Market-Bench: Benchmarking Large Language Models on Economic and Trade Competition

Yushuo Zheng[1,2], Huiyu Duan[1,*], Zicheng Zhang[1,2], Yucheng Zhu[1], Xiongkuo Min[1,*], Guangtao Zhai[1,2,*]

[1]Shanghai Jiao Tong University, [2]Shanghai Artificial Intelligence Laboratory

{yushuozheng, huiyuduan, zzc1998, zyc420, minxiongkuo, zhaiguangtao}@sjtu.edu.cn

[*]Corresponding author.

https://github.com/aiben-ch/Market-Bench

Abstract

The ability of large language models (LLMs) to manage and acquire economic resources remains unclear. In this paper, we introduce **Market-Bench**, a comprehensive benchmark that evaluates the capabilities of LLMs in economically-relevant tasks through economic and trade competition. Specifically, we construct a configurable multi-agent supply chain economic model where LLMs act as retailer agents responsible for procuring and retailing merchandise. In the **procurement** stage, LLMs bid for limited inventory in budget-constrained auctions. In the **retail** stage, LLMs set retail prices, generate marketing slogans, and provide them to buyers through a role-based attention mechanism for purchase. Market-Bench logs complete trajectories of bids, prices, slogans, sales, and balance-sheet states, enabling automatic evaluation with economic, operational, and semantic metrics. Benchmarking on 20 open- and closed-source LLM agents reveals significant performance disparities and winner-take-most phenomenon, *i.e.*, only a small subset of LLM retailers can consistently achieve capital appreciation, while many hover around the break-even point despite similar semantic matching scores. Market-Bench provides a reproducible testbed for studying how LLMs interact in competitive markets.

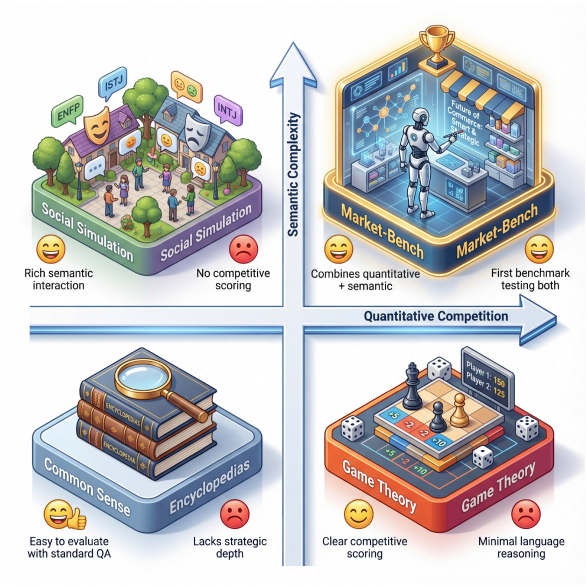

Figure 1: Existing LLM benchmarks focus on either semantic complexity or quantitative competition, but rarely both simultaneously under economic scarcity. Thus, we propose Market-Bench, coupling marketing slogans and operations to jointly evaluate mathematical optimization and language comprehension.

1 Introduction

The advancement of large language models (LLMs) has driven the application of AI in retail from passive analytics to active supply chain management and personalized marketing (Chui et al., 2023; Statista, 2024), requiring both the *computational precision* to manage scarce resources and the *semantic flexibility* to construct persuasive narratives. Marketing language is a decisive economic variable that stimulates purchase intentions by arousing consumers' latent motivations (Keller, 2003; Kohli et al., 2007), and recent work has begun to formalize the aesthetic evaluation of commercial imagery (Ji et al., 2026). However, current LLM benchmarks largely evaluate these capabilities in isolation. As shown in Figure 1, existing frameworks typically focus either on abstract strategic rationality ("Homo Economicus") or open-ended social role-play ("Homo Loquens"), rarely demanding both simultaneously. Social simulations (Park et al., 2023; Piao et al., 2025) model free-form dialogue but impose no hard economic constraints such as scarcity or bankruptcy, while economic simulators (Dwarakanath et al., 2024; Mao et al., 2023) test rigorous quantitative logic but abstract away linguistic complexity. Although LLM evaluation has expanded rapidly across visual quality assessment (Duan et al., 2025, 2022, 2023; Zhang et al., 2026a, 2025b; Jin et al., 2025; Wang et al., 2025),

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 39893–39906

multimodal reasoning (Zhang et al., 2025a, 2026b; Yang et al., 2025; Liu et al., 2024; Zheng et al., 2026, 2025a), and domain-specific fields such as medicine (Ji et al., 2025a,c,b, 2024; Wang et al., 2026), no existing benchmark unifies multi-turn buyer–seller interactions across both numeric and semantic dimensions under strict constraintsthat is, a testbed for *dual-process reasoning* within a closed-loop economy remains absent.

This raises a fundamental question: *Can contemporary LLMs effectively perform quantitative reasoning and semantic adaptation to survive in markets characterized by scarcity and competition?* The market domain is uniquely challenging because optimal pricing and procurement fail without persuasive marketing, while strong marketing fails if budgeting or inventory is mismanaged,a strict co-dependency absent from benchmarks that test either capability alone. Moreover, agents must infer hidden buyer preferences from limited numerical signals under multi-turn competitive dynamics, where errors compound into bankruptcy rather than minor accuracy drops.

We introduce **Market-Bench**, a configurable multi-agent supply-chain economy formulated as a Partially Observable Markov Game with financial and physical constraints. At each step, agents receive structured text observations,private funds and inventory, auction parameters, and public market history,and output strictly formatted JSON combining discrete mathematical actions (procurement bids, retail prices) with a free-form marketing slogan. Crucially, we propose **Persona-Gated Attention** (PGA), a mechanism grounded in consumer "consideration set" theory (Keller, 2003) that operationalizes free-form language as a computable, economically consequential variable. Unlike subjective LLM-as-judge heuristics, PGA provides a strict mathematical gatekeeper: it dynamically computes semantic alignment between generated slogans and hidden buyer personas to determine market visibility, directly linking language generation to financial outcomes.

Our contributions are as follows:

- We define **Market-Bench**, a closed-loop economic environment that enforces hard scarcity (finite funds, bankruptcy) and competitive exclusion.
- We propose **Persona-Gated Attention**, a mechanism that operationalizes the economic value of language, requiring agents to optimize semantic similarity to latent buyer personas to secure market access.
- We provide a reproducible benchmark implementation and report results across 20 LLM agents using automatically computed economic, supply-chain, and semantic metrics.

Work	Competitive Interaction	Hard Scarcity	Closed-Loop Economy	Free-Form Language	Text→ Payoff
Agent Societies					
(Park et al., 2023)	✗	✗	✗	✓	✗
(Zhou et al., 2024)	✗	✗	✗	✓	✗
(Piao et al., 2025)	✗	✗	✗	✓	✗
Strategic Games					
(Zhu et al., 2025)	✓	✗	✗	✓	✗
(Chen et al., 2025)	✓	✗	✗	✓	✗
(Mao et al., 2023)	✓	✓	✗	✓	✗
(Deng et al., 2024)	✓	✗	✗	✓	✓
Economic Simulators					
(Dwarakanath et al., 2024)	✗	✗	✓	✗	✗
Market-Bench	✓	✓	✓	✓	✓

Table 1: Comparison of representative LLM-agent benchmarks and economic simulators. **Market-Bench** is the only benchmark providing both a closed-loop multi-agent market under hard scarcity and free-form language that directly affects payoff via Persona-Gated Attention.

2 Related Work

2.1 LLM-Based Agent Societies

Generative agents have shifted evaluation from static QA to dynamic sandboxes. Park et al. (2023) showed LLMs can simulate believable social behavior, and recent frameworks like *AgentSociety* (Piao et al., 2025) and *MultiAgentBench* (Zhu et al., 2025) have scaled these simulations to urban environments and diverse tasks. However, these benchmarks primarily assess social believability or cooperative completion, rarely imposing the "scarcity conditions" (finite funds, bankruptcy) characteristic of real economies. **Market-Bench** bridges this gap by enforcing strict financial survival ($Funds_i < 0 \Rightarrow Bankruptcy$) within a competitive oligopoly.

2.2 Strategic Reasoning and Game Theory

A distinct line of research evaluates LLMs as "Homo Economicus", testing their rationality in classic game-theoretic setups. Chen et al. (2025) and Mao et al. (2023) subject agents to matrix games like Prisoner's Dilemma (Axelrod, 1984) or Trust Game (Berg et al., 1995) to measure cooperation, deception, and Nash equilibrium convergence. Similarly, Deng et al. (2024) explores bilateral bargaining, focusing on whether LLMs can

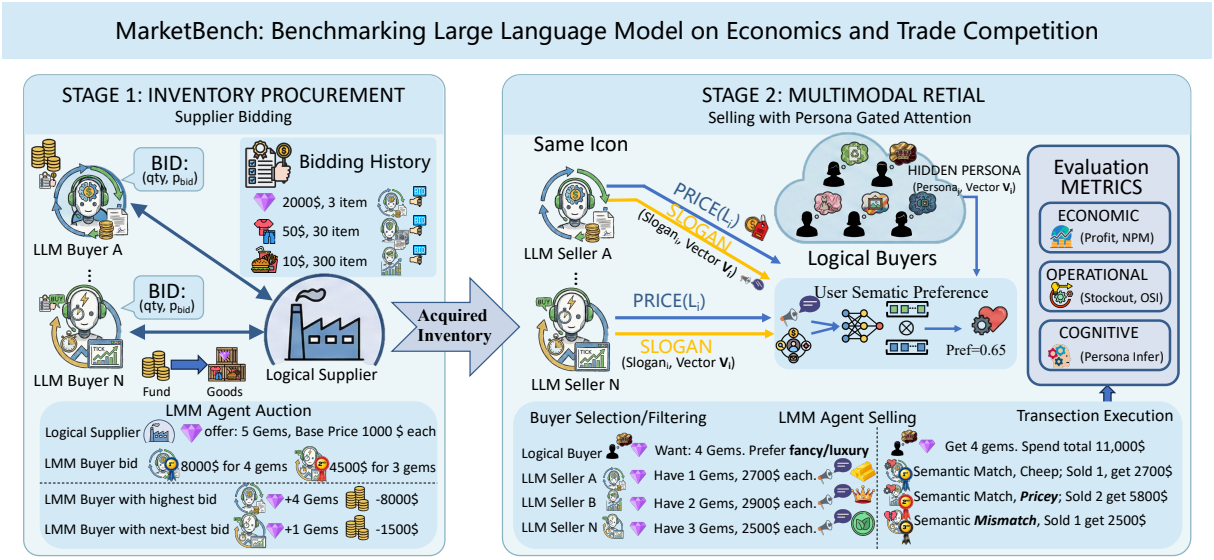

Figure 2: Overview of the Market-Bench environment. Agents operate in a competitive supply chain economy, making procurement, pricing, and marketing decisions under scarcity conditions. The Persona-Gated Attention mechanism requires agents to align their marketing slogans with latent buyer personas to convert demand into sales. Performance is evaluated across economic, supply chain, and cognitive metrics.

reach Pareto-optimal deals. Zheng et al. (2025b) further demonstrate the effectiveness of competitive game settings for revealing capability gaps among multimodal models. While these benchmarks test strategic logic, they often abstract away the linguistic complexity of commerce, reducing interaction to selecting a move (Cooper/Defect) or outputting a number. **Market-Bench** addresses this limitation by integrating free-form language as a first-class action, where agents must generate persuasive slogans to unlock market access through our novel **Persona-Gated Attention** mechanism.

2.3 Supply Chain and Business Optimization

LLMs are increasingly deployed as decision-support tools for supply chain management (Simchi-Levi et al., 2025; Dwarakanath et al., 2024). Dwarakanath et al. (2024) introduce *ABIDES-Economist* for macroeconomic simulation, and Anthropic's Project Vend (Anthropic, 2025b) deployed Claude to autonomously run a retail shop in a single-agent setting. These systems generally position the LLM as a backend optimizer rather than an autonomous agent surviving competitive pressure. **Market-Bench** unifies procurement, pricing, and marketing into a single scalable loop under active multi-agent competition.

3 Market-Bench

Market-Bench is a competitive supply-chain economy designed to test *dual-process* agent behavior in a single closed loop: agents must (i) optimize numeric decisions under budget and inventory constraints, and (ii) generate natural language marketing that determines which consumers can even *see* them. Crucially, this mechanism turns the simulation into both an evaluation environment and a *generative dataset*: each episode yields structured trajectories of observations, numeric actions, text actions, and economic outcomes, enabling fine-grained analysis beyond static QA-style benchmarks. Figure 2 summarizes the lifecycle.

3.1 Economic Setting and State Variables

We model a two-sided market with an upstream supplier and downstream retailers. Let

$$\mathcal{A} = \{A_1, \ldots, A_m\} \tag{1}$$

be the set of retailer agents, and let $\mathcal{X}$ denote the set of items. At step t, the environment state includes each agent's funds $Funds_i(t)$ and inventory $Inv_{i,x}(t)$ for every $x \in \mathcal{X}$. The supplier publishes an offer list

$$O_S(t) = \{(x, Q_x(t), P_{\text{base}}(x))\}_{x \in \mathcal{X}}, \tag{2}$$

where $Q_x(t)$ is available supply and $P_{\text{base}}(x)$ is a reserve (base) price. We also generate a set of

buyers $\mathcal{B}_t$ each step; each buyer B_j has a latent persona text $Persona_j$ and a buyer patience coefficient $\rho_j \in [0, 1]$. Importantly, personas are *not* revealed to agents: retailers must infer demand preferences only through public market outcomes, such as what sold and at what prices, which induces an incomplete-information setting.

3.2 Stage A: Procurement as a Multi-Unit Auction

Procurement implements a per-item, multi-unit, first-price auction with a reserve price. We deliberately choose first-price over second-price auctions because the latter reduces optimal policy to truthful bidding, whereas first-price auctions require bid shading, explicit competition modeling, and profit–win-rate tradeoffs,precisely the strategic reasoning we aim to evaluate. Each agent submits bids of the form

$$b_{i,x}(t) = (q_{i,x}(t), p^{\text{bid}}_{i,x}(t)) \tag{3}$$

where $q_{i,x}$ is the requested quantity and $p^{\text{bid}}_{i,x}$ is the bid price. Agents face a hard budget constraint (invalid bids receive zero allocation):

$$\sum_{x \in \mathcal{X}} q_{i,x}(t)\, p^{\text{bid}}_{i,x}(t) \;\le\; Funds_i(t). \tag{4}$$

The supplier allocates the available quantity of each item to the highest bids above the reserve price and charges winners their bid prices, updating inventories and funds.

3.3 Stage B: Retail as Price Competition with Persona-Gated Attention

After procurement, each agent chooses (i) a retail price $P_{i,x}(t)$ for any subset of items, and (ii) a short slogan $Slogan_i(t)$. The downstream market then matches buyers to retailers via a two-stage choice model: *attention* (who is visible) followed by *purchase* (who is cheapest).

Persona-gated attention (consideration gate). For each buyer B_j, we embed both the slogan and persona with an embedding function $\mathbf{E}(\cdot)$ and compute cosine similarity

$$\begin{aligned} &\text{Sim}(i, j) \\ &= \cos\big(\mathbf{E}(Slogan_i(t)), \mathbf{E}(Persona_j)\big) \end{aligned} \tag{5}$$

Buyers are sampled from a mixture of persona "tribes", including thrift, ethics, hype, and quality, each of which determines a persona template

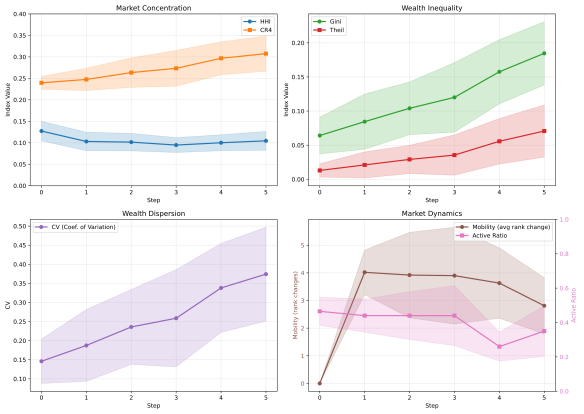

Figure 3: Market-level indices computed from logged trajectories in the default setting. Inequality increases (Gini, Theil, CV); the top-4 share rises (CR4) while concentration remains competitive (HHI); and the Active Ratio declines.

and a slogan sensitivity $\lambda_j \ge 0$. Attention weights are defined as

$$w_{i,j} = \exp\big(\lambda_j\, Sim(i, j)/\tau\big), \tag{6}$$

with temperature $\tau > 0$. The buyer then samples a consideration set $V_j(t) \subseteq \mathcal{A}$ in proportion to $w_{i,j}$; the set size is controlled by the buyer's patience coefficient ρ_j and an upper bound $K_{\max}$. Even a low-price seller cannot sell if it is not sampled into $V_j(t)$.

Purchasing rule (price competition under constraints). At each step, buyers arrive with item demands under a controllable scarcity level (set by scaling aggregate demand relative to supply). For each desired item, the buyer considers only sellers in $V_j(t)$ who posted a price and have inventory, and purchases from the lowest-priced available seller until demand is met or inventory is exhausted. This creates Bertrand-style pressure on pricing within each buyer's attention set, while persona alignment controls which sellers compete for each buyer.

3.4 Information Structure and Action Space

At each step, agents observe their funds and inventory, the supplier offer list, and a compact public market history (posted prices, slogans, and realized sales) to support persona inference. They output structured numeric actions (item-level bids and retail prices) plus a free-form slogan, making MarketBench a tightly coupled numeric-language decision problem.

Mode Type	Model	Economics				Operation				Cognitive
		NPM↑	Pi↑	RAR↑	IEI↑	Stockout Rate↓	Bid Eff.↑	OSI↑	Fill Rate↑	MMS↑
Closed Source Model	Gemini 2.5 Pro (Gemini Team, 2025)	0.167	36589	2.751	0.861	0.765	0.792	0.777	0.221	0.691
	Gemini 2.5 Flash (Gemini Team, 2025)	0.190	26104	1.569	0.645	0.791	0.795	0.820	0.194	0.675
	O3 (El-Kishky et al., 2025)	0.097	13368	1.016	0.901	0.573	0.699	1.000	0.418	0.690
	Sonnet 4.5 (Anthropic, 2025a)	0.158	10619	1.438	0.780	0.892	0.411	0.603	0.097	0.671
	GPT-4o (OpenAI, 2024)	0.117	7619	0.712	0.742	0.911	0.450	0.640	0.079	0.669
Open Source Model	Phi-4 (Abdin et al., 2024)	0.110	7565	0.822	0.867	0.898	0.386	0.636	0.091	0.677
	Qwen2.5 VL 72B (Team, 2025a)	0.058	3402	0.738	0.789	0.942	0.318	0.707	0.050	0.673
	Llama 3.1 70B (Team, 2024a)	0.068	2444	0.628	0.875	0.959	0.195	0.579	0.035	0.673
	QwenLong L1 32B (Wan et al., 2025)	0.085	2242	0.317	0.561	0.976	0.115	0.550	0.020	0.681
	Qwen2.5 32B (Team, 2024b)	0.053	1815	0.562	0.856	0.964	0.189	0.582	0.031	0.660
	Gemma 3 27B (Gemma Team, 2025)	0.041	1481	0.580	0.862	0.949	0.296	0.732	0.043	0.680
	Qwen2.5 VL 32B (Team, 2025a)	0.069	1409	0.656	0.609	0.967	0.197	0.540	0.026	0.639
	ERNIE 4.5 300B (Baidu ERNIE Team, 2025)	0.037	1360	0.614	0.849	0.942	0.418	0.741	0.050	0.685
	InternLM2.5 20B (Cai et al., 2024)	0.031	1154	0.440	0.843	0.959	0.441	0.801	0.035	0.675
	InternLM3 8B (Cai et al., 2024)	0.028	1022	0.318	0.845	0.963	0.280	0.659	0.031	0.660
	DeepSeek V3.2 (DeepSeek-AI, 2025)	0.072	616	0.292	0.796	0.985	0.237	0.829	0.013	0.685
	Qwen2.5 72B (Team, 2024b)	0.021	335	0.309	0.759	0.973	0.140	0.554	0.022	0.674
	Hunyuan A13B (Tencent Hunyuan Team, 2024)	0.014	298	0.134	0.941	0.975	0.152	0.591	0.020	0.683
	Qwen3 30B-A3B (Team, 2025b)	0.000	0	0.000	0.000	1.000	0.000	0.796	0.000	0.677
	ERNIE 4.5 21B (Baidu ERNIE Team, 2025)	-0.004	-58	-0.039	0.547	0.985	0.216	0.571	0.015	0.659

Table 2: Performance Metrics of Different Models. For each metric, an arrow indicates whether higher (↑) or lower (↓) values are better. The best and second-best performances for each metric are highlighted in red and blue, respectively.

3.5 Agent Objective and Economic Trade-offs

Market-Bench operationalizes standard firm objectives under scarce upstream supply. Let $y_{i,x}(t)$ be the number of units sold by agent i for item x at step t. Define procurement allocations $a_{i,x}(t)$ from Stage A and retail revenues from Stage B:

$$R_i(t) = \sum_x P_{i,x}(t)\, y_{i,x}(t) \tag{7}$$

Agents aim to maximize cumulative profit and avoid bankruptcy by maintaining nonnegative funds. We report automatically computed metrics spanning economic outcomes, such as profit and net profit margin; operational outcomes, such as stockout rate and fill rate, plus bounded stability indices such as IEI and OSI; and language-persona alignment.

This structure induces coupled economic tensions: bidding aggressively secures upstream supply but increases unit cost and reduces future liquidity; pricing aggressively increases sell-through but erodes margin; and slogans must simultaneously differentiate the agent semantically (to enter more buyers' consideration sets) while remaining consistent with the agent's pricing and inventory strategy.

3.6 Market-Bench as a Generative Dataset

Unlike static test sets, Market-Bench produces complete interaction traces that can be treated as a dataset of strategic behavior. Each run logs, per step and per agent, the procurement bids and allocations, posted prices and slogans, realized sales events, and resulting balance-sheet state (funds and inventory). This makes it possible to evaluate not only *final outcomes* but also *process-level* properties such as how agents revise bids across rounds, how language changes with observed market feedback, and how numeric policies respond to inventory dynamics. Beyond agent-level scores, these logs enable *dataset-level* analysis of market structure. We compute market-wide indices of inequality (Gini, Theil, CV), concentration (HHI, CR4), and participation (Active Ratio) directly from transaction outcomes. Figure 3 illustrates a typical evolution in our default setting: inequality rises quickly and participation declines, while the top-4 share increases even as HHI remains in the competitive range. These macro signals complement per-agent metrics and help characterize emergent dynamics in Market-Bench.

4 Experiments

4.1 Experimental Setup

Simulator. We run the Market-Bench simulator (Section 3) with m=20 retailer agents, each controlled by an LLM that outputs procurement bids, retail prices, and a marketing slogan. We evaluate the 20 LLM backends listed in Table 2 and log

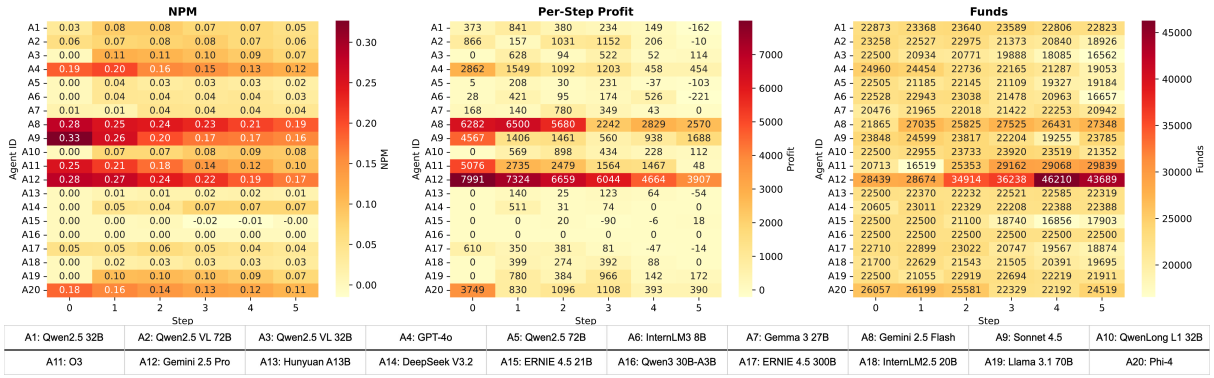

Figure 4: Temporal economic outcomes (mean across 10 runs). From left to right: net profit margin (NPM), per-step profit, and end-of-step funds. A small set of models, led by Gemini 2.5 Pro and Gemini 2.5 Flash, consistently earn high profits and compound capital, while many models remain near break-even.

complete transaction-level trajectories for reproducibility.

Environment and initialization. The default setting uses 8 items with tiered base prices (50 / 150 / 800 / 2000) and quantities (200, 200, 133, 133, 134, 75, 75, 50), with a horizon of 6 steps and 2 bidding rounds per step. Let total supplier supply be $S = \sum_x Q_x$ and total buyer demand be $D = rS$ with supply-demand ratio r (we use r=0.95). Initial funds are set proportional to the supplier catalog value:

$$K_{\text{init}} = \alpha \cdot \frac{\sum_x Q_x P_{\text{base}}(x)}{m} \tag{8}$$

In our default large-scale setting,

$$\sum_x Q_x P_{\text{base}}(x) = 300{,}000, \quad m{=}20, \quad \alpha{=}1.5, \tag{9}$$

giving K_{init}=22,500 per agent. Buyer volume is configured as $k = \beta m$, with k=200 and β=10. Persona-Gated Attention uses $K_{\max}$=20 and τ=1.0. All stress scenarios are disabled in these experiments to keep the economy stationary.

4.2 Experimental Results

We repeat the large-scale setting for 10 independent runs and report mean metrics across runs. Figures in this section report temporal economic, operational, and semantic dynamics aggregated over runs.

4.3 Analysis

Table 2 reveals performance dispersion that correlates with reasoning architecture rather than raw scale. Thinking-enabled models dominate: Gemini 2.5 Pro and Flash achieve the highest profits (Π=36,589 and 26,104) and margins (NPM=0.167 and 0.190), while O3 attains the best service level (FillRate=0.418, OSI=1.000) through explicit chain-of-thought reasoning. Among open-source models, Phi-4 presents a striking result: with only 14B parameters, it achieves Π=7,565, comparable to GPT-4o (Π=7,619), suggesting that reasoning-focused training transfers effectively to economic tasks. In contrast, larger MoE models show mixed outcomes: DeepSeek V3.2 (671B total) achieves high operational stability (OSI=0.829) but modest profits, while Hunyuan A13B attains the highest inventory efficiency (IEI=0.941) yet fails to convert this into profitability. Notably, model scale alone is a poor predictor: Phi-4 (14B) outperforms all open-source models including ERNIE 4.5 300B (Π=1,360), while Qwen3 30B-A3B achieves zero profit with zero BidEfficiency across all runs, indicating complete failure to produce valid bids. This mode of failure, where a single formatting error leads to total market exclusion, is unique to economic benchmarks with hard constraints. Beyond these aggregate patterns, Figure 4 reveals persistent stratification: a small set of models compounds capital early and maintains advantage, while others remain near break-even or drift into losses.

A rapid "entry shock" followed by stabilization. Across economic, operational, and semantic traces, the largest adjustments occur between steps 0 and 1. This is partly structural: OSI defaults to 1.0 at step 0 (insufficient points for variability), and then drops once order/sales variability becomes defined in step 1 (Figure 5). Semantically, slogan-persona similarities shift most in the first step and then stabilize (Figure 6), while the embedding clusters collapse quickly and remain tight by step 5 (Figure 7).

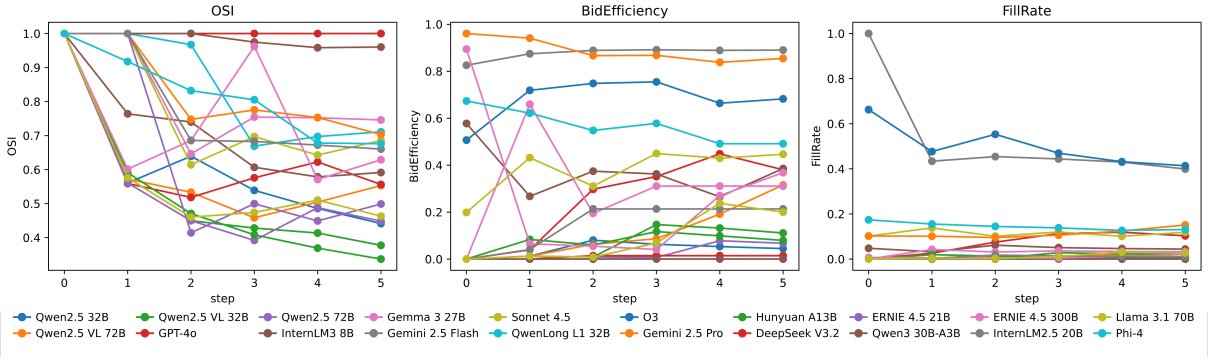

Figure 5: Operational dynamics over time (mean across 10 runs). Left: Order Stability Index (OSI). Middle: procurement BidEfficiency. Right: downstream FillRate. Agents with sustained auction success (high BidEfficiency) also achieve higher FillRate and tend to maintain higher OSI.

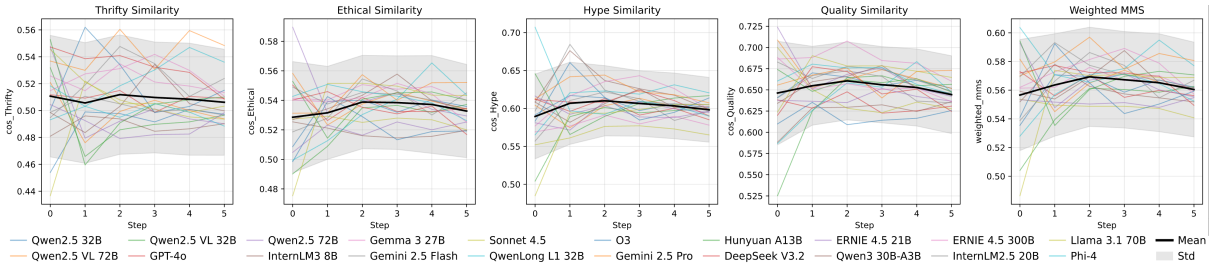

Figure 6: Slogan-persona similarity dynamics (mean ± std across 10 runs). Each line is an agent's average cosine similarity to each buyer tribe persona (Thrifty, Ethical, Hype, Quality) and the resulting weighted MMS. Across agents, similarities shift most between steps 0 and 1 and then stabilize.

Together, these patterns suggest that early market feedback rapidly compresses the strategy space: after first interaction, agents mostly refine within a settled regime rather than continually exploring.

Inequality rises without monopoly: a multi-winner oligopoly. Market indices (Figure 3) show wealth inequality tripling (Gini 0.07→0.21; Theil 0.02→0.10; CV 0.13→0.45), yet concentration stays competitive (HHI ≈ 0.08–0.10) while the top-4 share rises (CR4 0.23→0.33). This indicates an emergent "multi-winner" structure where a leading tier captures more share without monopoly. Rank mobility remains high, implying that relative ordering is contestable even as inequality grows, while the active ratio declines toward ≈ 0.4 as many agents become effectively inactive.

Scarcity makes procurement a threshold skill, producing bimodality. Profit correlates strongly with procurement success: Spearman ρ=0.68 with BidEfficiency, ρ=0.88 with FillRate, and ρ=−0.88 with StockoutRate. Under scarcity, failure to secure inventory is a first-order error that pricing and language cannot compensate for. Figure 5 reveals a *bimodal* regime: some agents sustain high BidEfficiency and meaningfully serve demand, while others remain near-zero. This separation is self-reinforcing,early success generates liquidity for future bids, whereas early failure produces revenue starvation,explaining the winner-take-most pattern in Figure 4.

Margin–volume trade-offs separate pricing from procurement. High FillRate does not guarantee high profit: o3 achieves the best FillRate but earns a lower margin than Gemini 2.5 Pro and Flash, suggesting aggressive bidding can trade margin for volume. Conversely, models with moderate service levels still profit through pricing discipline. Market-Bench thus distinguishes auction competence (acquiring supply) from retail competence (monetizing it). Figure 4 shows margin compression over time, consistent with intensified competition as early rents dissipate.

Operational indices require deconfounding with activity. Bounded indices such as IEI and OSI help summarize process properties, but can be misleading without conditioning on participation. For example, a low-activity agent that rarely procures or sells may appear stable (high OSI) or efficient

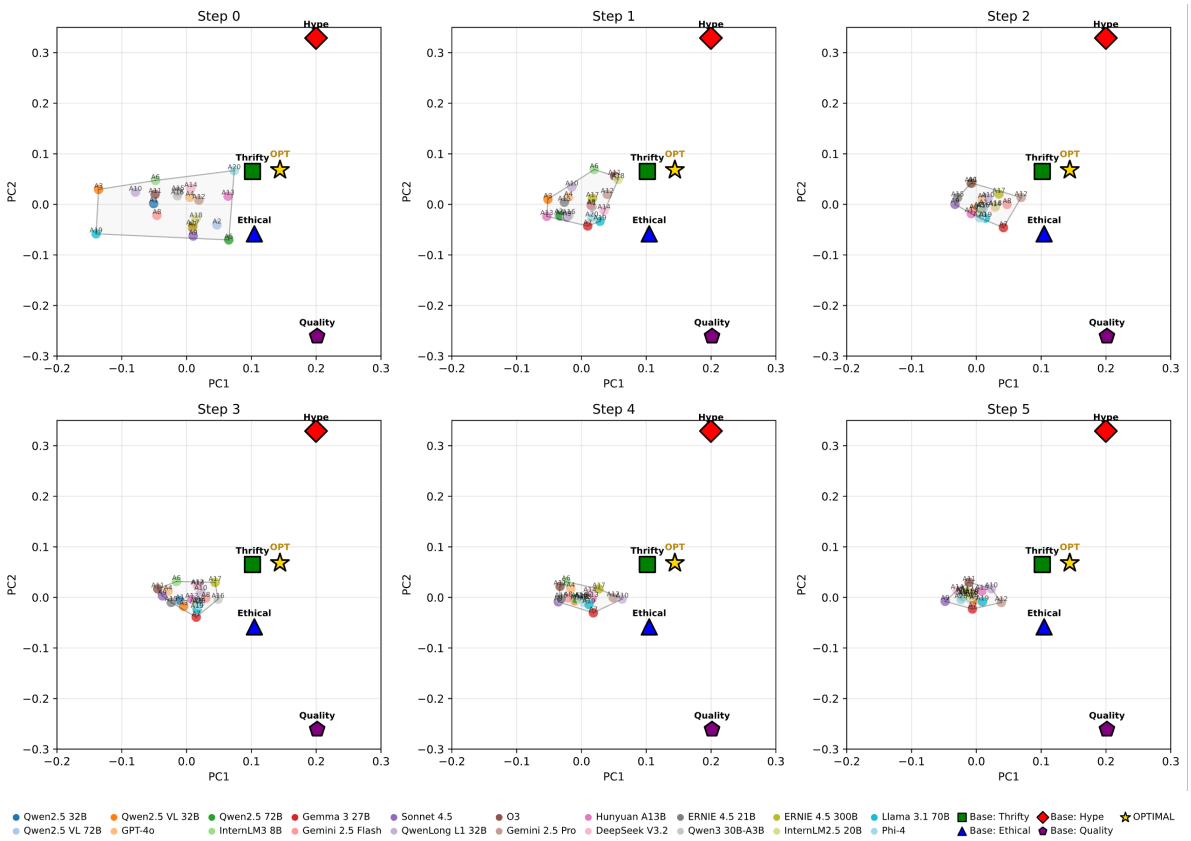

Figure 7: Evolution of slogan embedding clusters across steps (PCA projection of tribe-similarity features; mean across 10 runs). Agents start dispersed but rapidly converge to a tight region by step 5, suggesting that competitive feedback induces similar messaging strategies.

(high IEI) simply because there is little variability to measure. This motivates interpreting stability and efficiency jointly with market-access variables (BidEfficiency, FillRate, StockoutRate) and participation signals, rather than as standalone indicators of competence.

Language adapts quickly but converges toward a generic slogan optimum. MMS varies in a narrow band and is weakly correlated with profit (Spearman ρ=0.16). Figures 6 and 7 show that most slogan movement happens early and then converges, consistent with an emergent "messaging equilibrium." Since the buyer distribution places substantial slogan-sensitive mass on the Ethical and Hype tribes, agents converge near the Ethical-Hype axis but do not separate into tribe-specialized niches, suggesting slogans are learned through mimicry rather than buyer analysis. We view this rapid convergence not as a benchmark limitation but as an empirical finding about current LLM agents: under competitive pressure they collapse to a safe mimicry equilibrium, failing to sustain defensible niche positioning. Breaking this equilibrium,and maintaining semantic differentiation under competition,is a concrete open challenge that Market-Bench enables future work to study.

5 Conclusion

Market-Bench provides a reproducible market simulation for evaluating LLM agents where numeric decisions and language jointly determine economic outcomes. By coupling budget-constrained procurement with downstream price competition and semantic visibility, the benchmark produces logged trajectories that serve as a dataset of strategic behaviors. Our experiments reveal a stark "winner-take-most" dynamic. We observe a bimodal distribution of outcomes where a small elite of agents consistently compounds capital, while the majority struggle to break even. This suggests that while models may excel at isolated tasks, the synthesis of inventory management with market positioning remains a distinct frontier of difficulty. Market-Bench thus offers a necessary testbed for the next generation of agents designed to navigate the full complexity of the economic world.

Limitations

Simulation Scope. The experiments presented use a single market configuration with 20 agents, 6 time steps, and 8 item types. While Market-Bench's Hydra-based configuration system supports extensive customization, including agent count, episode length, supply-demand ratios, holding cost rates, and bidding round counts. We have not exhaustively explored all parameter combinations. Future work could systematically vary these parameters to characterize how LLM agent performance changes under different market conditions, such as highly competitive (supply $\ll$ demand) or relaxed (supply $\gg$ demand) scenarios.

Buyer Model. The buyer persona distribution in our experiments is synthetically generated with fixed weights (Thrifty 40%, Ethical 30%, Hype 20%, Quality 10%). Although the framework allows custom persona definitions and sensitivity parameters (λ, ρ), we did not explore the full space of buyer heterogeneity. Additionally, buyers currently make single-item purchases without memory of prior interactions. Extending the framework to support basket purchases, repeat customers, and buyer learning dynamics would provide richer evaluation scenarios for future investigation.

Future Directions. The modular architecture of Market-Bench enables several natural extensions: (1) dynamic supply with stochastic uncertainty to test adaptive procurement strategies; (2) multi-market scenarios with cross-market arbitrage opportunities; (3) longer episode horizons to study the emergence of complex trading strategies; and (4) non-English markets to evaluate multilingual LLM performance in economic contexts. These directions would further stress-test the economic reasoning and strategic planning capabilities of LLM agents beyond the current benchmark scope.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62225112, 62401365, 62522116, 62271312, 62132006, 62571324, and U24A20220; in part by the China Postdoctoral Science Foundation under Grants BX20250411 and 2025M773473; in part by the STCSM under Grant 22DZ2229005; and in part by the New Generation Artificial Intelligence-National Science and Technology Major Project (2025ZD0124104) in collaboration with the Shanghai Artificial Intelligence Laboratory.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Anthropic. 2025a. Claude sonnet 4.5 system card.

Anthropic. 2025b. Project vend: Can claude run a small shop? (and why does that matter?). `https://www.anthropic.com/research/project-vend-1`. In partnership with Andon Labs.

Robert Axelrod. 1984. *The Evolution of Cooperation*. Basic Books, New York.

Baidu ERNIE Team. 2025. Ernie 4.5 technical report.

Joyce Berg, John Dickhaut, and Kevin McCabe. 1995. Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122–142.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiao-wen Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, and 81 others. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Junhao Chen, Jingbo Sun, Xiang Li, Haidong Xin, Yuhao Xue, Yibin Xu, and Hao Zhao. 2025. Llmspark: A benchmark for evaluating large language models in strategic gaming contexts. *arXiv preprint arXiv:2509.16610*.

Michael Chui, Eric Hazan, Roger Roberts, Alex Singla, and Kate Smaje. 2023. The economic potential of generative ai.

DeepSeek-AI. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Yuan Deng, Vahab Mirrokni, Renato Paes Leme, Hanrui Zhang, and Song Zuo. 2024. Llms at the bargaining table. In *Agentic Markets Workshop at ICML 2024*. Poster.

Huiyu Duan, Qiang Hu, Jiarui Wang, Liu Yang, Zitong Xu, Lu Liu, Xiongkuo Min, Chunlei Cai, Tianxiao Ye, Xiaoyun Zhang, and Guangtao Zhai. 2025. Finevq: Fine-grained user generated content video quality assessment. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 3206–3217.

Huiyu Duan, Xiongkuo Min, Wei Sun, Yucheng Zhu, Xiao-Ping Zhang, and Guangtao Zhai. 2023. Attentive deep image quality assessment for omnidirectional stitching. *IEEE Journal of Selected Topics in Signal Processing*, 17(6):1150–1164.

Huiyu Duan, Xiongkuo Min, Yucheng Zhu, Guangtao Zhai, Xiaokang Yang, and Patrick Le Callet. 2022. Confusing image quality assessment: Toward better augmented reality experience. *IEEE Transactions on Image Processing*, 31:7206–7221.

Kshama Dwarakanath, Tucker Balch, and Svitlana Vyetrenko. 2024. Abides-economist: Agent-based simulator of economic systems with learning agents. *arXiv preprint arXiv:2402.09563*.

Ahmed El-Kishky, Alexander Wei, Andre Saraiva, Borys Minaiev, Daniel Selsam, David Dohan, Francis Song, Hunter Lightman, Ignasi Clavera, Jakub Pachocki, and 1 others. 2025. Competitive programming with large reasoning models. *arXiv preprint arXiv:2502.06807*.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Kaiyuan Ji, Yixuan Gao, Lu Sun, Yushuo Zheng, Zijian Chen, Jianbo Zhang, Xiangyang Zhu, Yuan Tian, Zicheng Zhang, and Guangtao Zhai. 2026. A^3: Towards Advertising Aesthetic Assessment. *arXiv preprint arXiv:2603.24037*.

Kaiyuan Ji, Yijin Guo, Zicheng Zhang, Xiangyang Zhu, Yuan Tian, Ning Liu, and Guangtao Zhai. 2025a. Medomni-45{\deg}: A safety-performance benchmark for reasoning-oriented llms in medicine. *arXiv preprint arXiv:2508.16213*.

Kaiyuan Ji, Jing Han, Guangtao Zhai, and Jiannan Liu. 2025b. Assessing the capabilities of generative pretrained transformer-4 in addressing open-ended inquiries of oral cancer. *International Dental Journal*, 75(1):158–165.

Kaiyuan Ji, Zhihan Wu, Jing Han, Jun Jia, Guangtao Zhai, and Jiannan Liu. 2024. Application of 3d nnunet with residual encoder in the 2024 miccai head and neck tumor segmentation challenge. In *Challenge on Head and Neck Tumor Segmentation for MRI-Guided Applications*, pages 250–258. Springer.

Kaiyuan Ji, Zhihan Wu, Jing Han, Guangtao Zhai, and Jiannan Liu. 2025c. Evaluating chatgpt-4's performance on oral and maxillofacial queries: Chain of thought and standard method. *Frontiers in Oral Health*, 6:1541976.

Jian Jin, Jiangyong Ying, Huiyu Duan, Liu Yang, Sijing Wu, Yunhao Li, Yushuo Zheng, Xiongkuo Min, and Guangtao Zhai. 2025. Rgc-vqa: An exploration database for robotic-generated video quality assessment. In *Proceedings of the ACM International Conference on Multimedia*.

Kevin Lane Keller. 2003. *Strategic Brand Management: Building, Measuring, and Managing Brand Equity*. Prentice Hall, Upper Saddle River, NJ.

Chiranjeev Kohli, Lance Leuthesser, and Rajneesh Suri. 2007. Got slogan? guidelines for creating effective slogans. *Business Horizons*, 50(5):415–422.

Lu Liu, Huiyu Duan, Qiang Hu, Liu Yang, Chunlei Cai, Tianxiao Ye, Huayu Liu, Xiaoyun Zhang, and Guangtao Zhai. 2024. F-bench: Rethinking human preference evaluation metrics for benchmarking face generation, customization, and restoration. *arXiv preprint arXiv:2412.13155*.

Shaoguang Mao, Yuzhe Cai, Yan Xia, Wenshan Wu, Xun Wang, Fengyi Wang, Tao Ge, and Furu Wei. 2023. Alympics: Llm agents meet game theory – exploring strategic decision-making with ai agents. *arXiv preprint arXiv:2311.03220*.

OpenAI. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

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

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2025. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*.

David Simchi-Levi, Konstantina Mellou, Ishai Menache, and Jeevan Pathuri. 2025. Large language models for supply chain decisions. *arXiv preprint arXiv:2507.21502*.

Statista. 2024. Artificial intelligence (ai) in retail market size worldwide from 2021 to 2029.

Llama Team. 2024a. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Qwen Team. 2024b. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Qwen Team. 2025a. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Qwen Team. 2025b. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Tencent Hunyuan Team. 2024. Hunyuan-large: An open-source moe model with 52 billion activated parameters by tencent. *arXiv preprint arXiv:2411.02265*.

Fanqi Wan, Weizhou Shen, Shengyi Liao, Yingcheng Shi, Chenliang Li, Ziyi Yang, Ji Zhang, Fei Huang, Jingren Zhou, and Ming Yan. 2025. Qwenlong-l1: Towards long-context large reasoning models with reinforcement learning. *ArXiv.org*.

Ruiyi Wang, Yushuo Zheng, Zicheng Zhang, Chunyi Li, Shuaicheng Liu, Guangtao Zhai, and Xiaohong Liu. 2025. Learning hazing to dehazing: Towards realistic haze generation for real-world image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Shuangqing Wang, Kaiyuan Ji, Yushuo Zheng, Zhihan Wu, Xiaorong Zhu, Zijian Chen, Lu Sun, Shuo Wang, Jianbo Zhang, Zicheng Zhang, and 1 others. 2026. Dental-qad: Reasoning-driven quality assessment and diagnosis in panoramic radiographs. *Displays*, page 103380.

Liu Yang, Huiyu Duan, Ran Tao, Juntao Cheng, Sijing Wu, Yunhao Li, Jing Liu, Xiongkuo Min, and Guangtao Zhai. 2025. Odi-bench: Can mllms understand immersive omnidirectional environments? *arXiv preprint arXiv:2510.11549*.

Zicheng Zhang, Ziheng Jia, Chunyi Li, Yingjie Zhou, Xiaohong Liu, Xiongkuo Min, and Guangtao Zhai. 2026a. Towards versatile multimedia quality assessment for visual communications. *Science China Information Sciences*, 69:122306.

Zicheng Zhang, Junying Wang, Yijin Guo, Farong Wen, Zijian Chen, Hanqing Wang, Wenzhe Li, Lu Sun, Yingjie Zhou, Jianbo Zhang, Bowen Yan, Ziheng Jia, Jiahao Xiao, Yuan Tian, Xiangyang Zhu, Kaiwei Zhang, Chunyi Li, Xiaohong Liu, Xiongkuo Min, and 2 others. 2026b. AIBench: Towards trustworthy evaluation under the 45° law. *Displays*, 91:103255.

Zicheng Zhang, Junying Wang, Farong Wen, Yijin Guo, Xiangyu Zhao, Xinyu Fang, Shengyuan Ding, Ziheng Jia, Jiahao Xiao, Ye Shen, Yushuo Zheng, Xiaorong Zhu, Yalun Wu, Ziheng Jiao, Wei Sun, Zijian Chen, Kaiwei Zhang, Kang Fu, Yuqin Cao, and 30 others. 2025a. Large multimodal models evaluation: A survey. *Science China Information Sciences*.

Zicheng Zhang, Haoning Wu, Ziheng Jia, Weisi Lin, and Guangtao Zhai. 2025b. Teaching lmms for image quality scoring and interpreting.

Yushuo Zheng, Huiyu Duan, Zicheng Zhang, Xiaohong Liu, and Xiongkuo Min. 2026. Learning to wander: Improving the global image geolocation ability of lmms via actionable reasoning. *Preprint*, arXiv:2603.10463.

Yushuo Zheng, Jiangyong Ying, Huiyu Duan, Chunyi Li, Zicheng Zhang, Jing Liu, Xiaohong Liu, and Guangtao Zhai. 2025a. Geox-bench: Benchmarking cross-view geo-localization and pose estimation capabilities of large multimodal models. *arXiv preprint arXiv:2511.13259*.

Yushuo Zheng, Zicheng Zhang, Xiongkuo Min, Huiyu Duan, and Guangtao Zhai. 2025b. Lm fight arena: Benchmarking large multimodal models via game competition. *Preprint*, arXiv:2510.08928.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Zhengyang Qi, Haofei Yu, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. Sotopia: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations (ICLR)*.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Robert Tang, Heng Ji, and Jiaxuan You. 2025. Multiagentbench: Evaluating the collaboration and competition of llm agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8580–8622, Vienna, Austria. Association for Computational Linguistics.

A Simulation Algorithm

The complete simulation loop for Market-Bench proceeds as follows. Each episode consists of T steps, with each step comprising a procurement stage (Stage A) and a retail stage (Stage B).

Simulation Loop.

1. **Initialize**: Set $\text{Funds}_i \leftarrow K_{\text{init}}$, $\text{Inv}_{i,x} \leftarrow 0$ for all agents $i \in \mathcal{A}$ and items $x \in \mathcal{X}$.
2. **For each step** $t = 0, \ldots, T-1$:
   (a) Prepare supplier offers $O_S(t)$.
   (b) **Stage A (Procurement)**:
   - For each bidding round $r = 1, \ldots, R_{\max}$:
   - Build bidding state with previous round results.
   - Each agent i submits bid b_i via LLM call (in parallel).
   - Validate budget constraints; reject overspending bids.

   (c) Settle final bids: allocate items to highest bidders above reserve price.
   (d) Update Funds_i and $\text{Inv}_{i,x}$ based on allocations.
   (e) **Stage B (Retail)**:
   - Each agent i outputs prices P_i and slogan via LLM call (in parallel).

   (f) Generate buyers $\mathcal{B}_t$ with persona embeddings.
   (g) For each buyer B_j: compute attention weights, sample consideration set, purchase from lowest-priced available seller.
   (h) Apply holding costs and check for bankruptcy.
   (i) Record metrics and market history.
3. **Return**: Logged trajectories and final metrics.

A.1 Bid Settlement Procedure

For each item x, bids are sorted by price descending with random tie-breaking. The available quantity Q_x is allocated greedily to highest bidders above the reserve price $P_{\text{base}}(x)$:

$$a_{i,x} = \min(q_{i,x}, \text{remaining}) \quad \text{if } p^{\text{bid}}_{i,x} \geq P_{\text{base}}(x) \tag{10}$$

B LLM Agent Prompts

B.1 Procurement Stage Prompt

The system prompt instructs the LLM to act as a retail agent in a sealed-bid auction:

```
You are a retail agent participating in a multi-round sealed-bid auction for supplier inventory. Output ONLY valid, parseable JSON. No prose.
Output schema (strict):

{
 "bids": {
  "<item_id>": {"qty":<int>,"price":<int>}
 }
}

Rules:
  1. Only bid on items in the supplier offers.
  2. qty and price must be non-negative integers.
  3. Do NOT overspend: total spend must not exceed Funds.
  4. Do NOT bid below base_price.
  5. JSON must be strict (no trailing commas).
```

Example Procurement Input.

```
Step: 0
Round: 1 of 2
Funds: 22500
Overspent on most recent bid: False
Supplier offers:
item_id,qty,base_price
item1,200,50
item2,200,50
item3,133,150
...
Current inventory: item1,0; item2,0;
...
Historical sales: (empty at step 0)
```

Example Procurement Output (Gemini 2.5 Pro).

```
{
  "bids": {
    "item1": {"qty": 110, "price": 51},
    "item2": {"qty": 110, "price": 51},
    "item3": {"qty": 75, "price": 151}
  }
}
```

B.2 Retail Stage Prompt

The system prompt for the retail stage:

```
You are setting retail prices and a marketing slogan for your current catalog. Output ONLY valid, parseable JSON.
Output schema (strict):

{
  "prices": { "<item_id>": <int>, ... },
  "slogan": "<string>"
}

Rules:
  1. You do NOT know buyer personas; infer from market_history.
  2. price must be a non-negative integer.
  3. Slogan should resonate with inferred personas.
  4. Keep slogan short (≤ 25 words).
```

Example Retail Input.

```
Step: 0
Funds: 10000
Current inventory:
item4: 19, item5: 19, item6: 3, item7: 3, item8: 1
Historical sales: (empty at step 0)
```

Example Retail Output (GPT-4o).

```
{
  "prices": {
    "item4": 200,
    "item5": 200,
    "item6": 1000,
    "item7": 1000,
    "item8": 2500
  },
 "slogan": "Luxury and value, redefined!"
}
```

Example Retail Output (Claude Sonnet 4.5).

```
{
  "prices": {
    "item2": 75,
    "item4": 225,
    "item6": 1200,
    "item7": 1200,
    "item8": 3000
  },
  "slogan": "Premium quality at prices
            that make sense."
}
```

C Evaluation Metric Definitions

Market-Bench computes 9 per-agent metrics spanning economic, operational, and cognitive dimensions. Table 2 presents results using these metrics.

C.1 Economic Metrics

Net Profit Margin (NPM). Measures profitability relative to revenue:

$$\text{NPM} = \frac{\sum_{t=0}^{T-1} \Pi_i(t)}{\sum_{t=0}^{T-1} R_i(t) + \epsilon} \tag{11}$$

where $\Pi_i(t) = R_i(t) - \text{COGS}_i(t) - H_i(t)$ is per-step profit, $R_i(t)$ is revenue, $\text{COGS}_i(t)$ is cost of goods sold, and $H_i(t)$ is holding cost. Higher is better.

Cumulative Profit (Pi). Total profit across all steps:

$$\Pi_i = \sum_{t=0}^{T-1} \Pi_i(t) \tag{12}$$

Higher is better.

Risk-Adjusted Return (RAR). Sharpe-like ratio of mean profit to profit volatility:

$$\text{RAR}_i = \frac{\mu(\Pi_i(t))}{\sigma(\Pi_i(t)) + \epsilon} \tag{13}$$

where $\mu(\Pi_i(t))$ and $\sigma(\Pi_i(t))$ are the mean and standard deviation of per-step profits. Higher is better.

Inventory Efficiency Index (IEI). Measures the fraction of goods sold relative to total throughput:

$$\text{IEI}_i = \frac{U_i^{\text{sold}}}{U_i^{\text{sold}} + \bar{I}_i + \epsilon} \tag{14}$$

where U_i^{sold} is total units sold and $\bar{I}_i$ is average inventory in units. IEI $\in [0, 1]$, higher is better.

C.2 Operational Metrics

Stockout Rate. Fraction of directed purchase attempts that failed due to zero inventory:

$$\text{StockoutRate}_i = \frac{N_i^{\text{stockout}}}{N_i^{\text{attempts}} + \epsilon} \tag{15}$$

Lower is better.

Bid Efficiency. Combined measure of procurement success and cost efficiency:

$$\text{BidEff}_i = \frac{Q_i^{\text{win}}}{Q_i^{\text{bid}} + \epsilon} \times \frac{V_i^{\text{base}}}{C_i^{\text{win}} + \epsilon} \tag{16}$$

where Q_i^{win} is won quantity, Q_i^{bid} is bid quantity, V_i^{base} is base value of won goods, and C_i^{win} is actual spend. Higher is better.

Order Stability Index (OSI). Measures how well order variability matches sales variability:

$$\text{OSI}_i = \frac{1}{1 + |\text{CV}_{\text{orders}} - \text{CV}_{\text{sales}}|} \tag{17}$$

where the coefficient of variation is:

$$\text{CV}_X = \frac{\sigma_X}{\mu_X + \epsilon} \tag{18}$$

OSI $\in [0, 1]$, higher is better.

Fill Rate. Fraction of directed demand (in units) successfully fulfilled:

$$\text{FillRate}_i = \frac{U_i^{\text{sold}}}{U_i^{\text{demand}} + \epsilon} \tag{19}$$

Higher is better.

C.3 Cognitive/Semantic Metric

Mean Match Score (MMS). Average cosine similarity between agent slogans and buyer personas:

$$\text{MMS}_i = \frac{1}{|\mathcal{B}_i^{\text{interact}}|} \sum_j \text{Sim}(i, j) \tag{20}$$

where $\text{Sim}(i, j) = \cos(\mathbf{E}(\text{Slogan}_i(t)), \mathbf{E}(\text{Persona}_j))$. Higher is better.

D Metric Replacement Rationale

Two traditional supply chain metrics were replaced with bounded alternatives due to numerical instability.

D.1 IEI

The Inventory Turnover Ratio (ITR = COGS / Avg. Inventory) becomes unbounded when average inventory approaches zero. Observed values ranged from 0 to 5.4×10^{12}, making comparison meaningless.

The Inventory Efficiency Index reformulates the metric:

$$\text{IEI} = \frac{U_{\text{sold}}}{U_{\text{sold}} + \bar{I} + \epsilon} \in [0, 1] \tag{21}$$

D.2 OSI

The Bullwhip ratio (Var(Orders) / Var(Sales)) explodes when sales variance is near zero. Observed values ranged from 0.27 to 7.3×10^{12}.

The Order Stability Index uses coefficient of variation:

$$\text{OSI} = \frac{1}{1 + |\text{CV}_{\text{orders}} - \text{CV}_{\text{sales}}|} \in [0, 1] \quad (22)$$

E Persona-Gated Attention Mechanism

The attention weight for buyer j considering agent i is:

$$w_{i,j} = \exp\left(\frac{\lambda_j \cdot \text{Sim}(i, j)}{\tau}\right) \quad (23)$$

where $\text{Sim}(i, j) = \cos(\mathbf{E}(\text{Slogan}_i(t)), \mathbf{E}(\text{Persona}_j))$, λ_j is slogan sensitivity, τ is semantic temperature, and $\mathbf{E}(\cdot)$ is the embedding function (Qwen3-Embedding-8B).

The consideration set size is:

$$|V_j| = \min\left(|\mathcal{A}|, \max(1, \lceil \rho_j \cdot K_{\max} \rceil)\right) \quad (24)$$

F Buyer Persona Configuration

Table 3 shows the default buyer persona distribution.

Tribe	Weight	λ	Keywords
Thrifty	0.4	0.2	(price-focused)
Ethical	0.3	0.8	green, fair, eco
Hype	0.2	0.9	exclusive, limited
Quality	0.1	0.5	quality, craft

Table 3: Default buyer persona distribution. λ denotes slogan sensitivity weight.

G Experimental Configuration

Table 4 summarizes the key simulation parameters.

Parameter	Value
Number of agents	20
Number of steps	6
Bidding rounds	2
Buyers per round	200
Initial funds	22,500
Item categories	8
Total supply units	1,000
Base prices	50–2,000
Supply-demand ratio	0.95
Holding cost rate	0.0
Buyer patience (ρ)	0.6
Embedding model	Qwen3-Embedding-8B
Temperature (τ)	1.0
$K_{\max}$	20
LLM Temperature	0.0
Response format	JSON

Table 4: Default experimental configuration.

H Supplier Offer Structure

Table 5 details the supplier offer configuration.

Item	Category	Qty	Base Price
item1	Commodity	200	50
item2	Commodity	200	50
item3	Standard	133	150
item4	Standard	133	150
item5	Standard	134	150
item6	Luxury	75	800
item7	Luxury	75	800
item8	Veblen	50	2,000
Total		1,000	–

Table 5: Supplier offer structure for 20-agent experiments.

I Evaluated Models

Table 6 lists the 20 LLM agents evaluated, spanning both open-weight and closed-source models.

Category	Models
Closed-source	Gemini 2.5 Pro, Gemini 2.5 Flash, O3, Claude Sonnet 4.5, GPT-4o
Open-weight	Phi-4, Qwen2.5 VL 72B, Llama 3.1 70B, QwenLong L1 32B, Qwen2.5 32B, Gemma 3 27B, Qwen2.5 VL 32B, ERNIE 4.5 300B, InternLM2.5 20B, InternLM3 8B, DeepSeek V3.2, Qwen2.5 72B, Hunyuan A13B, Qwen3 30B-A3B, ERNIE 4.5 21B

Table 6: LLM agents evaluated in Market-Bench experiments.