ceiling quality of any discovery mechanism built on them. These motivate our two-stage design – a Discovery & Matching stage produces the semantically eligible candidates, and the Sponsored Discovery mechanism operates exclusively within that set – with behavioral reputation as an observable quality proxy and delegation success (not click-through rate) as the primary evaluation metric.

3 Related Work

The Sponsored Discovery mechanism proposed in this study sits at the intersection of four research threads that both ground its design and situate its contribution; each solves part of the agent-promotion problem but leaves the combination open. We group prior work by the *approach* it takes, stating for each cluster what it grounds in our design and the gap our mechanism fills. Table 1 then positions MACS against the most comparable systems along seven dimensions.

Selecting an agent by capability. A first cluster decides *who handles a task* from declared capability. Classical multi-agent systems framed this as capability-based matchmaking [17]; the Semantic Web tradition sharpened it into *semantic matchmaking*, comparing ontology-based capability descriptions to requests and re-

turning graded compatibility – Paolucci *et al.* [14] showed OWL-S [15] matching beats keyword search, and Li and Horrocks [16] extended it to partial matches – establishing the principle we adopt: *filtering before ranking* (semantic eligibility precedes any economic ranking), with description richness bounding the achievable quality. Contemporary LLM orchestration frameworks (AutoGen, LangGraph, CrewAI) inherit the same logic *inside* a single, centrally designed workflow, selecting sub-agents by role or capability and judged on end-task completion rather than on the selection rule itself; a learned variant trains the routing function directly [18]. What unifies the cluster – and separates it from our setting – is that selection is cooperative and closed: no economic competition for visibility, no budget, no independent providers bidding to be surfaced. We keep capability matching as the *eligibility* stage but add a dynamic, economically mediated *ranking* stage on top of it.

Earning trust from behavior in adversarial populations. A second cluster resists bad actors by deriving trust from feedback. EigenTrust [19] propagates local trust transitively into a global peer score, evaluated against colluding peers in peer-to-peer file sharing; Beta reputation systems [20] model trust as a Beta distribution over positive/negative ratings with principled Bayesian updates. Both are adversarial-aware, yet both assume *binary, one-dimensional* feedback and a broadly stationary notion of trust. Our reputation signal departs on the two axes that matter for agentic tasks: it is *multi-dimensional* (accuracy, latency, cost) and grounded in *objective task outcomes*, and it uses a *sliding window* so that strategically drifting behavior is demoted quickly rather than diluted by a long history.

Building the marketplace substrate. A third cluster supplies the infrastructure on which agent commerce runs. Surveys identify discovery, identity, and payment as the missing pieces of current deployments [11], and note that no protocol (MCP, ACP, A2A, ANP) standardizes *capability advertisement* [5]. Concrete substrates fill these gaps: the Coral Protocol [2] treats economic transactions as first-class citizens with blockchain settlement; ledger-anchored A2A economies [21] add persistent, tamper-resistant identity and native micropayments; and Agent Exchange [22] – the closest work to ours in intent – proposes an auction platform for agents inspired by real-time bidding. These are presented as architectures and design visions rather than evaluated against adversarial selection. Crucially, they specify *where value settles*

and how identity persists, not *how visibility is allocated when providers compete on merit*. MACS is exactly that allocation rule: it could run as the selection policy *inside* Coral or Agent Exchange, and the persistent identity of ledger-anchored designs is the natural complement that closes the whitewashing/Sybil gap our reactive reputation cannot close alone (Section 6).

Measuring manipulability in the agentic economy. A fourth cluster studies the emerging *agentic economy* and how selection shapes its outcomes. Rothschild *et al.* [4] envision assistant and service agents transacting programmatically and single out advertising and discovery as mechanics to re-examine when the consumer is an agent. Empirically, Magentic Marketplace [8] shows the *search mechanism* – not just model capability – governs welfare and that agents are measurably manipulable; complementary testbeds confirm that institutional structure is decisive [9, 10], and that LLM agents are mis-calibrated about their own success and cost [3]. This line *documents* the vulnerability but stops short of a governed, merit-constraining promotion mechanism, and the specific threat of an agent *falsely advertising* domain expertise remains underexplored under competitive selection. Our work is the mechanism-design counterpart: we propose a selection rule designed to be robust to exactly this manipulation and evaluate it against a domain-falsifying LLM agent on a ground-truth benchmark.

Across these clusters, no prior approach combines an economic promotion signal (bids) layered on semantic eligibility, a match-gated composite that blocks *both* bid dominance and reputation exploitation, and an adversarial, ground-truth evaluation with real LLM agents. That combination is the contribution of this study. Table 1 makes the positioning explicit.

4 The Sponsored Discovery Mechanism

This section specifies the proposed Sponsored Discovery mechanism. We formalize the selection problem and the per-agent signals, define the semantic match score and the Match-gated Agent Composite Score (MACS) with its multiplicative gate, the multi-dimensional behavioral reputation that feeds it, the bid-adaptation rule, the winner-selection step, and the payment rule, and close with the mechanism's incentive properties. The concrete values used in our evaluation are deferred

Table 1: Comparison of the proposed mechanism with the most directly comparable prior work. ✓: supported; ✗: not supported; ∼: partial. Columns: Cap. MM [17]; GSP [12]; EigenTrust [19]; Coral [2]; Magentic [8]; AEX [22]; Ours.

Dimension	Cap. MM	GSP	EigenTrust	Coral	Magentic	AEX	**Ours**
Economic mechanism (bids/auctions)	✗	✓	✗	✓	✓	✓	✓
Semantic / capability-based eligibility	✓	✗	✗	✓	✓	✓	✓
Behavioral reputation (observed outcomes)	✗	✗	✓	✗	✗	∼	✓
Adversarial evaluation	✗	✗	✓	✗	✓	✗	✓
LLM agents as participants	✗	✗	✗	✓	✓	✓	✓
Ground-truth benchmark evaluation	✗	✗	✗	✗	✓	✗	✓
Newcomer / cold-start treatment	✗	✗	∼	✗	✗	✗	✓

to Section 5.

4.1 Problem Formulation

Let $\mathcal{A} = \{A_1, \ldots, A_N\}$ be a pool of registered agents. Each agent A_i is characterized by: a set of domain specializations $D_i \subseteq \mathcal{D}$, where $\mathcal{D}$ is the universe of task categories; a promotional bid $b_i \in [b_{\min}, b_{\max}]$ representing willingness to pay for visibility; a structural trust score $t_i \in [0, 1]$, reflecting verifiable properties such as operator attestations or protocol compliance; and a behavioral reputation $r_i \in [0, 1]$, computed from observed task outcomes (Section 4.4).

The system operates over discrete rounds $t = 1, \ldots, T$. In each round, the orchestrator receives Q task queries, each with a category $c \in \mathcal{D}$. For each query, the Sponsored Discovery mechanism: (i) computes a semantic match score m_i for each active agent A_i ; (ii) scores all agents according to the selection mechanism; (iii) selects the highest-scoring agent as the winner; and (iv) delegates the task. After the task completes, the outcome (correctness, latency, token cost) is measured and used to update r_i.

4.2 Semantic Match Score

The semantic match score m_i for agent A_i on a query of label ℓ reflects the degree to which A_i's declared specializations cover the task domain:

$$m_i(\ell) = \begin{cases} m_{\text{in}} & \text{if } \ell \in D_i \\ m_{\text{out}} & \text{otherwise} \end{cases} \tag{1}$$

where $0 < m_{\text{out}} < m_{\text{in}} \leq 1$: m_{in} is the in-domain match score (strong functional coverage) and m_{out} the out-of-domain score (partial coverage). In this simulation, m_i is a binary signal; in a production deployment it would be computed as the cosine similarity between the query embedding and the agent's capability-description embedding, yielding a continuous value in the same range.

4.3 MACS: Match-gated Agent Composite Score

At the core of the mechanism is the *Match-gated Agent Composite Score* (MACS), which scores each eligible agent by combining its signals under two design principles: behavioral reputation enters as an *observed* quality signal, and semantic match acts as a *multiplicative gate* rather than merely an additive term.

$$S_i^{\text{MACS}} = m_i\big(\alpha m_i + \beta b_i + \delta t_i + \varepsilon r_i\big) \tag{2}$$

where $\alpha, \beta, \delta, \varepsilon \geq 0$ weight semantic match, bid, structural trust, and behavioral reputation, respectively; their scale is immaterial (a common factor cancels in the arg max and the payment rule). The weights encode a deliberate priority ordering among signals: behavioral reputation is intended to be the dominant *observable* quality signal, semantic match the secondary merit component, the bid the economic competition signal, and structural trust a lightweight attestation-based prior. In the configuration we evaluate, t_i is a static, uniform prior, so the actively discriminating signals are match, bid, and reputation; the δt_i term is retained for architectural completeness. Note that m_i appears in two roles within the formula: as the outer multiplicative gate and as the inner additive term αm_i. This is intentional − the outer factor enforces a structural ceiling that no amount of bid or reputation can overcome, while the inner term lets semantic match additionally reward agents that are not merely eligible but strongly in-domain. Two senses of "dominant" coexist without contradiction: reputation is dominant *within* the additive composite (largest weight ε), while the match term is dominant *structurally* − the outer factor bounds the whole score − so the gate decides who may compete and reputation ranks those who clear it.

The multiplicative gate. Because m_i multiplies the entire parenthesis *and* also appears inside it, the score of an out-of-domain agent ($m_i = m_{\text{out}}$) relative to an

otherwise identical in-domain agent ($m_i = m_{\text{in}}$, same b_i, t_i, r_i) is

$$\frac{S^{\text{MACS}}_{\text{out}}}{S^{\text{MACS}}_{\text{in}}} = \frac{m_{\text{out}}}{m_{\text{in}}} \cdot \frac{\alpha\, m_{\text{out}} + X}{\alpha\, m_{\text{in}} + X}, \qquad X = \beta\, b_i + \delta\, t_i + \varepsilon\, r_i. \tag{3}$$

This ratio ranges from $(m_{\text{out}}/m_{\text{in}})^2$ when the other signals vanish ($X \to 0$) up toward $m_{\text{out}}/m_{\text{in}}$ as they dominate (X large), so the gate is *adaptive*: it penalizes an out-of-domain agent most harshly precisely when it has little else to offer, and relaxes toward a milder cap only for agents that are otherwise strong on bid, trust, and reputation. The design is motivated by the filtering-then-ranking pattern identified in semantic matchmaking [14]: semantic eligibility must *constrain* selection, not merely adjust it.

Proposition (gate bound). *Fix b_i, t_i, r_i and the weights. The score of an out-of-domain agent relative to an identical in-domain one obeys:*

$$\frac{S^{\text{MACS}}_{\text{out}}}{S^{\text{MACS}}_{\text{in}}} \in \left[\left(\frac{m_{\text{out}}}{m_{\text{in}}} \right)^2, \frac{m_{\text{out}}}{m_{\text{in}}} \right)$$

Proof. Let $g(X) = (\alpha m_{\text{out}} + X)/(\alpha m_{\text{in}} + X)$ and write the ratio of Eq. (3) as $(m_{\text{out}}/m_{\text{in}})\, g(X)$. As $m_{\text{out}} < m_{\text{in}}$, g increases in $X \geq 0$ from $g(0) = m_{\text{out}}/m_{\text{in}}$ to $g(X) \to 1$; multiplying by $m_{\text{out}}/m_{\text{in}}$ gives the bounds. □

A purely additive composite that still includes reputation can let an out-of-domain agent with enough accumulated reputation outscore an in-domain competitor; the multiplicative gate removes this failure by making match a hard-ceiling multiplier, which Section 5 isolates through a controlled additive ablation.

Figure 1 summarizes the mechanism as a single selection cycle: the four per-agent signals feed a weighted sum, the semantic match gates that sum multiplicatively to yield S_i, the orchestrator selects $\arg\max_i S_i$ and delegates the task, and the observed outcome is appended to the agent's history, from which its reputation is recomputed – closing the loop that demotes agents underperforming relative to their declarations.

4.4 Multi-dimensional Behavioral Reputation

The behavioral reputation r_i is computed over a sliding window of W recent task outcomes for agent A_i:

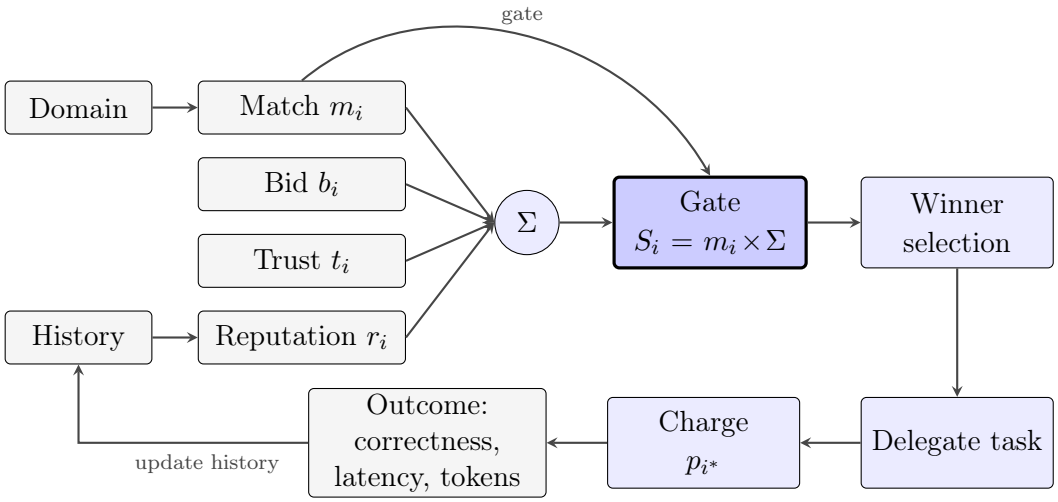

Figure 1: The MACS selection cycle. Each agent contributes a semantic match m_i (from its declared domain), a bid b_i, a structural trust t_i, and a behavioral reputation r_i *computed from* its history; these form a weighted sum Σ. The match additionally gates that sum multiplicatively, $S_i = m_i \times \Sigma$, bounding out-of-domain agents. The orchestrator selects $\arg\max_i S_i$, delegates the task, and charges the winner its GSP-style price p_{i^*} (Section 4.7); the observed outcome updates the agent's reputation – the closed loop that lets MACS demote agents that underperform relative to their declarations.

$$r_i = w_a\, a_i \;+\; \frac{1}{n_i} \sum_{k \in W} c_k \big(w_\sigma\, \sigma(\lambda_k) + w_\gamma\, \gamma(\nu_k) \big), \tag{4}$$

where the non-negative weights w_a, w_σ, w_γ sum to 1 (so that $r_i \in [0,1]$); W_i is the window of the agent's n_i most recent outcomes; $a_i = c_i/n_i$ is its accuracy (c_i correct answers); $c_k \in \{0,1\}$ marks whether outcome k was correct; λ_k, ν_k are its latency (ms) and token count; and

$$\sigma(\lambda) = e^{-\lambda/\tau_\lambda}, \qquad \gamma(\nu) = e^{-\nu/\tau_\nu} \tag{5}$$

are the *SLA score* and *cost score* (exponential decays with scales τ_λ, τ_ν; each equals 1 for an instantaneous, cost-free response and decays toward 0 as latency or tokens grow). Crucially, the SLA and cost terms are credited *only for successful delegations* ($c_k = 1$): being fast or cheap while answering *incorrectly* earns no reputation. This prevents a low-accuracy but cheap/fast agent (e.g. a fraudster) from inflating its

reputation through the SLA/cost terms.

At initialization, all agents have $r_i = 0$. Reputation builds exclusively from observed task outcomes, creating a *cold-start* effect for new entrants (discussed in Section 6).

4.5 Bid Adaptation

Agents adapt their promotional bids using a win-rate equilibrium rule:

$$b_i^{(t+1)} = \text{clamp}\Big(b_i^{(t)}\big(1 + \eta\big(\tfrac{1}{N_t} - \widehat{w}_i^{(t)}\big)\big), b_{\min}, b_{\max}\Big) \tag{6}$$

where $\text{clamp}(x, a, b)$ clamps x to $[a, b]$, $\widehat{w}_i^{(t)}$ is the agent's realized win rate over the last W_b rounds, N_t is the number of active agents, and $\eta > 0$ is the adaptation rate. Agents trending above their fair share ($1/N_t$) lower their bid; agents below their fair share raise it. This mirrors real-time ad-market bidding [13] and curbs persistent overbidding.

4.6 Winner Selection

Given scores $\mathbf{S} = (S_1, \ldots, S_N)$, the winner is the agent with the highest score:

$$i^* = \arg\max_{i \in \mathcal{A}_t} S_i \tag{7}$$

where $\mathcal{A}_t \subseteq \mathcal{A}$ is the set of agents active in round t (agents may enter the pool at a designated entry round, modeling newcomers). Ties are broken uniformly at random.

4.7 Payment Rule and Platform Revenue

The scoring rule above allocates the task; the mechanism must also specify what the winner *pays*. We adopt a critical-value payment: the winner pays the smallest bid that would still have secured the allocation, holding all other agents fixed. With a single advertising slot this is the Generalized Second-Price (GSP) rule [12] in its degenerate form, which coincides with the Vickrey (second-price) payment. Let $S_{(2)}$ be the highest score among the losers. The critical bid p_{i^*} for the winner i^* solves $S_{i^*}(p) = S_{(2)}$; under MACS,

$$p_{i^*} = \text{clamp}\left(\frac{1}{\beta}\left(\frac{S_{(2)}}{m_{i^*}} - \alpha m_{i^*} - \delta t_{i^*} - \varepsilon r_{i^*}\right), b_{\min}, b_{\max}\right). \tag{8}$$

Only the *bid* dimension is charged: the merit signals (m_i, t_i, r_i) shift the threshold but are not themselves paid. The winner is charged $p_{i^*} \leq b_{i^*}$ (it never pays more than it bid), per-round platform revenue is $\sum_Q p_{i^*}$, and the reserve $b_{\min}$ acts as a price floor. This rule anchors the revenue and incentive analysis of Section 6: as a high-quality agent accumulates reputation, the term $\varepsilon\, r_{i^*}$ grows and its critical price p_{i^*} falls − it retains the slot at a lower bid − which is the precise mechanism behind the bid-shading dynamic discussed in Section 6. In the static, one-shot single-slot game this critical-value payment is essentially strategy-proof in the bid − it coincides with the Vickrey payment, so no agent gains by misreporting b_i within a round. The bid-related non-truthfulness in our setting is instead *dynamic*: as an agent's reputation accumulates across rounds, the bid it needs to retain the slot falls, so high-reputation agents progressively shade their bids (Section 6). A genuine GSP incentive gap re-emerges only with multiple slots; its repeated-game analysis, and a VCG-style alternative that charges each winner the externality it imposes, are left to future work.

5 Experimental Evaluation

This section evaluates the proposed mechanism empirically. We first place MACS on a *complexity ladder* against three baselines (Section 5.1). We then describe the MMLU-Pro dataset (Section 5.2), the agent pool (Section 5.3), and our parameter choices (Section 5.4). Finally, we define the three adversarial scenarios (Section 5.5) and the evaluation metrics (Section 5.6), and report results in Section 5.7.

5.1 Selection Mechanisms: A Complexity Ladder

We evaluate the proposed Sponsored Discovery mechanism (Section 4) against a ladder of reference mechanisms of increasing complexity, summarized in Table 2. Each rung adds exactly one capability over the previous one, so the sequence isolates the contribution of each design decision and culminates in MACS.

BAR – Bid-Allocated Ranking. The simplest mechanism: agents are ranked purely by promotional bid, $S_i^{\text{BAR}} = b_i$. BAR is a pure pay-to-win auction – it imposes no quality constraint, so an agent willing to outbid all competitors always wins, regardless of competence.

QBAR – Quality-adjusted Bid-Allocated Ranking. Inspired by the quality score of sponsored search [12], QBAR adds the semantic match as an additive quality term, $S_i^{\text{QBAR}} = \alpha m_i + \beta b_i$, with mechanism-specific weight values (Table 4). This prevents pure bid dominance – an out-of-domain agent with a high bid now faces a semantic penalty – but QBAR has no behavioral memory, and because m_i is self-declared it can be spoofed.

ACS – Additive Composite Score. ACS adds structural trust and *behavioral reputation*, $S_i^{\text{ACS}} = \alpha\, m_i + \beta\, b_i + \delta\, t_i + \varepsilon\, r_i$. It thus gains outcome-grounded memory (unlike QBAR), but it still treats semantic match as one additive term among others – there is no gate.

MACS – the proposed mechanism (Section 4). MACS adds the *multiplicative gate*, $S_i^{\text{MACS}} = m_i\,(\alpha\, m_i + \beta\, b_i + \delta\, t_i + \varepsilon\, r_i)$ (Eq. (2)). It is exactly ACS multiplied by the match factor, so the pair (ACS, MACS) is a *controlled ablation*: the two differ *only* in whether m_i gates the composite, so any performance gap between them is attributable to the gate alone. This directly addresses the objection that a quality-adjusted score with behavioral memory might already suffice – ACS *is* that score. Structurally, under ACS an out-of-domain agent that accumulates enough reputation can win (its εr_i term offsets its low m_i), whereas under MACS it remains bounded by the gate (Eq. (3)) regardless of reputation.

5.2 Dataset

We use the MMLU-Pro benchmark [23], a collection of 12,032 multiple-choice questions spanning 14 academic categories. We group these categories into three domain clusters: **STEM** (mathematics, physics, chemistry, engineering, computer science), **Bio** (biology, health, psychology), and **Social** (economics, law, business, history, philosophy, other).

Table 2: The complexity ladder of selection mechanisms. Each rung adds one capability; MACS (the proposed mechanism) is the only rung with the multiplicative match gate.

Mechanism	Score S_i	Capability added
BAR	b_i	bid only (pay-to-win)
QBAR	$\alpha m_i + \beta b_i$	+ semantic match (additive, self-declared)
ACS	$\alpha m_i + \beta b_i + \delta t_i + \varepsilon r_i$	+ trust and behavioral reputation
MACS	$m_i \left(\alpha m_i + \beta b_i + \delta t_i + \varepsilon r_i\right)$	+ multiplicative match gate

Table 3: Agent pool configuration. All agents start with initial bid $b_0 = 0.35$ except where noted. Strategy: FS = few-shot schema, CoT = chain-of-thought schema, ZS_s = zero-shot schema, ZS_r = zero-shot raw. "All" domains means STEM + Bio + Social.

ID	Model	Strategy	Domains	Entry round
A_1	GPT-4o-mini	FS	STEM	0
A_2	GPT-4o-mini	CoT	Bio + Social	0
A_3	GPT-3.5-turbo	ZS_s	STEM	0
A_4	GPT-3.5-turbo	ZS_r		0
A_5	GPT-4o-mini	FS	All	100

MMLU-Pro provides ground-truth correct answers, enabling objective measurement of agent accuracy without requiring a judge model. This is particularly important for our evaluation: we need to distinguish between what an agent *claims* it can do (declared domain) and what it *actually* does (task outcome), which requires unambiguous correctness labels.

5.3 Agent Pool

The pool comprises five agents with varying models, prompting strategies, and domain specializations, as shown in Table 3. Three agents use GPT-4o-mini (a high-quality, cost-efficient model) and two use GPT-3.5-turbo (a weaker baseline model). Prompting strategies range from few-shot schema-guided prompting to zero-shot raw completion.

A_5 enters the pool at round 100 with zero reputation, modeling the cold-start challenge faced by newcomers to an established marketplace. Agent A_4 declares *no* domain specialization: it is therefore always scored out-of-domain ($m_i = m_{\text{out}}$)

Table 4: Parameters used in the evaluation (rationale for the key choices – ε, the gate ratio, and the decay scales – is given in the text). Weights are mechanism-specific: QBAR uses (α, β), while ACS and MACS use $(\alpha, \beta, \delta, \varepsilon)$; their scale is immaterial (it cancels in the arg max).

Parameter	Symbol	Value
In-domain match	m_{in}	0.90
Out-of-domain match	m_{out}	0.55
QBAR weights	α, β	$0.50, 0.50$
MACS weights	$\alpha, \beta, \delta, \varepsilon$	$0.25, 0.20, 0.15, 0.40$
Structural trust	t_i	0.50
Reputation weights	w_a, w_σ, w_γ	$0.60, 0.15, 0.25$
SLA decay scale	τ_λ	2000 ms
Cost decay scale	τ_ν	1000 tok
Reputation window	W	25 rounds
Bid window	W_b	25 rounds
Adaptation rate	η	0.10
Initial bid	b_0	0.35
Bid bounds	$b_{\min}, b_{\max}$	$0.05, 1.00$
Rounds; queries/round	$T; Q$	$200; 5$
Expensive overhead	–	+5000 ms, +3000 tok

and, under MACS, permanently gate-limited. It plays the role of an honest, weak generalist – a low-quality reference distinct from the fraudster, which instead *claims* every domain.

5.4 Parameter Selection

The mechanism of Section 4 is defined over free parameters. Table 4 lists the concrete values used throughout our evaluation; the rationale for the most consequential choices follows. Unless a parameter is varied as part of a sensitivity analysis, these values are held fixed across all mechanisms and scenarios.

Composite weights $(\alpha, \beta, \delta, \varepsilon)$. The MACS weights encode the priority ordering motivated in Section 4.3: reputation is the dominant *observable* signal ($\varepsilon = 0.40$) because it is the only term grounded in verified outcomes rather than self-declaration; semantic match is the secondary merit term ($\alpha = 0.25$); the bid is a genuine but

bounded competition lever ($\beta = 0.20$); and structural trust is a light attestation prior ($\delta = 0.15$), kept small because it is static in this study. Concretely, every agent is assigned the same $t_i = 0.50$ held fixed for the whole run (Table 4), so the $\delta\, t_i$ term is a constant offset that does not differentiate agents within a round; it is retained for architectural completeness and would acquire discriminative weight only once attestation-based trust varies across agents. We chose $\alpha = 0.25$ rather than, say, 0.30 to keep the bid and trust terms jointly able to break ties among equally in-domain agents while still leaving reputation clearly dominant. Because the conclusions should not hinge on a single value, we sweep ε (the reputation weight) over $\{0.2, \ldots, 0.6\}$ while renormalizing the remaining weights proportionally; the analysis is reported with the main results (Table 8).

Match scores ($m_{\text{in}}, m_{\text{out}}$). We set $m_{\text{in}} = 0.90$ rather than 1.0 so that a perfect declared match still leaves headroom, anticipating a continuous embedding matcher; $m_{\text{out}} = 0.55$ represents partial topical overlap rather than total mismatch. Their ratio fixes the gate strength: $m_{\text{out}}/m_{\text{in}} \approx 0.61$ means an out-of-domain agent is capped at 61% of an equivalent in-domain agent's score – and as low as $(m_{\text{out}}/m_{\text{in}})^2 \approx 37\%$ when its other signals are weak (Eq. (3)). This is strong enough to deter domain spoofing, yet mild enough not to hard-exclude agents on near-miss categories.

Windows and adaptation rate (W, η). The sliding window $W = 25$ trades reaction time against stability: too short and reputation is noisy round-to-round; too long and a defecting agent is demoted too slowly. With $Q = 5$ queries per round, $W = 25$ aggregates ≈ 125 recent outcomes. The bid adaptation rate $\eta = 0.10$ is deliberately conservative: larger values cause bids to oscillate as agents over-correct to transient win-rate swings.

Expensive-agent overheads. The +5,000 ms and +3,000-token overheads in Scenario 3 are calibrated against the decay scales: with a typical base latency of ≈ 1 s and $\approx$ 1k tokens, the inflated values land at ≈ 6 s and $\approx$ 4k tokens, giving $\sigma(6{,}000) = e^{-3} \approx 0.05$ and $\gamma(4{,}000) = e^{-4} \approx 0.018$. These near-zero scores make the expensive agent a clean stress test of whether a mechanism penalizes resource footprint independently of accuracy.

5.5 Experimental Scenarios

We evaluate all four mechanisms – BAR, QBAR, ACS, and MACS – under three scenarios designed to test different failure modes. The ACS–MACS pair is the ablation that isolates the multiplicative gate (Section 5.1). For each scenario we state the *ideal quality ranking* – the order a perfectly quality-aware mechanism would produce – which follows directly from each agent's underlying model and prompting strategy (Table 3).

Scenario 1 – Baseline (honest agents). All five agents report domain expertise truthfully and complete tasks with their actual underlying LLM call. The ideal quality ranking, taken from the agents' measured standalone accuracy, is $A_2 \succ A_1 \succ A_4 \succ A_3$ (first 100 rounds) and $A_2 \succ A_1 \succ A_5 \succ A_4 \succ A_3$ thereafter – A_2 (GPT-4o-mini with chain-of-thought) is the most accurate and the GPT-3.5-turbo agents (A_3, A_4) the least; A_1 and A_5 (GPT-4o-mini few-shot) are comparable. Among the two GPT-3.5-turbo agents, the zero-shot *raw* A_4 measured marginally more accurate than the schema-prompted A_3 in our runs, hence $A_4 \succ A_3$. Under this scenario, a good mechanism should gradually surface the most accurate agents without suppressing provider diversity.

Scenario 2 – Fraudster. A_3 is replaced by FRAUD, an agent that *always* reports m_{in} regardless of query category (claiming universal expertise) but has no genuine capability: it answers at random (accuracy $\approx$ chance – about 10%, since MMLU-Pro offers up to ten options per question). FRAUD sets its initial bid to $b_0 = 0.70$ (the honest weak generalist A_4 starts lower, at $b_0 = 0.20$, in this and the expensive scenario). The ideal quality ranking is $A_2 \succ A_1 \succ A_4 \succ$ FRAUD (and $A_2 \succ A_1 \succ A_5 \succ A_4 \succ$ FRAUD once A_5 enters), with FRAUD last.

This scenario tests each mechanism's ability to detect and demote an agent that exploits the semantic match signal. Under BAR, FRAUD's high bid wins immediately. Under QBAR, FRAUD is doubly rewarded: its spoofed match score plus its high bid produce the highest composite score. Under MACS, FRAUD's behavioral reputation collapses after delivering poor answers, and the reputation penalty propagates back through the multiplicative gate.

Scenario 3 – Expensive agent. A_3 is replaced by EXP: a GPT-4o-mini few-shot agent (same real accuracy as A_1) whose *reported* latency and token count are inflated by +5,000 ms and +3,000 tokens per call, simulating a heavily chained or proprietary-wrapped agent whose quality is high but whose resource footprint is prohibitive. The ideal quality ranking under MACS-aware criteria is $A_2 \succ A_1 \succ A_4 \succ \text{EXP}$ (and $A_2 \succ A_1 \succ A_5 \succ A_4 \succ \text{EXP}$ once A_5 enters); A_2 is the strongest and inexpensive, and the SLA/cost penalties place EXP last despite its high accuracy.

5.6 Evaluation Metrics

We evaluate all mechanism-scenario combinations over $T = 200$ rounds with $Q = 5$ queries per round, with reputation computed over a sliding window of $W = 25$ rounds. To account for stochasticity, every configuration is run for 10 independent seeds; we report the mean with 95% confidence intervals, and all time series are shown with error bands (Figure 2).

Delegation Success Rate (DSR). The fraction of all delegated queries answered correctly, cumulated up to round t:

$$\text{DSR}_t = \frac{\sum_{\tau \le t} c_\tau}{tQ} \tag{9}$$

Here c_τ counts the correct answers in round τ. A higher DSR indicates that the mechanism consistently selects more accurate agents.

Provider Diversity Index (PDI). The Shannon entropy of the selection distribution across agents:

$$\text{PDI}_t = -\sum_{i=1}^{N} \frac{x_i(t)}{\sum_j x_j(t)} \log \frac{x_i(t)}{\sum_j x_j(t)} \tag{10}$$

where $x_i(t)$ is agent i's cumulative selection count. A higher PDI (up to $\log N$) indicates that no single agent monopolizes the pool. Monopolization reduces resilience and may indicate the mechanism is exploitable by a single dominant bidder.

Statistical methodology. We call a difference *significant* when the two 95% CIs do not overlap – a conservative criterion. Because seeds are *paired* (a given seed feeds

Table 5: Final metrics at round $T = 200$, mean ± 95% CI over 10 seeds. DSR: Delegation Success Rate; PDI: Provider Diversity Index (Shannon entropy − higher means *less* concentrated, reported as a diversity indicator rather than a quantity to maximize). **Bold** marks the best DSR per scenario.

Scenario	Mechanism	DSR	PDI
Baseline	BAR	0.439 ± 0.008	1.60 ± 0.00
	QBAR	0.448 ± 0.009	1.51 ± 0.01
	ACS	0.447 ± 0.008	1.45 ± 0.04
	MACS	$\mathbf{0.462 \pm 0.007}$	1.15 ± 0.03
Fraudster	BAR	0.366 ± 0.013	1.57 ± 0.01
	QBAR	0.350 ± 0.009	1.56 ± 0.01
	ACS	0.412 ± 0.012	1.49 ± 0.04
	MACS	$\mathbf{0.434 \pm 0.008}$	1.19 ± 0.04
Expensive	BAR	0.459 ± 0.009	1.59 ± 0.01
	QBAR	0.461 ± 0.009	1.49 ± 0.01
	ACS	0.462 ± 0.009	1.44 ± 0.03
	MACS	$\mathbf{0.473 \pm 0.008}$	1.16 ± 0.04

the identical query stream to every mechanism), we also report paired-difference 95% CIs for the principal MACS contrasts (Table 6), which remain significant even where marginal CIs overlap; with 10 seeds and no multiple-comparison correction, borderline gaps warrant caution.

5.7 Experimental Results

Table 5 reports the final metric values at round $T = 200$ for all four mechanisms across the three scenarios, and Figure 2 shows the corresponding time series (delegation success and reputation under MACS); we describe each scenario's dynamics below, and analyze cost and revenue in Section 5.8 (Table 7).

Scenario 1 – Baseline. With all agents honest, the mechanisms start close, since bids are equal and reputation is zero. As reputation accumulates, MACS concentrates selections on the highest-accuracy agents (chiefly A_2, then A_1/A_5), reaching the highest delegation success, significantly above BAR (non-overlapping CIs; Table 5); QBAR and ACS fall in between. Its provider diversity is correspondingly the lowest, reflecting deliberate concentration on quality rather than monopolization by a single

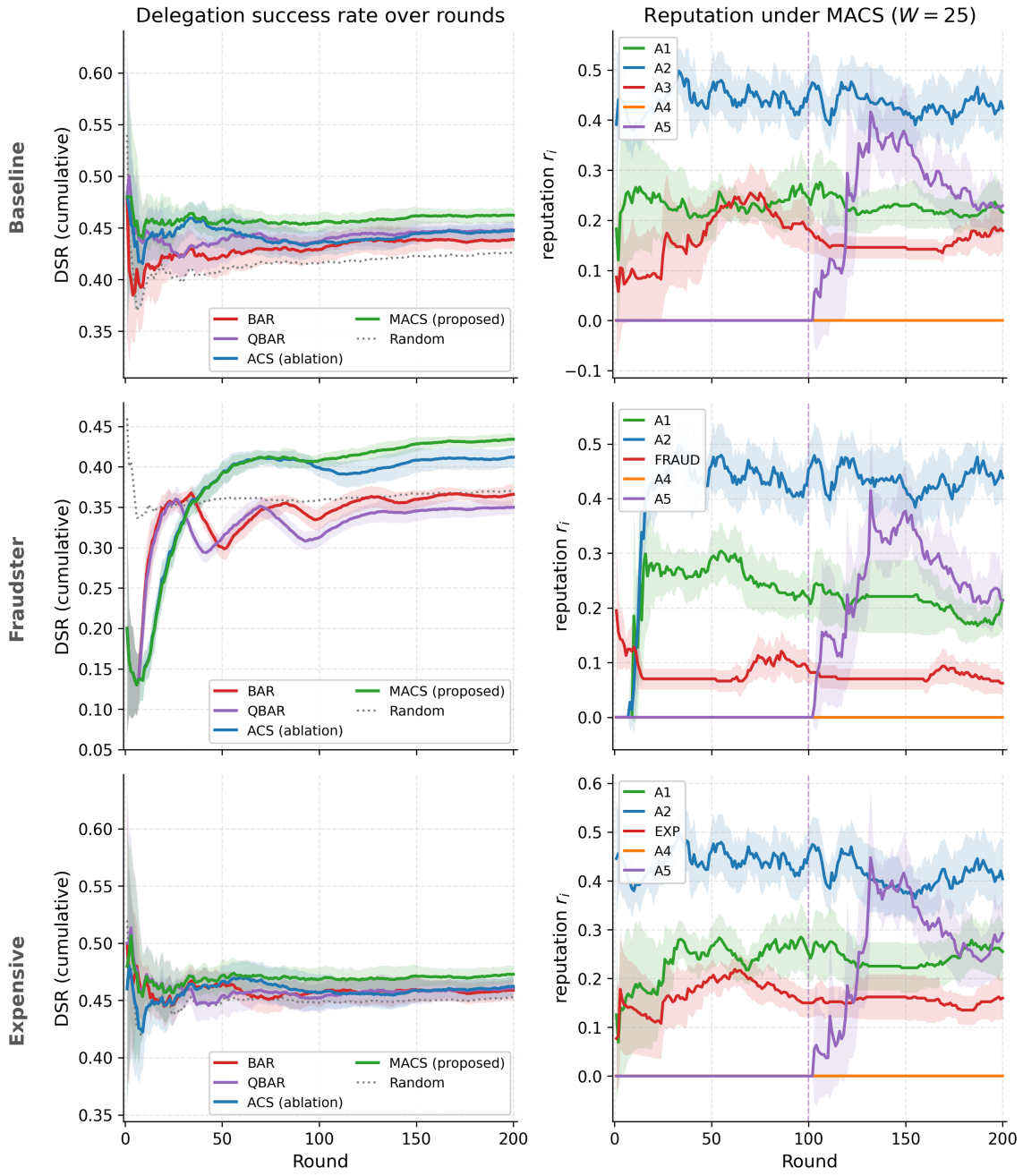

Figure 2: Per-scenario dynamics under all mechanisms (mean over 10 seeds; shaded 95% bands). *Left:* delegation success rate over rounds – MACS and ACS climb while BAR/QBAR are exploited (most visibly in the fraudster row), with the random policies as dotted references. *Right:* behavioral reputation under MACS (one curve per agent); the fraudster row shows FRAUD's reputation collapsing within roughly $W = 25$ rounds, and the honest newcomer A_5's cold-start entry at round 100 is visible in every row.

bidder. Because every agent is honest here, the gaps between mechanisms are modest, as expected.

Scenario 2 – Fraudster. This scenario differentiates the mechanisms most clearly. QBAR is worst: FRAUD's spoofed match score combines with its high bid to keep it selected, and BAR is likewise exploited under a pure pay-to-win rule. Adding behavioral reputation recovers much of the loss (ACS), because FRAUD's random answers earn near-zero reputation once SLA/cost credit is conditioned on success. MACS is best and its margin over ACS is statistically significant (non-overlapping 95% CIs; all values in Table 5), approaching the honest baseline: the reputation collapse demotes FRAUD under both, while the multiplicative gate additionally routes each query to an in-domain specialist, lifting MACS above the purely additive ACS.

Scenario 3 – Expensive agent. BAR and QBAR are largely blind to resource footprint, so EXP (high accuracy but +5,000 ms and +3,000 tokens per call) keeps being selected, inflating latency and token cost without any accuracy gain. Because SLA and cost are credited only on success and decay sharply ($\sigma(6{,}000) \approx 0.05$, $\gamma(4{,}000) \approx 0.018$), EXP's reputation stays low under ACS and MACS, which demote it. MACS attains the highest delegation success and the lowest PDI, concentrating on the cheap, high-quality A_2; the marginal-CI gap over ACS/QBAR is small, but the paired test (shared seeds) confirms MACS>ACS even here ($+0.011 \pm 0.005$; Table 6). The cost this scenario is designed to expose surfaces in the *cost per correct answer* (Section 5.8), not in DSR.

Ablation – isolating the gate (ACS vs. MACS). Because ACS carries the same behavioral reputation as MACS but omits the multiplicative gate, the ACS–MACS gap isolates the gate's contribution. Its value is largest under the fraudster (0.434 vs 0.412, non-overlapping marginal 95% CIs), where gating each query to an in-domain specialist matters most. In the baseline and expensive scenarios the *marginal* CIs of ACS and MACS overlap, but the *paired* difference – appropriate because the seeds are shared – is positive and its 95% CI excludes zero in all three scenarios (Table 6). The gate therefore contributes *beyond* the mere presence of reputation in every scenario, most strongly under adversarial domain misreporting. Across both metrics, DSR – the primary objective – favors MACS in every scenario, while PDI reflects MACS's

Table 6: Paired MACS–ACS DSR difference across the 10 shared seeds (the gate ablation): mean ± 95% CI. Because seeds are paired, this is more powerful than comparing marginal CIs; every CI excludes 0, so the gate helps significantly in each scenario. MACS also beats BAR/QBAR by wider, always-significant margins (+0.024/+0.014 baseline; +0.068/+0.084 fraudster; +0.014/+0.012 expensive).

Scenario	ΔDSR (MACS−ACS)	CI excl. 0
Baseline	$+0.015 \pm 0.006$	✓
Fraudster	$+0.022 \pm 0.010$	✓
Expensive	$+0.011 \pm 0.005$	✓

deliberate concentration on high-quality agents (a guardrail against monopolization rather than a quantity to maximize).

Sensitivity to the reputation weight ε. To confirm that the results do not hinge on the specific $\varepsilon = 0.40$, we swept ε over $\{0.2, 0.3, 0.4, 0.5, 0.6\}$ on the fraudster scenario (5 seeds per setting), renormalizing (α, β, δ) proportionally (Table 8). MACS's delegation success is essentially flat across the range (0.427–0.438, overlapping CIs) and remains above the additive ablation ACS at *every* ε, and far above the ε-independent baselines BAR (0.366) and QBAR (0.350). The gate's advantage and MACS's robustness to adversarial misreporting therefore do not depend on a finely tuned reputation weight.

5.8 Cost-awareness, Revenue, and Gate Robustness

Cost, revenue, and the critical price. DSR counts only correctness, whereas "delegation success" was defined to include cost and SLA; Table 7 therefore reports the *cost per correct answer* ($\bar{\nu}$/DSR, mean tokens per correct answer) alongside the platform economics. In the expensive scenario – built to expose cost – MACS answers correctly at the lowest token cost (1,507 vs. 2,135 for BAR, −29%), because it demotes the token-inflated EXP; the raw DSR hides this burden. In the benign baseline the ordering reverses, as MACS routes to the most accurate agent A_2 (chain-of-thought, hence costlier per call in both tokens and latency, the *ms/correct* column) – an accuracy-versus-latency trade-off rather than waste, while it still eliminates the pathological +5,000 ms overhead of EXP. On the market side, the GSP-style rule is now measured: MACS charges the lowest mean critical price and collects the lowest

Table 7: Cost-awareness and revenue (mean over 10 seeds). *tok/correct* and *ms/-correct*: total tokens and latency per correct delegation ($\bar{\nu}$/DSR, $\bar{\lambda}$/DSR); *revenue*: total platform revenue per run; *price*: mean per-query critical price. MACS is most token-efficient exactly in the cost-adversarial (expensive) scenario, and charges the lowest price everywhere.

Scenario	Mechanism	tok/correct	ms/correct	revenue	price
Baseline	BAR	825	4747	93.6	0.094
	QBAR	853	5525	116.4	0.116
	ACS	853	5501	140.7	0.141
	MACS	886	6665	69.7	0.070
Fraudster	BAR	799	3863	202.1	0.202
	QBAR	805	4615	164.1	0.164
	ACS	851	5549	175.0	0.175
	MACS	885	6853	80.1	0.080
Expensive	BAR	2135	6012	118.4	0.118
	QBAR	1719	6656	104.3	0.104
	ACS	1646	6727	133.4	0.133
	MACS	**1507**	7556	66.8	0.067

revenue in every scenario (e.g. 80 vs. 202 for BAR under the fraudster), and that price *declines* as reputation accumulates – the reputation-for-price substitution of Eq. (8), and the flip side of lower orchestrator cost (Section 6).

Sensitivity to the gate strength. Beyond ε we sweep the gate's single parameter, $m_{\text{out}}/m_{\text{in}}$, holding $m_{\text{in}} = 0.90$ fixed (Table 8, fraudster). MACS is essentially invariant ($0.434 \to 0.431$) because it routes to in-domain agents, whereas the additive ACS degrades as the gate weakens ($0.432 \to 0.403$); the MACS–ACS gap widens accordingly. The headline result is thus robust to the gate calibration (the main experiments use ≈ 0.61), and the gate's value *grows* as the semantic signal weakens.

6 Discussion

Reactive vs. proactive fraud detection. MACS is a *reactive* mechanism: it demotes bad actors only after observing their behavior. In Scenario 2, there is an initial window – on the order of the reputation window ($W = 25$) – during which FRAUD

Table 8: Robustness of the fraudster-scenario result to the two key weights (final DSR, mean ± 95% CI; 5 seeds per setting). *Top:* reputation weight ε, with (α, β, δ) renormalized. *Bottom:* gate strength $m_{\text{out}}/m_{\text{in}}$, with $m_{\text{in}} = 0.90$ fixed. MACS stays essentially flat and above the additive ACS throughout; the match-independent BAR (0.366) and QBAR (≈ 0.35) are low references.

Swept parameter	value	MACS DSR	ACS DSR
Reputation weight ε	0.20	0.427 ± 0.009	0.415 ± 0.010
	0.30	0.433 ± 0.012	0.417 ± 0.012
	0.40	$\mathbf{0.434 \pm 0.008}$	0.412 ± 0.012
	0.50	0.433 ± 0.013	0.428 ± 0.016
	0.60	0.438 ± 0.008	0.414 ± 0.012
Gate $m_{\text{out}}/m_{\text{in}}$	0.4	$\mathbf{0.434 \pm 0.008}$	0.432 ± 0.007
	0.6	$\mathbf{0.434 \pm 0.008}$	0.412 ± 0.009
	0.8	0.431 ± 0.007	0.403 ± 0.012

is selected before its reputation collapses, delivering low-quality answers. For high-stakes orchestration pipelines, this cold-start vulnerability may be unacceptable. A proactive complement – such as capability probing (sending a small set of test queries before admitting an agent to the pool) or operator-issued capability attestations anchored to verifiable credentials [21] – could reduce this window at the cost of added complexity and upfront resource expenditure.

Identity, whitewashing, and cold start. Because reputation is reactive and starts at zero, two identity-linked limitations follow. *Whitewashing*: a bad actor whose reputation has collapsed can re-register under a fresh identity to reset its history, letting FRAUD evade demotion indefinitely – a known attack on distributed reputation systems [19, 20] against which MACS has no structural defense without persistent identity; ledger-anchored identifiers [21] make re-registration as costly as funding a new on-chain identity and close this gap. The mirror-image problem is the honest *cold-start barrier*: A_5 enters at round 100 with zero reputation and initially ranks below incumbents regardless of its true quality. Provisional trust injection, shadow evaluation before admission, or time-decayed reputation (which ages incumbents' advantage) would mitigate it. Both stem from the same zero-start, identity-bound design and warrant dedicated treatment in future work.

The DSR–diversity trade-off. Delegation success and provider diversity are not jointly maximizable: always routing to the single best agent maximizes DSR but minimizes PDI, and uniform routing does the reverse. MACS therefore operates on a DSR–PDI *frontier*; we treat PDI as a monopolization *guardrail* rather than an objective, and the bid-adaptation rule keeps the operating point off the degenerate single-winner corner (a deployment can move along the frontier by tuning $b_{\min}$ and η). Reporting both metrics makes the trade-off explicit, echoing evidence that stronger competitive selection can paradoxically *degrade* market performance [9].

From declared to computed match. In our simulation the fraudster *declares* m_{in}; in the production architecture m_i is instead the platform-computed similarity between the query and the agent's *published* capability description, which an agent cannot inflate without simultaneously diluting its match on the queries it genuinely covers. The computed matcher is thus itself a structural defense against crude spoofing, so our binary-declaration fraudster is a worst case; what survives is the subtler *calibrated misrepresentation* discussed below – which is why outcome-grounded reputation, not the self-reported match, carries the dominant weight [3].

Incentive properties and their limits. The proposed mechanism does not achieve full strategy-proofness in the mechanism-design sense [13]; two constraints in particular bound its robustness. First, *bid shading*: once an agent accumulates high behavioral reputation it can lower its bid and still win the majority of in-domain queries. The win-rate equilibrium rule (Section 4.5) curbs runaway overbidding, but its fixed point is heuristic rather than provably incentive-compatible, so high-quality agents tend to shade bids over time – reducing platform revenue but also the orchestrator's cost, an alignment that may be desirable in agentic markets. Empirically this is exactly what we observe: MACS charges the lowest price and collects the lowest platform revenue in every scenario (Section 5.8). This invites a fair objection – why would a revenue-maximizing platform adopt a rule that drives per-slot price toward the floor? The reserve $b_{\min}$ is a direct lever for targeting revenue, and a low per-query price is the cost of sustaining the high delegation success and provider trust that grow transaction *volume*, whereas a pay-to-win rule maximizes short-term per-slot revenue while degrading quality and eroding the marketplace – the death spiral MACS is designed to prevent. The win-rate rule also creates two second-order effects: a race toward lower

bids among consistent winners, and a structural tension around generalism, since the gate deliberately caps an *undeclared* generalist such as A_4 − the intended defense against spoofing, but also a penalty on legitimately broad agents that a continuous matcher rewarding genuinely broad *published* descriptions would soften. Second, *calibrated misrepresentation*: a subtler adversary than our fraudster could declare expertise only in categories where its actual accuracy is just sufficient to sustain a non-negligible reputation, while still extracting the match multiplier m_{in} on queries it cannot genuinely handle. MACS offers no structural defense against this calibrated attack; proactive capability probing or cryptographic capability attestations [21] are required to close the gap.

Simulation scope and generalizability. Our simulation is deliberately small: a pool of $N = 5$ agents over $T = 200$ rounds with $Q = 5$ queries per round − modest relative to real production traffic. This scale limitation deserves emphasis: our own motivating evidence shows that agentic-market performance can *degrade sharply with scale* [8], so our findings should be read as established in a small-pool regime, with validation at larger agent counts and horizons a necessary next step. Within this regime, results are averaged over 10 independent seeds with 95% confidence intervals (Table 5); the metric that most cleanly separates the mechanisms is delegation success, with provider diversity (PDI) read as a guardrail against monopolization rather than a quantity to maximize. Additionally, the semantic match signal is binary in this implementation; a production system would use embedding-based continuous similarity, which would modulate the gate continuously rather than in a step function. Structural trust (t_i) is static throughout the simulation; a realistic implementation would update it based on cryptographic attestations, operator vouching, or protocol compliance signals.

Reproducibility. All randomness is seed-controlled (seed = $137\,r$) and answers come from a cache keyed by (strategy, model, question), so a run is reproducible given the cache; we fix the model versions and MMLU-Pro question set and will release code, prompts, and per-seed logs. As the underlying models are closed and evolving (GPT-3.5-turbo is being deprecated), absolute accuracies may drift, so our reproducible claim is the *relative* ordering of mechanisms, not the absolute DSR.

Future work. The most promising next steps follow directly from the limitations above: continuous embedding-based semantic matching for graded eligibility, proactive capability probing at admission time to shrink the fraudster cold-start window, dedicated handling of the honest-newcomer cold-start barrier (provisional trust, time-decayed reputation), and evaluation on open-ended tasks via LLM-as-a-judge.

7 Conclusion

As autonomous agents proliferate across web infrastructure, the mechanisms that govern their competition for visibility will shape the long-term quality, fairness, and resilience of agentic markets. This study proposed a Sponsored Discovery mechanism, built around the Match-gated Agent Composite Score (MACS), to govern how agents gain visibility in the *Web of Open Agents*. Its central insight is that semantic domain match must act as a *multiplicative gate* – not merely an additive weight. In this sense, out-of-domain agents face a structural ceiling that neither a high bid nor accumulated reputation can overcome, while a multi-dimensional behavioral reputation grounds selection in observed accuracy, latency, and cost. Evaluated with real LLM agents on the MMLU-Pro benchmark across honest, domain-falsifying, and cost-inflated scenarios, MACS handled the non-adaptive failure modes that systematically exploited the bid-only and quality-adjusted baselines. Although its robustness remained bounded by calibrated misrepresentation and Sybil whitewashing – limits that call for complementary proactive probing and a persistent-identity mechanism.

References

[1] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 1–22, 2023.

[2] Robert J. Georgio, Chris Forder, Saikat Deb, Anar Rahimov, Patrick Carroll, and Onur Gürcan. Coral protocol: Open infrastructure connecting the internet of agents. arXiv preprint arXiv:2505.00749, 2025.

[3] Andrey Fradkin and Rohit Krishnan. MarketBench: Evaluating AI agents as market participants. arXiv preprint arXiv:2604.23897, 2026.

[4] David M. Rothschild, Markus Mobius, Eleanor W. Dillon, Daniel G. Goldstein, and Jake M. Hofman. The agentic economy. arXiv preprint arXiv:2505.15799, 2025.

[5] Arham Ehtesham, Aditi Singh, Gaurav Kumar Gupta, and Saket Kumar. A survey of agent interoperability protocols: MCP, ACP, A2A, and ANP. arXiv preprint arXiv:2505.02279, 2025.

[6] MIT Media Lab. NANDA: Networked agents and decentralized AI. Technical report, Massachusetts Institute of Technology, 2024.

[7] Agntcy. Agent discovery standard (ADS). Open-source specification, 2025.

[8] Gagan Bansal, Wenyue Hua, Zezhou Huang, Adam Fourney, Amanda Swearngin, et al. Magentic marketplace: An open-source environment for studying agentic markets. arXiv preprint arXiv:2510.25779, 2025.

[9] Xuan Liu, Haoyang Shang, and Haojian Jin. When agent markets arrive. arXiv preprint arXiv:2604.06688, 2026.

[10] Nenad Tomasev, Matija Franklin, Joel Z. Leibo, Julian Jacobs, William A. Cunningham, Iason Gabriel, and Simon Osindero. Virtual agent economies. arXiv preprint arXiv:2509.10147, 2025.

[11] Nawaf A. Murad and Saad Baloch. Governed by agents: A survey on the role of agentic AI in future computing environments. arXiv preprint arXiv:2509.16676, 2025.

[12] Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American Economic Review*, 97(1):242–259, 2007.

[13] Hal R. Varian. Position auctions. *International Journal of Industrial Organization*, 25(6):1163–1178, 2007.

[14] Massimo Paolucci, Takahiro Kawamura, Terry R. Payne, and Katia Sycara. Semantic matching of web services capabilities. In *Proceedings of the 1st International Semantic Web Conference (ISWC)*, pages 333–347, 2002.

[15] David Martin et al. OWL-S: Semantic markup for web services. W3C Member Submission, 2004.

[16] Lei Li and Ian Horrocks. A software framework for matchmaking based on semantic web technology. *International Journal of Electronic Commerce*, 8(4):39–60, 2004.

[17] Michael Wooldridge and Nicholas R. Jennings. Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, 10(2):115–152, 1995.

[18] Kushagra Agrawal and Nisharg Nargund. Neural orchestration for multi-agent systems: A deep learning framework for optimal agent selection in multi-domain task environments. In *Pattern Recognition and Machine Intelligence (PReMI)*, volume 16357 of *Lecture Notes in Computer Science*, pages 513–521. Springer, 2026.

[19] Sepandar D. Kamvar, Mario T. Schlosser, and Hector Garcia-Molina. The EigenTrust algorithm for reputation management in P2P networks. In *Proceedings of the 12th International Conference on World Wide Web (WWW)*, pages 640–651, 2003.

[20] Audun Jøsang and Roslan Ismail. The beta reputation system. In *Proceedings of the 15th Bled Electronic Commerce Conference*, 2002.

[21] Arash Vaziry, Sara R. Garzon, and Axel Küpper. Towards multi-agent economies: Enhancing the A2A protocol with ledger-anchored identities and x402 micropayments for AI agents. arXiv preprint arXiv:2507.19550, 2025.

[22] Yingxuan Yang, Ying Wen, Jun Wang, and Weinan Zhang. Agent exchange: Shaping the future of AI agent economics. arXiv preprint arXiv:2507.03904, 2025.

[23] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. arXiv preprint arXiv:2406.01574, 2024.