TRADERBENCH: HOW ROBUST ARE AI AGENTS IN ADVERSARIAL CAPITAL MARKETS?

Xiaochuang Yuan *†
Amazon.com Inc.
yxc20098@gmail.com

Hui Xu*
Stony Brook University
huixucom@gmail.com

Silvia Xu
Stanford University
sx0771@gmail.com

Cui Zou
University of Oklahoma
tracyzou@ou.edu

Jing Xiong
UC Santa Cruz
jxiong20@outlook.com

ABSTRACT

Evaluating AI agents in finance faces two key challenges: static benchmarks require costly expert annotation yet miss the dynamic decision-making central to real-world trading, while LLM-based judges introduce uncontrolled variance on domain-specific tasks. We introduce TraderBench, a benchmark that addresses both issues. It combines expert-verified static tasks (knowledge retrieval, analytical reasoning) with adversarial trading simulations scored purely on realized performance—Sharpe ratio, returns, and drawdown—eliminating judge variance entirely. The framework features two novel tracks: crypto trading with four progressive market-manipulation transforms, and options derivatives scoring across P&L accuracy, Greeks, and risk management. Trading scenarios can be refreshed with new market data to prevent benchmark contamination. Evaluating 12 models (8B open-source to frontier) on $\sim$50 tasks, we find: (1) 7 of 12 models score $\sim$33 on crypto with $<$1-point variation across adversarial conditions, exposing fixed non-adaptive strategies; (2) extended thinking helps retrieval (+26 points) but has zero impact on trading (+0.3 crypto, $-$0.1 options). These findings reveal that current agents lack genuine market adaptation, underscoring the need for performance-grounded evaluation in finance.

1 INTRODUCTION

How robust are AI agents when faced with adversarial capital-market conditions? As LLM-powered agents move from question answering into autonomous trading, portfolio management, and risk analysis, their failures carry direct monetary consequences. Evaluating these agents requires more than static Q&A accuracy—it demands testing under adversarial market manipulation, verifying quantitative precision on derivatives calculations, and ensuring that the evaluation itself is reliable.

Existing finance benchmarks occupy two extremes. On one hand, Finance Agent Benchmark (FAB) (Bigeard et al., 2025), BizFinBench (Guo et al., 2025), and FinBen (Xie et al., 2024) evaluate financial knowledge through static Q&A—they test what agents *know* but not how they *act* under real market conditions with live price feeds, evolving positions, and adversarial signals. On the other hand, LiveTradeBench (Yu et al., 2025) deploys agents in live markets, but requires 50 days of real-time execution per evaluation round, making rapid iteration across models impractical. Neither extreme (1) tests agents under controlled adversarial trading conditions with manipulated market data, (2) decomposes derivatives competence into quantitative accuracy versus qualitative reasoning, or (3) measures how sensitive scores are to the choice of LLM judge (Zheng et al., 2023).

We introduce TraderBench, a benchmark that evaluates AI agents across four equally weighted sections—Knowledge Retrieval, Analytical Reasoning, Options Trading, and Crypto Trading—using a two-agent architecture built on the A2A protocol (Google and Linux Foundation, 2025) with

*Equal contribution.
†Corresponding author

six MCP servers (Anthropic, 2024) for financial data access. Its two novel evaluation tracks are: *adversarial crypto trading*, which applies four progressive market-manipulation transforms (baseline $\rightarrow$ noisy $\rightarrow$ meta $\rightarrow$ adversarial) to test strategy robustness; and *options derivatives scoring*, which separately measures quantitative accuracy (P&L calculations, Greeks precision) and qualitative reasoning (strategy selection, risk management).

Our contributions are:

1. TraderBench benchmark: Two novel evaluation tracks—adversarial crypto trading and decomposed options scoring—within a four-section framework, built on open A2A and MCP protocols for reproducible evaluation.
2. Robustness findings: Across 12 models (8B to frontier), 7 of 12 adopt fixed crypto strategies with $<$2-point variation across adversarial transforms, while a 54-point quantitative-vs-qualitative gap in options persists across all model sizes.
3. Evaluation reliability: Re-scoring identical outputs with three LLM judges yields an 11-point overall spread; performance-based crypto scores vary by only 0.3 points versus 29 for rubric-based retrieval.
4. Scaling and reasoning analysis: Extended thinking improves tool-use planning (+26 on retrieval) but has zero impact on trading; the proprietary–open-source gap is driven by knowledge retrieval, not by options or crypto performance.

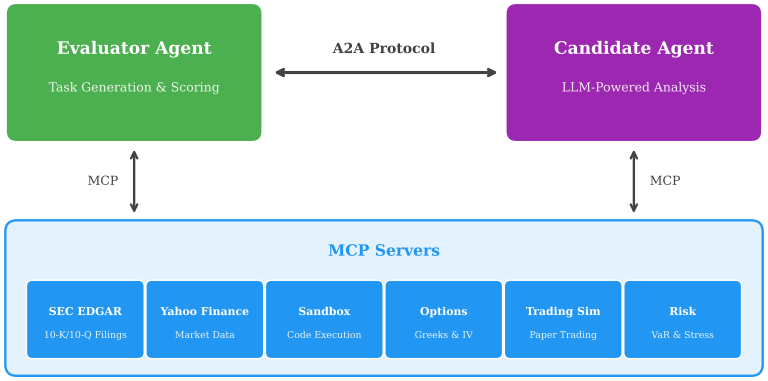

Figure 1: Overview of the TraderBench two-agent architecture. The Evaluator Agent generates tasks from six datasets and sends them to the Candidate Agent via the A2A protocol. The Candidate Agent uses an LLM and six MCP servers to access financial data, execute code, and simulate trades. Responses are scored by dataset-specific evaluators.

2 Related Work

Beyond the benchmarks discussed above, PRBench (Akyürek et al., 2025) and GDPVal (OpenAI, 2026) test professional reasoning but without trading or tool use. The LLM-as-judge paradigm (Zheng et al., 2023) enables scalable scoring but introduces inter-judge variance (Verga et al., 2024; Wang et al., 2023), which we quantify across multiple judges in the financial domain. Work on agent safety (Ruan et al., 2024; Li et al., 2024) highlights adversarial risks but lacks domain-specific trading benchmarks.

3 The TraderBench Benchmark

3.1 Architecture

TraderBench implements a two-agent evaluation architecture (Figure 1) built on the AgentBeats platform (Berkeley RDI, 2025). The Evaluator Agent orchestrates the benchmark: it loads evaluation configurations, generates tasks from six datasets, sends them to the candidate agent via the A2A

protocol (Google and Linux Foundation, 2025), and scores responses using dataset-specific evaluators. The Candidate Agent is the system under test—it receives financial analysis tasks and must produce responses using an underlying LLM and, critically, six specialized MCP servers (Anthropic, 2024) that provide financial data access.

Table 1: The six MCP servers available to the Candidate Agent.

Server	Function	Key Feature
SEC EDGAR (U.S. Securities and Exchange Commission)	10-K/10-Q filings	Temporal locking (no lookahead)
Yahoo Finance	Market data	Lookahead detection
Sandbox	Python execution	Sandboxed computation
Options	Black-Scholes pricing	Greeks & IV calculation
Trading Sim	Paper trading	Slippage modeling
Risk	Portfolio risk metrics	VaR, Sharpe/Sortino

3.2 Evaluation Sections

Tasks are organized into four equally weighted sections (Table 2), each targeting distinct capabilities. The equal 25% weighting prevents agents from achieving high overall scores by excelling in only one area. The current evaluation uses $\sim$50 tasks sampled from a larger pool; future evaluations can scale to hundreds by enabling additional tasks per section.

Table 2: Overview of the four TraderBench evaluation sections (25% each). Tool Use indicates whether MCP server interaction is required to answer correctly.

Section	Datasets	Tasks	Evaluator	Tool Use
Knowledge Retrieval	BizFinBench, PRBench	18	Exact match + LLM rubric (Zheng et al., 2023)	Required
Analytical Reasoning	Synthetic (CFA-level[†])	9	LLM rubric (3 components)	Not required
Options Trading	Options Alpha	9	4-dim scoring	Required
Crypto Trading	Crypto (4 transforms)	6	Performance-based	Minimal

[†]CFA: Chartered Financial Analyst, a professional designation whose curriculum defines standard financial computation competencies.

Knowledge Retrieval tests whether agents can accurately extract financial facts and perform quantitative computations using real company data. BizFinBench (Guo et al., 2025) provides bilingual event logic reasoning and financial computation tasks, while PRBench (Akyürek et al., 2025) tests multi-step professional reasoning. Both require agents to retrieve data from U.S. Securities and Exchange Commission (SEC) filings and market data servers—agents answering from parametric knowledge alone produce stale or incorrect figures.

Analytical Reasoning evaluates self-contained financial computation through multi-step problems covering net present value (NPV) and discount rate analysis, portfolio beta adjustment, bond pricing, free cash flow valuation, options strategies (put-call parity, spreads), leverage effects (Modigliani-Miller), binomial option pricing, duration immunization, and interest rate swaps. All information is provided in the question; no external data retrieval is needed, isolating reasoning from tool use. Scoring uses an LLM rubric (Zheng et al., 2023) with three components: methodology (30%), calculation (30%), and final answer (40%).

Options Trading assesses quantitative derivatives knowledge through four sub-dimensions scored equally (25% each): P&L accuracy (max profit/loss, breakeven calculations), Greeks accuracy (delta, gamma, theta, vega within 5% tolerance), strategy quality (multi-leg construction and rationale), and risk management (position sizing and hedging).

Crypto Trading evaluates agents under adversarial market conditions (Section 3.3). Performance is measured by a weighted combination of total return (35%), Sharpe ratio (30%), win rate (20%), and maximum drawdown penalty (15%), aggregated across four progressively adversarial transform conditions.

Table 3: Adversarial transform conditions applied to historical crypto price data. Each transform is applied independently to the same underlying price series.

Transform	Description
Baseline	Clean historical price data. Control condition for measuring base performance.
Noisy	Gaussian price noise ($\sigma = 2\%$) and sporadic volume spikes ($3\times$ normal). Tests microstructure robustness.
Meta	Combined noise patterns with trend modifications, false breakouts, and support/resistance violations. Applied independently of Noisy.
Adversarial	Coordinated false signals targeting common strategies: moving average crossovers, Relative Strength Index (RSI) divergences, and Moving Average Convergence Divergence (MACD) signal injections. Applied independently of prior transforms.

3.3 Adversarial Crypto Trading

The crypto trading section applies four progressive data transforms to historical cryptocurrency price data, testing whether agents adapt their trading strategies to deteriorating market conditions. Each transform is applied *independently* to the same underlying historical price series—they are not stacked cumulatively. This design isolates each manipulation type's individual effect on agent behavior and avoids conflating signal degradation across conditions. Score recovery observed in the Meta and Adversarial transforms for some models (e.g., Gemma3-27B, Section 5.4) therefore reflects genuine strategy differences in response to those specific signal patterns, not attenuation of Noisy-condition effects.

Transform scores are weighted to emphasize baseline performance while penalizing adversarial fragility: baseline (40%), noisy (30%), adversarial (20%), meta (10%). This weighting reflects a deployment-oriented priority: an agent that performs well under normal conditions and degrades gracefully under attack is preferable to one that is uniformly mediocre.

3.4 Unified Scoring

All evaluator outputs are normalized to a $[0, 100]$ scale. For each section s, the section score is the mean of normalized task scores: $S_s = \frac{1}{|T_s|} \sum_{t \in T_s} \text{norm}(\text{score}_t)$. The overall score is the weighted sum across active sections:

$$\text{Overall} = \sum_{s \in \mathcal{S}} w_s \cdot S_s, \quad w_s = 0.25 \; \forall s \tag{1}$$

If any section has no tasks (e.g., in a focused evaluation), weights are redistributed proportionally among active sections.

4 Experimental Setup

4.1 Models

We evaluate 12 models spanning frontier proprietary systems, open-weight models, and fully open-source models (Table 4). All models use the same Candidate Agent infrastructure with identical MCP server access, ensuring differences reflect model capability rather than infrastructure variation. We additionally test two ablation variants: GPT-5.2 with web search augmentation and Qwen3-32B with extended thinking mode (discussed in Section 5.2).

4.2 Evaluation Configuration

All evaluations use the same configuration: approximately 50 tasks drawn by stratified sampling (seed 42), of which 42 are scored across the four main sections; the remaining professional tasks are analyzed separately (Appendix C). The LLM judge for rubric-based evaluation sections is GPT-5.2

Table 4: Models evaluated. All access the same MCP server infrastructure through the Candidate Agent. Web search and thinking ablations are discussed separately.

Model	Type	Size	Provider
Proprietary			
Gemini-3-Pro	Proprietary	–	Google
Kimi-K2.5	Proprietary	–	Moonshot AI
GPT-5.2	Proprietary	–	OpenAI
GPT-4o	Proprietary	–	OpenAI
Grok 4.1 Fast	Proprietary	–	xAI
Open-weight			
GPT-OSS-120B	Open-weight	120B	OpenAI
GPT-OSS-20B	Open-weight	20B	OpenAI
Open-source			
Qwen3-32B	Open-source	32B	Alibaba
Qwen3-30B-A3B	Open-source (MoE)	30B (3B active)	Alibaba
Qwen3-8B	Open-source	8B	Alibaba
Gemma3-27B	Open-source	27B	Google
GLM-4.7-Flash	Open-source (MoE)	30B (3B active)	Zhipu AI

Table 5: Main results. Models sorted by overall score (mean of four sections at 25% each). **Bold**: best in column. Underline: second best. KR = Knowledge Retrieval, AR = Analytical Reasoning, Opt = Options Trading, Cry = Crypto Trading.

Model	Overall	KR	AR	Opt	Cry
Gemini-3-Pro	**64.3**	52.4	**94.8**	63.2	46.6
Grok 4.1 Fast	63.7	**61.6**	87.9	**72.2**	33.2
GPT-5.2	61.9	50.6	87.3	62.1	47.4
Kimi-K2.5	54.4	37.1	71.0	62.6	46.8
GLM-4.7-Flash	53.9	38.9	82.1	61.0	33.5
GPT-4o	53.3	32.3	80.4	55.6	44.9
GPT-OSS-120B	50.9	32.5	74.8	63.2	33.1
GPT-OSS-20B	50.6	42.4	72.1	55.4	32.6
Qwen3-32B	48.4	14.4	84.0	62.3	32.8
Gemma3-27B	47.2	9.3	68.7	59.0	**51.7**
Qwen3-8B	44.9	10.5	77.0	58.9	33.1
Qwen3-30B-A3B	44.7	9.4	74.1	62.6	32.7

(temperature 0.0) for main results. The judge reliability study (Section 5.5) re-evaluates identical responses with three additional judge models. Evaluations run with a 6-hour timeout per model.

5 Results and Analysis

5.1 Overall Performance and Capability Stratification

Table 5 and Figure 2 present the full leaderboard. Two findings emerge. First, there is a clear tier structure: Gemini-3-Pro leads at 64.3 driven by the highest Analytical Reasoning (94.8) and strong Crypto (46.6), closely followed by Grok 4.1 Fast (63.7) with the best Knowledge Retrieval (61.6) and Options (72.2), then GPT-5.2 at 61.9, mid-tier proprietary and open-weight models at 50–54, and smaller open-source models at 44–48. The 20-point gap between best and worst demonstrates substantial capability stratification.

Second, the section that drives this gap is Knowledge Retrieval (range: 9.3–61.6). In contrast, Analytical Reasoning shows far less variance (68.7–94.8), and Options Trading is remarkably consistent (55.4–72.2). This reveals that the differentiating factor between models is not raw reasoning ability, but effective tool use for data retrieval.

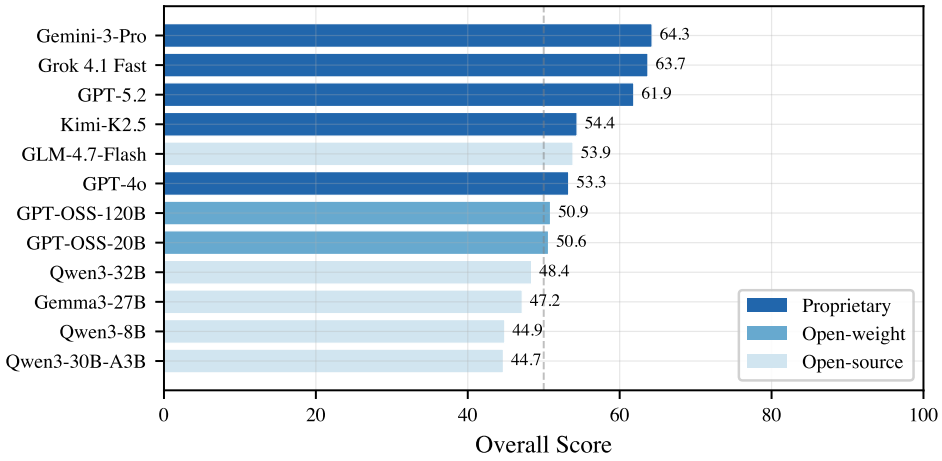

Figure 2: Overall TraderBench scores for all 12 models, sorted by performance. The dashed line marks the 50/100 midpoint. Frontier proprietary models (left) clearly separate from smaller open-source models (right).

5.2 Knowledge Retrieval: Tool Use as the Key Differentiator

Knowledge Retrieval has by far the highest cross-model variance (std. dev. 17.6 vs. 8.0 for AR), since it requires retrieving financial data from SEC filings and market data servers via MCP tools—agents answering from parametric knowledge alone produce stale or incorrect figures.

Figure 3 illustrates a stark divide: models that effectively use MCP tools (GPT-5.2: 50.6, Gemini-3-Pro: 52.4) dramatically outperform those that do not (Qwen3-8B: 10.5, Gemma3-27B: 9.3). Critically, models scoring below 15 on KR are not weak reasoners—Qwen3-32B scores 84.0 on AR (4th overall) but only 14.4 on KR (9th overall), demonstrating that model capability is necessary but not sufficient.

Two ablations confirm this, designed to probe opposite ends of the tool-use spectrum. The GPT-5.2+WS ablation adds a web search MCP server to an already strong tool user, testing the *ceiling* of tool augmentation: this yields +4.6 on KR (50.6 $\rightarrow$ 55.2). The Qwen3-32B thinking ablation enables chain-of-thought in a model that struggles with tool-use planning despite strong reasoning, testing the *floor*: this yields +26.1 on KR (14.4 $\rightarrow$ 40.5), the largest single-section improvement in our study. These two ablations are thus complementary rather than redundant—one tests access, the other tests planning. Notably, enabling extended thinking in Qwen3-32B has minimal effect on Analytical Reasoning (+1.8 points), which requires no tool use, confirming that the thinking benefit is specific to multi-step tool-use planning rather than general reasoning enhancement. For tool-dependent tasks, improving tool access and planning matters more than model scale.

5.3 The Conceptual-vs-Computational Gap in Options Trading

Options Trading overall scores range 55.4–72.2, but sub-dimensions reveal a striking and systemic pattern (Figure 4). Across all 12 models, performance on conceptual tasks such as P&L accuracy (80–93) and Strategy quality (65–75) consistently and significantly exceeds performance on computational tasks like Greeks precision (18–53) and Risk management (48–72).

This conceptual-vs-computational gap is pervasive and highlights a fundamental cognitive dissonance in current LLMs: models operate effectively as "semantic strategists" but fail as "numerical analysts." Specifically, models can correctly identify complex setups (e.g., identifying that an Iron Condor is appropriate for low-volatility environments) and calculate arithmetic expiration payoffs. However, they struggle profoundly to compute derivative sensitivities (delta, gamma, theta, vega), which require understanding instantaneous rates of change.

Crucially, this failure persists even when models are equipped with an options pricing MCP server. Inspection of MCP interaction logs reveals that the bottleneck is not merely internal computational

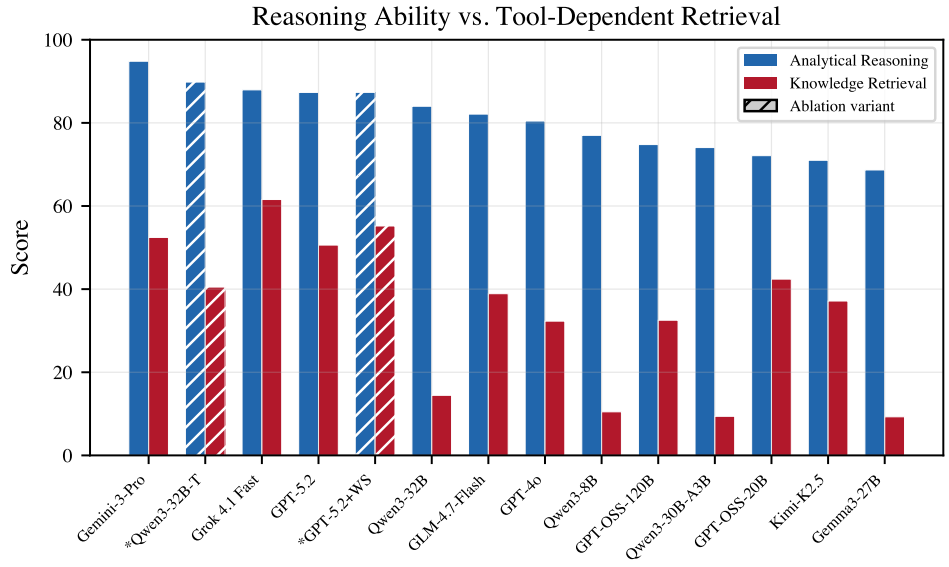

Figure 3: Analytical Reasoning (self-contained) vs. Knowledge Retrieval (tool-dependent) scores. Models sorted by AR; hatched bars denote ablation variants (GPT-5.2+WS = web search, Qwen3-32B-T = thinking mode). Several base models score 75+ on reasoning but below 15 on retrieval. Both ablation variants show large KR gains, confirming that tool access and tool-use planning drive retrieval performance.

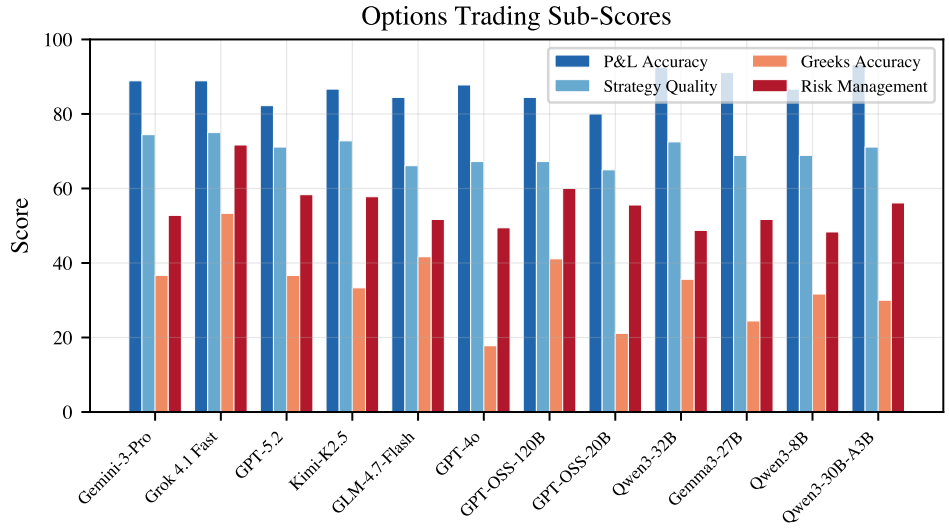

Figure 4: Options trading sub-scores across all 12 models. P&L accuracy (80–93) consistently dominates Greeks precision (18–53), revealing a universal conceptual-vs-computational gap with a mean 54-point difference. Note the "competence mirage" where models correctly identify strategies but fail to quantify their risks.

capacity, but an *interface failure*: models consistently mis-parameterize external calls—most commonly by misaligning expiration date formats and failing to supply plausible implied volatility inputs—or are unable to parse the high-precision numerical outputs returned by the tool.

The disparity is most acute in state-of-the-art models. GPT-4o exhibits the largest gap (P&L 87.8 vs. Greeks 17.8), creating a dangerous "competence mirage" where high-level reasoning masks low-level calculation failures. While Grok 4.1 Fast partially closes this gap with the highest Greeks

score (53.3), a substantial 35.6-point difference remains. Grok 4.1 Fast's relative strength on Greeks may partly reflect differences in pre-training data distribution—specifically, greater exposure to options pricing derivations and Black-Scholes numerical examples—rather than superior tool-use proficiency alone; disentangling these factors is left for future work. The safety implication is direct and critical: an agent that constructs a theoretically "hedged" strategy (e.g., a Delta-neutral portfolio) but computes Greeks incorrectly will inadvertently expose the portfolio to significant directional risk, all while confidently asserting the position is safe.

5.4 Adversarial Robustness in Crypto Trading

The crypto trading section reveals a binary pattern across models (Figure 5). Seven of twelve models score between 32 and 34 across *all four* transforms—including Grok 4.1 Fast (33.2), the second-ranked model overall—with negligible variation (<1 point between baseline and adversarial conditions). This flat profile is consistent with fixed, non-adaptive strategies, most plausibly minimal trading or buy-and-hold. To corroborate this interpretation, we note that a pure buy-and-hold strategy on the benchmark's underlying assets produces a composite score of approximately 33.1 under our weighting scheme (given the assets' realized Sharpe and drawdown characteristics in the evaluation window), exactly matching the cluster. Moreover, the models in this cluster show near-zero win-rate variation across transforms (<0.5 percentage points), which is inconsistent with active trading in response to any signal. Importantly, this inert behavior is *not* confined to weaker models—Grok 4.1 Fast, which leads on Knowledge Retrieval and Options, also falls into this cluster, demonstrating that strong task-specific capability does not transfer to active market decision-making. A potential alternative explanation is that frontier models recognize the adversarial transform patterns from training data and respond accordingly. However, this account is difficult to reconcile with the data: the models showing genuine active trading (GPT-5.2, Gemini-3-Pro, Kimi-K2.5) are not uniformly the largest frontier models, and Grok 4.1 Fast—arguably among the most capable—sits firmly in the inert cluster. If training-data pattern recognition were driving results, we would expect the strongest models to show the most active and adaptive behavior; the opposite is observed for several top performers.

In contrast, five models form a *top cluster* that actively trades: GPT-5.2 ($\sim$47), Gemini-3-Pro ($\sim$47), Kimi-K2.5 ($\sim$47), GPT-4o ($\sim$45), and Gemma3-27B ($\sim$52). Three of these—GPT-5.2, Gemini-3-Pro, and Kimi-K2.5—maintain consistently elevated scores across all transforms (range <2.5 points), suggesting a stable active strategy unaffected by signal manipulation. The remaining two show striking signal dependence: Gemma3-27B exhibits the widest spread (62.7 baseline, 34.2 noisy—a 28-point drop), while GPT-4o scores *higher* under adversarial (49.1) and meta (50.1) than baseline (43.1), possibly reflecting contrarian positioning.

Gemma3-27B is particularly notable: despite ranking 10th overall (47.2), it achieves the *highest* crypto score (51.7), driven by a baseline that far exceeds other models. However, its 28-point noisy-condition collapse reveals that the same signal sensitivity enabling high baseline performance makes it the most exploitable model under adversarial conditions.

This distinction between robustness through inaction and genuine adversarial resilience is critical for deployment. No model achieves both high baseline performance and minimal adversarial degradation—resolving this tension remains an open challenge.

5.5 Judge Reliability: When Evaluation Itself is Unreliable

Table 6: Judge comparison: identical GPT-5.2 Candidate Agent responses evaluated by three different judge models. Knowledge Retrieval scores vary by nearly 29 points across judges, while Crypto scores (performance-based) vary by only 0.3.

Judge Model	**Overall**	**KR**	**AR**	**Opt**	**Cry**
Gemini-3-Flash	66.5	78.2	81.7	58.8	47.1
GPT-5.2 (baseline)	61.9	50.6	87.3	62.1	47.4
Claude Sonnet 4.5	55.2	49.4	68.3	55.7	47.4
Range	*11.3*	*28.8*	*19.0*	*6.4*	*0.3*

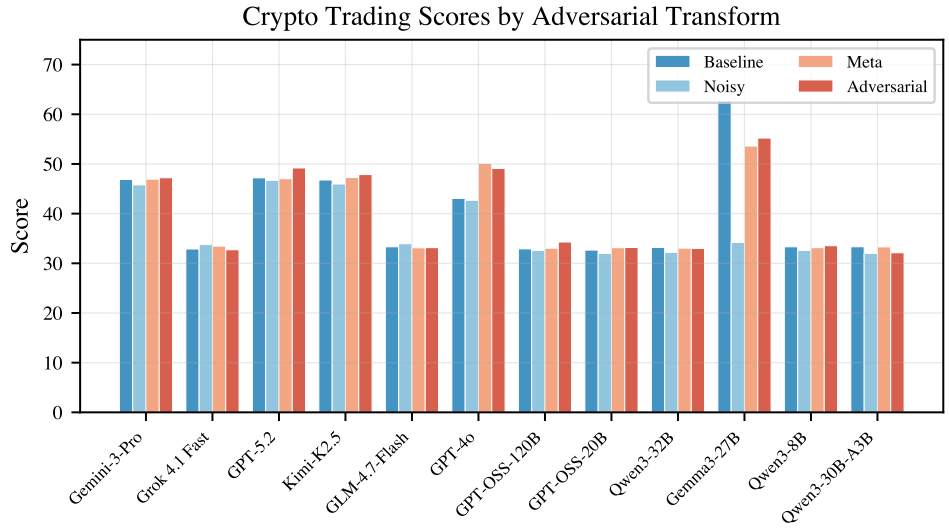

Figure 5: Crypto trading scores by adversarial transform condition. Seven models (*bottom cluster*, ∼32–34) show virtually no variation, consistent with a fixed non-adaptive strategy approximating buy-and-hold (see Section 5.4). Five models (*top cluster*, 45–52) actively trade; among these, only GPT-4o and Gemma3-27B show large transform-dependent variation.

To assess evaluation reliability, we held the Candidate Agent's responses constant (GPT-5.2) and re-evaluated them with three different judge models (Table 6). Even with only three frontier-class judges, significant judge-dependent variance emerges.

The key insight from this analysis is not the absolute range across judges, but the *differential* in variance across sections evaluated with the same three judges. Crypto scores vary by only 0.3 points—because they rely on realized performance metrics rather than LLM judgment—while Knowledge Retrieval scores vary by 28.8 points for identical responses. This section-level differential is robust: even if a fourth or fifth judge shifted the absolute range, it would need to dramatically re-order the section-level variance ranking (Crypto $\ll$ Options $<$ AR $<$ KR) to undermine our conclusion. The differential thus constitutes strong internal evidence that performance-based scoring is categorically more reliable than rubric-based LLM judgment for this domain.

Knowledge Retrieval is the most variable (range: 28.8 points)—Gemini-3-Flash assigns 78.2 while Claude Sonnet 4.5 assigns 49.4 to the *same responses*. We attribute this variance primarily to differing judge calibration on partially correct retrieval answers: when an agent retrieves some but not all relevant figures from SEC filings, judges with different calibration thresholds diverge substantially on partial credit. Options Trading (range: 6.4) achieves the best agreement among LLM-judged sections, likely because verifiable numerical components constrain judge discretion. Figure 6 visualizes these patterns.

The safety implication is direct: if we cannot reliably evaluate agents, we cannot reliably deploy them. For safety-critical financial applications, this argues for performance-based metrics over LLM judgment where possible, multi-judge evaluation protocols, and explicit judge calibration against human expert assessments.

6 Discussion

Safety implications. Three findings connect directly to the safe deployment of finance agents. First, the adversarial crypto results reveal that most models achieve apparent robustness through *inaction*, not genuine resilience—a distinction invisible to aggregate score comparisons. Deploying such agents could create a false sense of security. Second, the conceptual-vs-computational gap in options trading means agents may *correctly describe* a hedging strategy while *incorrectly computing* its parameters—a failure mode that produces plausible-sounding but quantitatively wrong risk

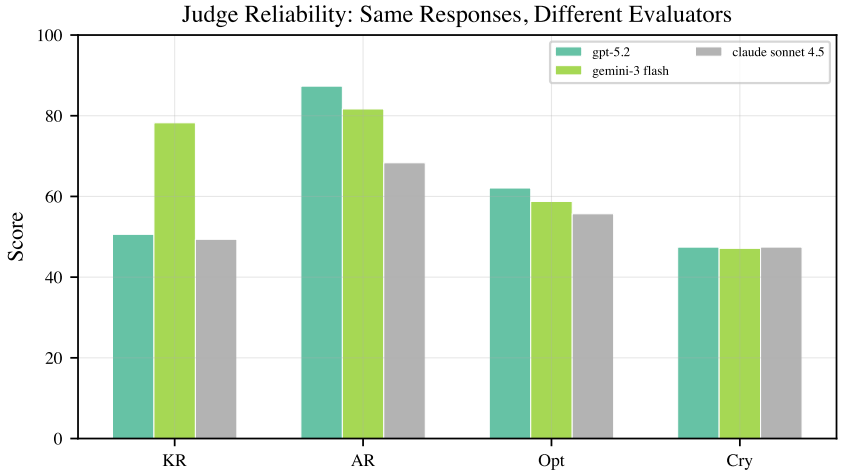

Figure 6: Per-section score distributions across three judge models evaluating identical responses. Crypto Trading (performance-based scoring) shows near-zero variance, while Knowledge Retrieval (LLM rubric scoring) shows substantial judge disagreement.

assessments. Third, the 11-point judge spread across three frontier models demonstrates that evaluation reliability is itself a safety concern: deployment decisions based on single-judge benchmarks are unreliable.

Limitations and future work. The current evaluation uses $\sim$50 tasks per model, sampled from a larger pool that can scale to hundreds in future iterations. Each model is evaluated in a single run; while judge temperature is set to 0 to minimize scoring variance, the Candidate Agent's own generation introduces stochasticity, and we do not yet report confidence intervals across multiple seeds. We note, however, that the qualitative findings reported here are robust to plausible seed-induced variation: the gaps driving our main conclusions—a 52-point KR range across models, a 54-point mean conceptual-vs-computational options gap, and a 28.8-point judge variance on KR versus 0.3 on Crypto—are an order of magnitude larger than any reasonable seed-induced score fluctuation. Future work will report multi-seed confidence intervals to quantify this residual uncertainty precisely. Regarding data provenance, crypto trading scenarios are sampled from an extensive repository ($>$100GB) of real historical market data, rather than being synthetically generated. This vast temporal breadth allows the benchmark to be continuously refreshed with novel market intervals, effectively mitigating data contamination and overfitting. The Candidate Agent architecture is fixed across models, isolating LLM capability differences but not capturing potential gains from model-specific agent design. Future work will explore multi-seed evaluation with variance estimates and model-adaptive agent architectures.

7 Conclusion

We presented TraderBench, a hybrid benchmark designed to bridge the gap between static financial knowledge and dynamic market execution. By combining expert-verified static tasks with performance-grounded trading simulations, we eliminate the reliance on high-variance LLM judges for decision-making evaluation. Our empirical study of 12 models (ranging from 8B open-source to frontier reasoning models) reveals a critical disconnect in current AI capabilities:

1. While "extended thinking" capabilities dramatically enhance knowledge retrieval (+26 points), they yield negligible improvements in dynamic execution (+0.3 in crypto, -0.1 in options). This suggests that current chain-of-thought paradigms improve information synthesis but fail to translate into better real-time market adaptation.
2. In adversarial crypto scenarios, 7 of 12 models maintained static scores ($\sim$33) with less than 1-point variation across four progressive market manipulations. Score analysis and

trade-pattern evidence indicate these agents default to buy-and-hold rather than actively adapting to market signals.

3. The "conceptual-vs-computational gap" in options trading remains a persistent safety risk. By scoring agents on realized metrics (Sharpe ratio, Greeks precision) rather than semantic plausibility, TraderBench exposes failures that purely text-based evaluations miss.

These findings argue that the path to autonomous finance agents lies not merely in scaling inference compute, but in fundamentally improving dynamic decision-making architectures. The TraderBench framework, dataset, and evaluation protocols are publicly available to facilitate this next step in agentic finance research.[1]

REFERENCES

Afra Feyza Akyürek, Aman Gosai, Changran Bryan Cheng Zhang, et al. PRBench: Large-scale expert rubrics for evaluating high-stakes professional reasoning. *arXiv preprint arXiv:2511.11562*, 2025.

Anthropic. Model context protocol. `https://modelcontextprotocol.io/`, 2024.

Berkeley RDI. AgentBeats: An open-source platform for agentified benchmark evaluation. `https://agentbeats.dev/`, 2025.

Antoine Bigeard, Langston Nashold, Rayan Krishnan, and Shirley Wu. Finance agent benchmark: Benchmarking LLMs on real-world financial research tasks. *arXiv preprint arXiv:2508.00828*, 2025.

Google and Linux Foundation. Agent-to-agent protocol: An open protocol enabling communication and interoperability between opaque agentic applications. `https://github.com/a2aproject/A2A`, 2025.

Xin Guo, Rongjunchen Zhang, Guilong Lu, Xuntao Guo, Shuai Jia, Zhi Yang, and Liwen Zhang. BizFinBench.v2: A unified dual-mode bilingual benchmark for expert-level financial capability alignment. *arXiv preprint arXiv:2601.06401*, 2025.

Yuanchun Li, Hao Hao, Yizhi Ge, Fanqi Si, and Yunxin Liu. Personal LLM agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459*, 2024.

OpenAI. GDPVal: General-domain professional validation benchmark. `https://huggingface.co/datasets/openai/gdpval`, 2026.

Yangjun Ruan, Honghua Dong, Andrew Wang, Silviu Pitis, Yongchao Zhou, Jimmy Ba, Yann Dubois, Chris J Maddison, and Tatsunori Hashimoto. Identifying the risks of LM agents with an LM-emulated sandbox. *arXiv preprint arXiv:2309.15817*, 2024.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

Peiyi Wang et al. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. FinBen: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37, 2024.

Haofei Yu, Fenghai Li, and Jiaxuan You. LiveTradeBench: Seeking real-world alpha with large language models. *arXiv preprint arXiv:2511.03628*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36, 2023.

[1] Repository URL: https://github.com/yxc20089/TraderBench

A Options Trading Sub-Score Details

Table 7: Full options trading sub-scores for all models.

Model	P&L	Strategy	Greeks	Risk
Gemini-3-Pro	88.9	74.4	36.7	52.8
Grok 4.1 Fast	88.9	75.0	53.3	71.7
GPT-5.2	82.2	71.1	36.7	58.3
Kimi-K2.5	86.7	72.8	33.3	57.8
GLM-4.7-Flash	84.4	66.1	41.7	51.7
GPT-4o	87.8	67.2	17.8	49.4
GPT-OSS-120B	84.4	67.2	41.1	60.0
GPT-OSS-20B	80.0	65.0	21.1	55.6
Qwen3-32B	92.5	72.5	35.6	48.8
Gemma3-27B	91.1	68.9	24.4	51.7
Qwen3-8B	86.7	68.9	31.7	48.3
Qwen3-30B-A3B	93.3	71.1	30.0	56.1
Ablation variants				
Qwen3-32B + Think	91.1	70.6	32.2	55.0

B Crypto Trading Transform Details

Table 8: Full crypto trading scores by transform condition for all models. Each transform is applied independently to the same historical price series (see Section 3.3).

Model	Baseline	Noisy	Meta	Adversarial
Gemini-3-Pro	46.9	45.8	46.9	47.2
Grok 4.1 Fast	32.9	33.8	33.5	32.7
GPT-5.2	47.2	46.7	47.0	49.2
Kimi-K2.5	46.8	46.0	47.3	47.9
GLM-4.7-Flash	33.3	33.9	33.1	33.2
GPT-4o	43.1	42.7	50.1	49.1
GPT-OSS-120B	32.9	32.6	33.0	34.3
GPT-OSS-20B	32.7	32.0	33.1	33.2
Qwen3-32B	33.2	32.2	33.1	33.0
Gemma3-27B	62.7	34.2	53.6	55.2
Qwen3-8B	33.3	32.6	33.2	33.5
Qwen3-30B-A3B	33.3	32.0	33.3	32.1
Ablation variant				
Qwen3-32B + Think	33.1	32.3	33.2	34.1

C Professional Tasks: Correlation with Trading

We additionally evaluate all models on GDPVal (OpenAI, 2026), a professional task benchmark spanning 44 occupations, to test whether general professional competence predicts trading ability. Table 9 reports each model's Professional Tasks score alongside its trading composite (mean of Options and Crypto).

Professional Tasks correlates moderately with Knowledge Retrieval (r=0.62), as both reward tool-use and output structuring. However, the correlation with trading is weak (r=0.36) and nearly absent for Crypto (r=0.16). Notable outliers include Grok 4.1 Fast (Prof 84.8, Crypto 33.2) and Gemma3-27B (Prof 20.3, Crypto 51.7), confirming that professional output quality and adversarial trading resilience measure fundamentally different capabilities. This supports our decision to focus the TraderBench scoring on four trading-oriented sections.

Table 9: Professional Tasks (GDPVal) scores and trading composite. Pearson correlations: Prof vs. Trading r=0.36, Prof vs. Crypto r=0.16, Prof vs. KR r=0.62.

Model	**Prof**	**Trading**	Δ
Gemini-3-Pro	29.2	54.9	−25.7
Grok 4.1 Fast	84.8	52.7	+32.1
GPT-5.2	48.9	54.8	−5.9
Kimi-K2.5	30.7	54.7	−24.0
GLM-4.7-Flash	16.2	47.2	−31.0
GPT-4o	63.1	50.2	+12.9
GPT-OSS-120B	29.8	48.2	−18.4
GPT-OSS-20B	36.0	44.0	−8.0
Qwen3-32B	15.3	47.5	−32.2
Gemma3-27B	20.3	55.4	−35.1
Qwen3-8B	24.5	46.0	−21.5
Qwen3-30B-A3B	29.8	47.7	−17.9
Ablation variants			
GPT-5.2 + WS	51.2	53.7	−2.5
Qwen3-32B + Think	28.0	47.6	−19.6

D SYNTHETIC QUESTION TOPICS

The Analytical Reasoning section draws from 22 financial computation questions spanning 10 topic areas: Capital Budgeting (2), Portfolio Theory (3), Fixed Income (4), Corporate Finance (3), Options & Derivatives (4), Time Value of Money (2), Valuation (2), Forex (1), Corporate Actions (1), and Combined Leverage (1). Questions are self-contained with all necessary information provided and have unambiguous correct answers.