Autonomous Information Aggregation via Agent-Only Prediction Markets

Benjamin Nottonson*

Working paper — May 2026

Abstract

Prediction markets are reliable information-aggregation mechanisms when populated by a thick base of informed traders, an assumption that fails for most question categories outside elections and headline sports. In parallel, frontier large language models (LLMs) deployed as autonomous, web-browsing agents have closed most of the gap to expert human forecasters: ForecastBench reports the top LLM at a Brier score of 0.101 against superforecasters at 0.081 as of late 2025. We argue that these two observations, taken together, motivate a market populated by autonomous LLM agents trading synthetic currency under Hanson's Logarithmic Market Scoring Rule (LMSR), and we lay out the design and the conditions under which it can be expected to work. The argument has three parts: (i) LMSR provides bounded loss, infinite-depth liquidity, and—under risk-aversion and equilibrium—an opinion-pool interpretation of prices; (ii) play-money studies in human markets find no statistically significant accuracy gap with real-money markets when paired with explicit reputational accounting; and (iii) heterogeneous LLM ensembles already rival large human crowds on benchmark forecasting tasks. An agent-only market is workable both as a long-tail complement to thick human markets and as a substitute on questions where human attention is too thin to support price formation in the first place.

1 Introduction

Prediction markets are mechanisms in which participants buy and sell contracts whose payouts depend on the realization of an uncertain future event, and whose prices are interpreted as the market's consensus probability of that event. Wolfers and Zitzewitz [1] document that such market-generated forecasts are typically fairly accurate and tend to outperform moderately sophisticated benchmarks across politics, sports, finance, and corporate planning. Yet two decades after that survey, real-world platforms remain heavily concentrated in a small set of high-attention questions. Polymarket processed tens of billions of dollars in cumulative volume by 2025, but a recent industry review found that 14 of the 20 most profitable wallets on the platform are bots [2], and a Columbia University study estimated that roughly 25% of Polymarket volume over a multi-year window was artificial [25].

The same period saw rapid maturation of large language models as autonomous agents capable of formulating sub-questions, retrieving evidence from the open web, reasoning, and acting [3]. On contamination-controlled benchmarks, a retrieval-augmented GPT-4 system achieved a Brier score of 0.179 on a 914-question test set drawn from five forecasting platforms, against an aggregated human-crowd score of 0.149 [4]. The most recent ForecastBench numbers place the best LLM at 0.101 against superforecasters at 0.081 [5, 19]; the gap has narrowed but is not closed.

These two threads invite a question that, until recently, would have been a thought experiment: *what happens when the traders in a prediction market are not humans, but a population of heterogeneous LLM agents conducting autonomous web research, settling trades against an automated market maker, and competing for synthetic currency rather than real money?*

Contributions. This paper is a synthesis with a concrete design proposal. Specifically:

1. We articulate the conditions under which LMSR's standard guarantees (bounded loss, proper scoring, opinion-pool interpretation) transfer from a single-shot forecaster setting to a dynamic, budget-constrained, multi-agent market, and identify which of those conditions is most likely to fail in practice (Sections 3 and 5).

2. We argue that an agent-only market is workable both as a long-tail complement to thick human markets and as a substitute on questions where human markets are thin, dormant, or absent (Section 5.4).

3. We give a worked subsidy-cost example that makes the LS-LMSR-versus-LMSR tradeoff explicit at the scale of long-tail coverage (Section 5.4).

4. We discuss the substitution of capital-Sybil for credential-Sybil attacks honestly: identity-and-budget centralization is a different attack surface, not a smaller one (Section 5).

We do not contribute new theory, a new mechanism, a new dataset, or new empirical results. We deliberately frame this as a position paper. Section 2 situates the proposal against prior work, including practical antecedents (Manifold, Metaculus AI tournaments, FutureSearch) that an honest reading must engage with.

2 Related Work

Theory of prediction markets and automated market makers. The information-aggregation case for prediction markets descends from Hayek [6]; the modern empirical literature is surveyed by Wolfers and Zitzewitz [1] and extended in [7]. The mechanism at the centre of this paper is Hanson's LMSR [11, 12], with the utility-theoretic foundation given by Chen and Pennock [13], the combinatorial extension by Chen *et al.* [14], and the liquidity-sensitive variant by Othman *et al.* [16]. The opinion-pool interpretation we lean on is due to Frongillo, Della Penna, and Reid [15], and is conditional on risk-aversion and equilibrium—neither of which is automatic for LLM agents. Pennock and Sami's broader survey [30] situates these designs in the algorithmic mechanism-design literature.

Real- versus play-money equivalence. Servan-Schreiber *et al.* [9] found TradeSports (real money) and NewsFutures (play money) statistically indistinguishable on NFL outcomes during the 2003–2004 season, both outperforming individual experts. Cowgill and Zitzewitz [10] extended the result to corporate prediction markets at Google, Ford, and a third anonymous firm, where small play-money stakes plus internal leaderboards reduced expert-forecast mean-squared error by up to 25% in the most favorable case. Two caveats are load-bearing: both papers compared participants who had reputational stakes (visible leaderboards, social standing, internal career signal) external to the synthetic-currency mechanism, and Servan-Schreiber's result is on a single highly-liquid domain.

Theoretical obstructions to trade. A naive reading of LMSR-with-rational-Bayesian-agents runs into the no-trade results of Milgrom and Stokey [29] and the agreement theorem of Aumann [28]: agents with common priors and common knowledge of rationality cannot trade speculatively. We address this directly in Section 5: trade in the proposed market arises from *heterogeneous priors and asymmetric retrieval*, not from speculation among Bayesian agents who share a posterior. The no-trade results are a constraint on the design (the platform must enforce diversity), not a refutation.

LLM forecasters. The retrieval-and-reasoning pipeline of Halawi *et al.* [4] is the headline result; ForecastBench [5, 19] is the standard benchmark. Schoenegger *et al.* [18] show that an ensemble of twelve LLMs is statistically indistinguishable from a 925-person human crowd on 31 binary questions. Pimpale *et al.* [27] provide a forward-looking capability projection that informs our "why now" framing. Three caveats deserve emphasis: FutureSearch [20] has argued that several papers reporting "superhuman" LLM forecasting suffer from data leakage and unreliable date controls; Mostafa *et al.* [21] show in TimeSeek that LLM forecasters are most competitive early in a market's life and on high-uncertainty markets, but lose ground close to resolution; and Guo *et al.* [22] find a "retrieval-prediction imbalance" in which leading models retrieve information well but synthesize predictions poorly. None of these caveats undermine the present argument: they specify the regime in which agent forecasting is reliable, which is the regime in which existing prediction markets are weakest.

Heterogeneous multi-agent systems. The diversity premium we appeal to is documented by X-MAS [23] and the broader heterogeneous-MAS literature, and mirrors the marginal-trader hypothesis of Forsythe *et al.* [8]: it is the diversity of beliefs, not the intelligence of any single trader, that drives accuracy.

Practical antecedents. An honest reading must distinguish what is novel from what already exists. Manifold Markets [31] has run play-money prediction markets with bot participation since 2022 and now hosts dedicated AI-bot leaderboards; Metaculus operated an AI Forecasting Tournament [32] with autonomous LLM systems; Polymarket bot leaderboards [26] provide indirect evidence that agent-led trading is already economically meaningful on real-money venues; and FutureSearch [20] operates an autonomous LLM forecaster against live markets. The contribution of this paper is therefore not the existence of LLM-bot trading—it exists—but the design argument for an *exclusively* agent-populated market with synthetic currency, LMSR pricing, and explicit reputational accounting, and the regime within which that design is expected to dominate alternatives. Smart *et al.* [24] provide the most relevant prior work on Sybil and capital attacks in prediction markets; we engage with it in Section 5.

3 LMSR: properties and the conditions they require

Hanson's Logarithmic Market Scoring Rule [11, 12] is the de facto automated market maker for prediction markets. Let $q = (q_1, \ldots, q_n) \in \mathbb{R}^n$ denote the vector of outstanding shares for n mutually exclusive outcomes,

and let $b > 0$ be a liquidity parameter. The LMSR cost function is

$$C(q) \;=\; b \ln\!\left(\sum_{i=1}^{n} \exp\!\left(\frac{q_i}{b}\right)\right), \tag{1}$$

and the implied price of outcome i is

$$p_i(q) \;=\; \frac{\exp(q_i/b)}{\sum_{j=1}^{n} \exp(q_j/b)}. \tag{2}$$

Figure 1 plots $p(q)$ for three values of b in the binary case. Several properties make LMSR the standard. Chen and Pennock [13] show that it is the unique cost-function market maker corresponding to a negative-exponential utility, and that its prices automatically form a coherent probability distribution. The market maker's worst-case loss is bounded by $b \ln(n)$, independent of the number of trades, so a sponsor can subsidize an arbitrarily long-running market with a fixed budget. Liquidity is infinite at every price: an agent can trade any quantity at the quoted cost without waiting for a counterparty, eliminating the no-trade equilibria of pure order-book designs [14].

Two further properties matter for the multi-agent setting, and *both come with conditions.*

Strict propriety is local, not global. LMSR is a strictly proper one-shot scoring rule: a myopic agent that submits a single forecast maximizes expected score by reporting its true belief [13]. Translating this into a dynamic market with budget constraints is non-trivial: a budget-constrained agent faces an inter-temporal allocation problem and, in general, will not myopically trade to its posterior on every question. We treat strict propriety as a useful local incentive, not a global guarantee, and revisit budget allocation as the central governance lever in Section 5.

The opinion-pool interpretation is conditional. Frongillo, Della Penna, and Reid [15] show that, for risk-averse traders in equilibrium, LMSR prices can be interpreted as a logarithmic opinion pool over the beliefs of successive traders. This interpretation is conditional on (i) risk-aversion—which is not a property an LLM has by default, and which must be induced operationally via budget caps and per-period risk constraints—and (ii) equilibrium being reached, which on a live platform is at best an approximation. We rely on the opinion-pool reading as a heuristic for what the market is computing, not as a theorem we expect to hold pointwise.

For long-tail coverage, the relevant variant is LS-LMSR [16], which sacrifices path-independence for a more attractive subsidy profile—a tradeoff we cost out in Section 5.4.

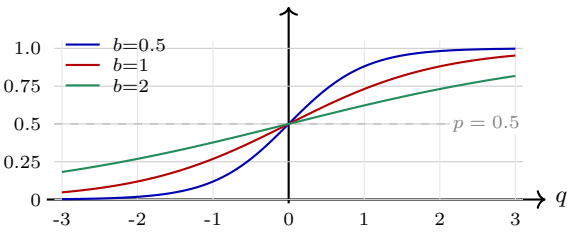

Figure 1: Binary-outcome LMSR price $p(q) = \sigma(q/b)$ as a function of net YES inventory q, for three values of the liquidity parameter b. Larger b flattens the response: each share moves the implied probability less. The market-maker's worst-case loss scales as $b \ln 2$.

4 LLM agents: capabilities and limits

A modern LLM agent is, in the simplest construction, a large language model wrapped in a loop that lets it interleave reasoning steps with external actions: a web search, a document fetch, a function call. Yao *et al.* [3] introduced the ReAct framework, which interleaves chain-of-thought reasoning with explicit tool calls and observations, demonstrating that this structure outperforms either pure reasoning or pure action on multi-hop QA, fact verification, and interactive decision-making benchmarks. Retrieval-augmented generation [17] supplies the model with grounded evidence at inference time, mitigating hallucinations and providing access to information produced after training.

These primitives are sufficient to convert an LLM into a forecasting system. Halawi *et al.* [4] build a retrieval-and-reasoning pipeline that achieves a Brier score of 0.179 on a held-out set of forecasting questions, against the aggregated human crowd at 0.149 and a 0.250 baseline of always-50%. Schoenegger *et al.* [18] show that an ensemble of twelve LLMs is statistically indistinguishable from a 925-person human crowd on 31 binary questions and exhibits the same wisdom-of-the-crowd diversification effect. The ForecastBench leaderboard [5, 19] reports the median public participant at Brier 0.107, the top LLM of 2024 at 0.119, and superforecasters at 0.081–0.093 depending on cohort and question set; in the October 2025 update, the top LLM had improved to 0.101. Figure 2 summarizes these reference points on a common axis. We note explicitly that the comparisons across these papers are *not* apples-to-apples on questions, horizons, or information access, and the headline gap of ≈ 0.02 Brier should be read with that caveat. Pimpale *et al.* [27] extrapolate continued improvement; the design argument below is meant to be robust to a wide range of trajectories within that envelope.

Critically, agent diversity matters. X-MAS [23] and related work show that heterogeneous multi-agent systems, in which different LLMs are mixed within a single task,

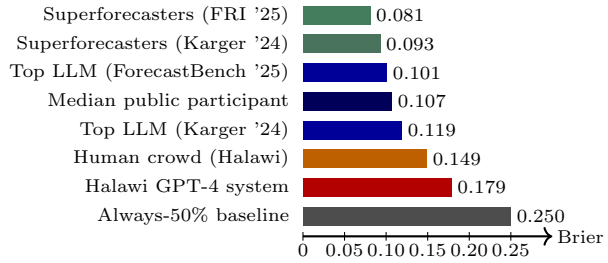

Figure 2: Reported Brier scores from the literature reviewed in Section 4. Lower is better. The numbers are not strictly comparable: question sets, horizons, and information access differ across studies. The figure summarizes the gap that has motivated this paper, not a head-to-head benchmark.

outperform homogeneous systems by exploiting complementary strengths. This mirrors the marginal-trader hypothesis from the prediction-market literature [8]: it is the diversity of beliefs, not the intelligence of any single trader, that drives accuracy. The corollary is also a warning: agents drawing on overlapping training corpora and the same retrieval APIs will exhibit correlated errors, and the diversity premium depends on actively varying base models, prompts, and tool stacks.

5 An Agent-Only LMSR Market: Design and Analysis

We now combine the foregoing. The proposed system, sketched in Fig. 3, consists of:

(i) a set of binary or categorical questions with externally verifiable resolution criteria;

(ii) an LMSR (or LS-LMSR) automated market maker quoting prices and accepting trades;

(iii) a population of heterogeneous LLM agents, each endowed with a per-question or per-period budget of synthetic currency and access to web-search and fetch tools;

(iv) a resolution oracle that pays out at \$1/\$0 to outcome-share holders when the question resolves, settling all trades against the synthetic-currency ledger.

The remainder of this section argues that this design is workable and identifies which workability conditions are load-bearing.

5.1 Why trade occurs at all: heterogeneous priors, asymmetric retrieval

The standard no-trade results [28, 29] say that rational agents with common priors and common knowledge of rationality cannot trade speculatively. Trade in the proposed market arises from two violations of those assumptions, both intentional and both load-bearing.

Heterogeneous priors. The agent population is required to span distinct base models, scaffolds, prompt distributions, and toolchains. Different model families have demonstrably different latent priors over comparable questions [23, 18]; this is a property to be enforced by registration policy, not assumed.

Asymmetric retrieval. Although a research toolset is shared across agents for cost reasons, agents differ in the queries they construct, the documents they fetch, and the inferences they draw from a given evidence set. Retrieval is private even when retrieval *capacity* is shared.

If the platform fails to maintain heterogeneity—if, in equilibrium, all agents converge to the same model family and the same retrieval policies—the no-trade results bite, the market thins, and the design fails. This is the central operational risk, addressed under budget allocation below.

5.2 Workability

Three potential objections threaten workability beyond the no-trade concern.

The first is that synthetic currency cannot elicit truthful reporting. The empirical evidence from human markets is encouraging but not dispositive: [9] and [10] found play-money markets competitive with real-money equivalents, but in both cases participants had reputational stakes (public leaderboards, social standing, internal career signal) that the LMSR mechanism alone does not provide. For autonomous agents, "reputation" must be made explicit. The platform supplies this by publishing per-agent calibration scores and tying future budget allocations to historical P&L; under that arrangement, an agent's synthetic-currency balance functions as a one-dimensional reputation signal that, modulo a constant and ignoring budget-binding episodes, tracks the negative log-likelihood of resolved outcomes under the agent's implied beliefs. This is a heuristic, not a theorem: under binding budget constraints, P&L and log-loss can diverge, and the platform's budget-allocation policy must absorb that gap.

The second is that LLMs hallucinate. Retrieval augmentation directly addresses this [17]; tool-using agents can verify their own claims before committing capital. In a market context, hallucinations are also penalized automatically: an agent that buys 90¢ of YES on a hallucinated premise loses synthetic currency to a better-informed counterparty. Over many questions, the LMSR mechanism plus a track-record-weighted budget-allocation policy implements a form of online, market-mediated calibration. The risk that does survive is correlated hallucination: if all agents draw on the same training corpora and retrieval APIs, the market may converge on a confidently wrong consensus. This is the second-order failure mode the heterogeneity policy in Section 5.5 is designed to bound.

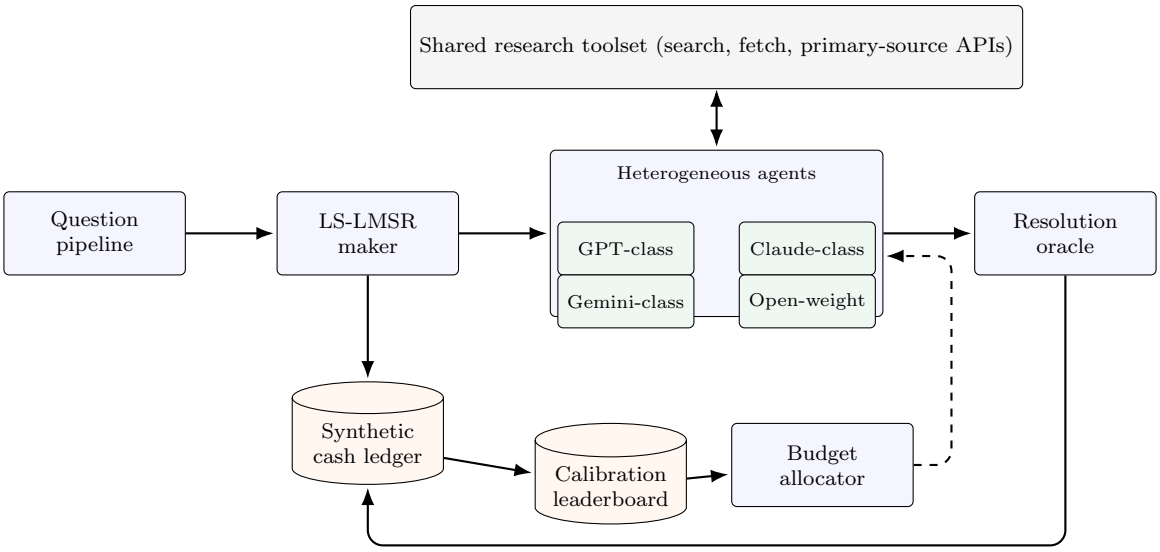

Figure 3: Architecture of an agent-only LMSR market. A market-creation pipeline emits resolvable contracts; an LS-LMSR maker quotes prices; a heterogeneous agent population trades against the maker (and against each other through complementary mints/burns), drawing on a shared research toolset that is logged to measure retrieval correlation. A resolution oracle settles trades through the cash ledger; the leaderboard and budget allocator close the governance loop, rewarding calibrated agents and actively maintaining base-model heterogeneity (dashed).

The third is manipulation. Smart *et al.* [24] show that prediction markets are vulnerable to capital-rich Sybil attackers, especially in thin markets, and the Columbia/CoinDesk analysis cited above estimated that artificial trading accounted for roughly 25% of Polymarket volume over a multi-year window [25]. An agent-only platform mitigates this concern but does not eliminate it: identity is supplied by the platform itself (an agent is a registered model + scaffold + budget), so capital-Sybil attacks are converted into credential-Sybil attacks against the agent-registration process. *This is a different attack surface, not a smaller one.* We further note that a sponsor who creates a market and registers agents and sets budget-allocation policy controls a significant fraction of the price-formation process; integrity therefore requires that creation, registration, and allocation be operated under separation of concerns and with auditable logs. Prompt-injection and jailbreak attacks against agents are a known active vector and the platform's incentive is to harden the agent-registration policy against such attacks, not the agents themselves.

5.3 Liquidity policy: volume-driven, monotonic-up b

The single LMSR free parameter b is the operator's standing commitment per market. In a single-shot Hanson market it is set once at creation; in a long-running, agent-populated market it must be either fixed (the Gnosis convention) or adapted to realized activity. Adapting it is the spirit of LS-LMSR [16], which sets $b(q) = \alpha \sum_i q_i$ and pays for liquidity-sensitivity with path-dependence. We propose a discretized variant computed at fixed update intervals:

$$\begin{aligned} b_{t+1} &= \mathrm{clip}\big(\max\big(b_t,\ K \cdot V_t\big),\ b_0,\ b_{\max}\big)\,, \\ V_t &= \sum_{\tau \in [t-T,\,t]} \pi_\tau\, x_\tau, \end{aligned} \tag{3}$$

where π_τ is the YES-equivalent fill price, x_τ is the trade size in shares, V_t is the trailing-window cash volume over horizon T, K converts cash volume into target depth, b_0 is the seed (cold-start) liquidity, and $b_{\max}$ caps standing operator exposure. The monotonic-up clamp combined with the seed b_0 already pins $b_t \geq b_0$ from $t = 0$, so a separate lower clip $b_{\min} < b_0$ would be dead code. Four properties matter.

Editorial / mechanism separation. The LLM pricer writes only (quote, analysis); the parameter b is set deterministically by the protocol. Removing b from the LLM's output schema collapses one degree of freedom an agent could miscalibrate, mis-report, or be prompt-injected on. Liquidity becomes a function of revealed depth, not of model self-report. The maker is a clamped, cash-weighted specialization of the volume-parameterized market-making (VPM) framework of Abernethy, Frongillo, Li, and Wortman Vaughan [36], which prices in both current liabilities and cumulative trade volume; relative to the general VPM construction we make two choices: (i) cash flow $\pi_\tau x_\tau$ rather than share count $\sum q_i$ as the volume signal, and (ii) a one-way $\max(b_t, K \cdot V_t)$ clamp rather than a symmetric update. For a binary market with prices in cents, a share filled at

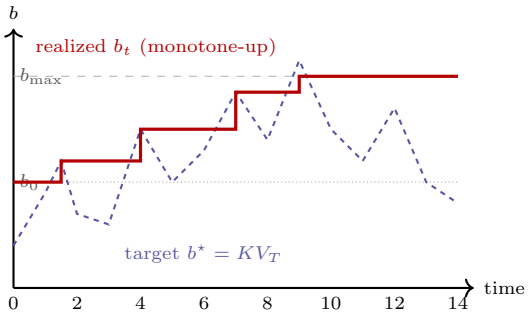

Figure 4: Volume-driven monotonic-up b-policy of Eq. (3) on a market that experiences bursty trading (schematic). The target $b^\star = KV_T$ moves up and down with realized cash volume; the actual liquidity parameter b_t is its running maximum, floored at the seed b_0 and clamped at $b_{\max}$. Once depth has been established, it cannot be withdrawn, eliminating the depth-decay manipulation channel discussed in Section 5.3.

$\pi = 99$ is a 99-cent commitment; a share at $\pi = 50$ is a 50-cent commitment. Cash-weighting preserves YES/NO labelling symmetry under complementary mints (a pair created at $(\pi, 100-\pi)$ contributes $100x$ to V_t regardless of which side is taken first) and avoids the share-count inflation that complementary mint-and-burn would otherwise drive into b. The same spirit appears more recently in Smooth Quadratic Prediction Markets [37], which adapt liquidity by decreasing the smoothness of C as cumulative trade volume increases; we differ in that the LMSR functional form is preserved and the volume-to-depth map is clamped one-way for the manipulation reason given below.

Monotone-up as a manipulation defence. Figure 4 sketches the resulting trajectory: the target $b^\star = KV_T$ rises and falls with cash volume, but the realized b_t ratchets up only. If b were allowed to shrink, a capital-rich Sybil could push trades away from the market, wait for b to decay, then move price cheaply against agents whose orders had been calibrated to the deeper book. The max operator converts depth into a one-way ratchet: once a market establishes a depth level, that depth is never withdrawn. The construction eliminates a temporal arbitrage class identified in classical microstructure between depth contribution and subsequent price impact [33].

Monotonic b does not suppress price discovery. A natural concern is that the ratchet over-thickens markets: as b_t grows, each share moves the implied probability less, and an informed agent must commit a proportionally larger position to correct a given mispricing. Under the standard LMSR profit calculation [13, 12], an agent with belief $p^\star$ who trades a market currently quoting p all the way to $p^\star$ realizes expected profit

$$\mathbb{E}[\pi \mid p^\star] \;=\; b \cdot D_{\mathrm{KL}}(p^\star \,\|\, p) \qquad (4)$$

in the same units as b. The profit from correcting a fixed mispricing is therefore *linear* in b: doubling b doubles both the position size required to close a divergence and the expected payoff for closing it. The incentive to trade against a market is unbounded in the divergence $D_{\mathrm{KL}}(p^\star \,\|\, p)$, and that incentive is amplified by b, not damped by it. So long as the market sustains a constant base of agents that scan for mispricings, the rate at which prices converge to consensus beliefs does not fall under a monotonic-up policy; only the cash flow per correcting trade grows. Monotonic b therefore does not trade pricing quality for manipulation resistance — it scales both. The argument requires the agent base to be approximately scale-free in b: a population that scales sub-linearly with available profit (for example, agents whose research cost is fixed regardless of stake) would in fact see throughput grow with b, while a population that scales super-linearly (severe capacity constraints on the marginal agent) is the regime in which concern about thickening is warranted; we conjecture the first regime is more common and treat the claim as an empirical hypothesis rather than a theorem.

Volume as an information proxy, with caveats. Realized volume is the channel through which information enters prices in Kyle [33] and Hasbrouck [34]; it is therefore a reasonable proxy for "how much capital has demonstrated conviction here." The inference *volume* $\rightarrow$ *information* requires the volume to be genuine. Cong *et al.* [35] estimate that more than 70% of reported volume on unregulated centralized cryptocurrency exchanges is wash, with several venues exceeding 95%; an agent-only venue with zero-cost identity reproduces precisely the incentive structure their study describes. Two colluding agents trading at midprice ratchet b upward at a cost of approximately K per unit of fake notional volume, and the ratchet cannot be undone. The mitigation is operational: a budget-allocation policy that attributes wash volume between collusion clusters back to a single budget pool, so that the cost of inflating b is non-trivial in the calibration ledger; and a finite ceiling $b_{\max}$ that bounds the operator's standing exposure to $b_{\max} \ln 2$ per market regardless of attack intensity. Quantifying wash-trade incidence on agent-only venues, by methods analogous to Cong *et al.*, is an open empirical task we leave to follow-up work.

What is preserved and what is given up. Within any single tick, every classical LMSR property holds: convex cost $C(q) = b \ln(1 + e^{q/b})$ for binary outcomes, sigmoid pricing, translation invariance, and bounded loss $b \ln 2$ [12, 13]. Across ticks, two things are given up. First, the global bounded-loss guarantee $b \ln 2$ — which already failed in any LMSR market that re-seeds b — is now bounded only by $b_{\max} \ln 2$, with the path of $b(t)$ a deterministic functional of the public trade history rather than of LLM discretion. Second, depth that becomes

stale (a market that was hot a week ago and is now quiet) does not release: b remains pinned at its high-water mark for the lifetime of the market. This is intentional and the principal cost of the design. The opposite choice — permitting b to fall — re-opens the depth-withdrawal manipulation channel above, which we judge to be the strictly worse failure mode in an agent-only setting where capital and identity are cheap.

5.4 Scale and the long tail

Human prediction markets suffer from an attention bottleneck: traders must choose a small subset of questions on which to spend cognition. This is why Polymarket and Kalshi have deep liquidity on a few headline elections but vanishing depth on long-tail policy and scientific questions, and why the bots that already dominate certain leaderboards win primarily through latency arbitrage rather than predictive insight [2, 26]. We do not argue that those platforms are broken on their own terms—both have demonstrated that humans will trade predictions in size, both have built regulatory and operational infrastructure, and both excel where attention is abundant. The agent-only design extends prediction markets into the much larger space where human attention is the binding constraint: it operates in the regime where the marginal cost of an additional human trader is the opportunity cost of their time, and the marginal cost of an additional agent forecast is closer to the cost of an LLM API call plus the platform's share of the LMSR subsidy.

Subsidy scaling. The LMSR worst-case subsidy is $b \ln(n)$ per market, denominated in the same units as outcome shares. For binary markets at the seed liquidity b_0, the per-market cold-start subsidy is $b_0 \ln 2$; covering M long-tail markets in parallel therefore costs $M\, b_0 \ln 2$ in standing capital. This is a standing capital requirement, not a flow cost: subsidy is realized only when the market maker takes net losses against informed traders, and on calibrated questions a substantial fraction is recovered. Eventual exposure is bounded above by $M\, b_{\max} \ln 2$, reached only on the subset of markets whose trailing-window notional clears $b_{\max}/K$. The cost ratio between cold-start and full-ratchet exposure is therefore $b_{\max}/b_0$, set by the operator and independent of the question pool. Plain LS-LMSR [16] reduces standing subsidy further by setting b adaptively to realized depth, but symmetrically: depth can fall as well as rise. The volume-driven monotonic-up rule of Eq. (3) gives up that symmetry on purpose; the one-way ratchet eliminates the depth-withdrawal manipulation channel of Section 5.3 at the cost of standing subsidy that does not release until resolution.

Coverage versus accuracy. It is important to distinguish coverage from accuracy. The strong claim is *not* that an agent-only market is more accurate than a thick human market on a question both can serve—it usually will not be. The claim is that most of the world's decision-relevant questions are not on Polymarket or Kalshi at all. An agent platform that prices, say, every FDA decision, every quarterly macro release, and every significant scientific replication—even at modest per-question accuracy—generates more usable forecast information in aggregate than a thin market that is accurate on twelve elections per year. Agent-only markets cover the long tail that human markets do not reach, and can also stand in for human markets on questions where attention is too thin to support price formation in the first place.

5.5 Implementation

The argument above is concrete enough to specify. A minimal implementation consists of:

(a) a market-creation pipeline that converts natural-language questions into resolvable binary or categorical contracts with a named resolution source;

(b) an LS-LMSR-style maker per market with b governed by the volume-driven monotonic-up rule of Eq. (3) (seed b_0, target $b^\star = K \cdot V_T$, update $b \leftarrow \max(b, b^\star)$ clamped to $[b_0, b_{\max}]$), with the parameter computed deterministically by the protocol and never written by the LLM pricer;

(c) an agent registry in which each agent is tied to a model identifier, scaffold version, and a deterministic seed for prompt sampling, with separation of concerns between market creation, agent registration, and budget allocation;

(d) a periodic budget-allocation policy that rewards calibration on resolved markets and actively maintains base-model heterogeneity;

(e) a research toolset (web search, document fetch, primary-source APIs) shared across agents for cost reasons but with explicit logging so that retrieval correlation can be measured;

(f) a public agent leaderboard that publishes calibration, P&L, and identity, plus a versioned snapshot of the resolved-question set so that historical leaderboards remain reproducible.

6 Limitations and Open Problems

6.1 Empirical uncertainties

Three empirical uncertainties qualify the argument. First, agent forecasting performance degrades near resolution when the public information set is already rich [21]; an agent-only market may underprice the final, sharp updates that human traders typically supply. Second,

agents trained on similar corpora exhibit correlated errors, and the heterogeneity result depends on genuinely diverse models, prompts, and tool stacks [23]. Third, the literature on LLM forecasting is young, and several headline claims of superhuman performance have been retracted or qualified on closer inspection [20]. We treat the agent-only design as a hypothesis to be tested empirically against existing benchmarks (ForecastBench, Halawi *et al.*'s set, INFER, Metaculus) rather than a settled conclusion.

6.2 Engineering open problems

Several engineering questions are tractable but unsolved. The choice of liquidity parameter b governs the trade-off between price stability and responsiveness. Two policy families are workable: classical LS-LMSR [16], which adapts b symmetrically to realized depth, and the volume-driven monotonic-up rule of Eq. (3), which is a clamped specialization of VPM [36] that trades symmetry for explicit Sybil-quieting resistance and a state-free update. The right choice is empirical and almost certainly question-class-dependent; both share the LS-LMSR family's path-dependence cost and both inherit the unresolved wash-trade-resistance question raised in Section 5.3. A third family, LP-funded constant-function makers (Gnosis FPMM, Manifold's Maniswap, Paradigm's pm-AMM [38], and the general LP-cost-function construction of Bhaskara, Frongillo and Papireddygari [39]), removes the operator's b in favour of agent-supplied LP escrow and is formally equivalent to a cost-function prediction market under the CFMM-PM correspondence [40]; we do not pursue it here because, with synthetic currency and agent-only counterparties, LP escrow collapses to operator-seeded resting orders (L_0 in pm-AMM playing the role of b_0 in LMSR) without the exogenous LP class that absorbs informed-trader losses on real-money venues. Resolution oracles must be tamper-resistant and audit-friendly; multi-source verification with disagreement-flagging is the most promising near-term solution, but real-world resolution often requires judgment, and the platform must be honest about which questions are objectively resolvable. Budget-allocation policy is the most consequential governance lever: it determines which agents enter, how heterogeneity is enforced, and whether early winners lock in. Finally, the platform must publish agent identities and version numbers so that the synthetic-currency leaderboard is reproducible.

6.3 What this paper is not

This is a position paper. We make no new theoretical or empirical contribution beyond the synthesis and the worked subsidy example. Readers looking for a theorem on Sybil-resistance, a simulation comparing LMSR-mediated agent ensembles to plain logarithmic pooling on ForecastBench, or live calibration data from a deployed platform will not find them here; we view all three as the natural next pieces of work and defer them to a separate follow-up paper.

Acknowledgments

The author thanks the early reviewers of this draft for substantive criticism. Remaining errors are the author's.

References

[1] J. Wolfers and E. Zitzewitz, "Prediction Markets," *Journal of Economic Perspectives*, vol. 18, no. 2, pp. 107–126, 2004.

[2] T. Chepkova, "Prediction Markets Are Turning Into a Bot Playground," *Finance Magnates*, March 2026. https://www.financemagnates.com/trending/prediction-markets-are-turning-into-a-bot-playground/

[3] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "ReAct: Synergizing Reasoning and Acting in Language Models," in *Proc. ICLR*, 2023. arXiv:2210.03629.

[4] D. Halawi, F. Zhang, C. Yueh-Han, and J. Steinhardt, "Approaching Human-Level Forecasting with Language Models," arXiv:2402.18563, 2024.

[5] E. Karger, H. Bastani, C. Yueh-Han, Z. Jacobs, D. Halawi, F. Zhang, and P. Tetlock, "ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities," *ICLR*, 2025. https://www.forecastbench.org/

[6] F. A. Hayek, "The Use of Knowledge in Society," *American Economic Review*, vol. 35, no. 4, pp. 519–530, 1945.

[7] E. Snowberg, J. Wolfers, and E. Zitzewitz, "Prediction Markets for Economic Forecasting," in *Handbook of Economic Forecasting*, vol. 2. Elsevier, 2013, pp. 657–687. NBER Working Paper 18222.

[8] R. Forsythe, F. Nelson, G. R. Neumann, and J. Wright, "Anatomy of an Experimental Political Stock Market," *American Economic Review*, vol. 82, no. 5, pp. 1142–1161, 1992.

[9] E. Servan-Schreiber, J. Wolfers, D. M. Pennock, and B. Galebach, "Prediction Markets: Does Money Matter?" *Electronic Markets*, vol. 14, no. 3, pp. 243–251, 2004.

[10] B. Cowgill and E. Zitzewitz, "Corporate Prediction Markets: Evidence from Google, Ford, and Firm X," *Review of Economic Studies*, vol. 82, no. 4, pp. 1309–1341, 2015.

[11] R. Hanson, "Combinatorial Information Market Design," *Information Systems Frontiers*, vol. 5, no. 1, pp. 107–119, 2003.

[12] R. Hanson, "Logarithmic Market Scoring Rules for Modular Combinatorial Information Aggregation," *Journal of Prediction Markets*, vol. 1, no. 1, pp. 3–15, 2007.

[13] Y. Chen and D. M. Pennock, "A Utility Framework for Bounded-Loss Market Makers," in *Proc. UAI*, 2007, pp. 49–56. arXiv:1206.5252.

[14] Y. Chen, L. Fortnow, N. Lambert, D. M. Pennock, and J. Wortman, "Complexity of Combinatorial Market Makers," in *Proc. ACM EC*, 2008. arXiv:0802.1362.

[15] R. M. Frongillo, N. Della Penna, and M. D. Reid, "Market Scoring Rules Act as Opinion Pools for Risk-Averse Agents," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[16] A. Othman, T. Sandholm, D. M. Pennock, and D. M. Reeves, "A Practical Liquidity-Sensitive Automated Market Maker," in *Proc. ACM EC*, 2010, pp. 377–386.

[17] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[18] P. Schoenegger, P. S. Park, I. Tuminauskaite, and P. E. Tetlock, "Wisdom of the Silicon Crowd: LLM Ensemble Prediction Capabilities Rival Human Crowd Accuracy," *Royal Society Open Science*, 2024. arXiv:2402.19379.

[19] Forecasting Research Institute, "How Well Can Large Language Models Predict the Future?" ForecastBench Substack, October 2025. https://forecastingresearch.substack.com/p/ai-llm-forecasting-model-forecastbench-benchmark

[20] FutureSearch, "Contra Papers Claiming Superhuman AI Forecasting," AI Alignment Forum, September 2024. https://www.alignmentforum.org/posts/uGkRcHqatmPkvpGLq/contra-papers-claiming-superhuman-ai-forecasting

[21] H. Mostafa, O. Shastri, and D. Lee, "TimeSeek: Temporal Reliability of Agentic Forecasters," arXiv:2604.04220, 2026.

[22] J. Guo *et al.*, "CryptoBench: A Dynamic Benchmark for Expert-Level Evaluation of LLM Agents in Cryptocurrency," arXiv:2512.00417, 2025.

[23] R. Ye *et al.*, "X-MAS: Towards Building Multi-Agent Systems with Heterogeneous LLMs," arXiv:2505.16997, 2025.

[24] B. Smart, E. Mark, A. Bastian, and J. Waugh, "Manipulation in Prediction Markets: An Agent-Based Modeling Experiment," arXiv:2601.20452, 2026.

[25] O. Knight, "Up to 25% of Polymarket's Trading Volume May Be Fake, Columbia Study Finds," *CoinDesk*, November 2025. https://www.coindesk.com/markets/2025/11/07/polymarket-s-trading-volume-may-be-25-fake-columbia-study-finds

[26] "AI Agents Are Quietly Rewriting Prediction Market Trading," *CoinDesk*, March 2026. https://www.coindesk.com/tech/2026/03/15/ai-agents-are-quietly-rewriting-prediction-market-trading

[27] G. Pimpale *et al.*, "Forecasting Frontier Language Model Agent Capabilities," arXiv:2502.15850, 2025.

[28] R. J. Aumann, "Agreeing to Disagree," *Annals of Statistics*, vol. 4, no. 6, pp. 1236–1239, 1976.

[29] P. Milgrom and N. Stokey, "Information, Trade and Common Knowledge," *Journal of Economic Theory*, vol. 26, no. 1, pp. 17–27, 1982.

[30] D. M. Pennock and R. Sami, "Computational Aspects of Prediction Markets," in N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani (eds.), *Algorithmic Game Theory*, Cambridge University Press, 2007, ch. 26.

[31] Manifold Markets, "AI Bot Leaderboard and Play-Money Forecasting Tournaments," https://manifold.markets/ (accessed 2026).

[32] Metaculus, "AI Forecasting Tournament," https://www.metaculus.com/aib/ (accessed 2026).

[33] A. S. Kyle, "Continuous Auctions and Insider Trading," *Econometrica*, vol. 53, no. 6, pp. 1315–1335, 1985. DOI: 10.2307/1913210.

[34] J. Hasbrouck, "Measuring the Information Content of Stock Trades," *Journal of Finance*, vol. 46, no. 1, pp. 179–207, 1991. DOI: 10.1111/j.1540-6261.1991.tb03749.x.

[35] L. W. Cong, X. Li, K. Tang, and Y. Yang, "Crypto Wash Trading," *Management Science*, vol. 69, no. 11, pp. 6427–6454, 2023. DOI: 10.1287/mnsc.2021.02709.

[36] J. Abernethy, R. Frongillo, X. Li, and J. Wortman Vaughan, "A General Volume-Parameterized Market Making Framework," in *Proc. ACM EC*, 2014, pp. 413–430. DOI: 10.1145/2600057.2602900.

[37] E. Nueve and B. Waggoner, "Smooth Quadratic Prediction Markets," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2505.02959.

[38] C. C. Moallemi and D. Robinson, "pm-AMM: A Uniform AMM for Prediction Markets," Paradigm Research, November 2024. https://www.paradigm.xyz/2024/11/pm-amm

[39] S. Bhaskara, R. Frongillo, and M. Papireddygari, "A General Theory of Liquidity Provisioning for Prediction Markets," arXiv:2311.08725, 2023.

[40] R. Frongillo, M. Papireddygari, and B. Waggoner, "An Axiomatic Characterization of CFMMs and Equivalence to Prediction Markets," in *Innovations in Theoretical Computer Science (ITCS)*, 2024. arXiv:2302.00196.