Table 3: Performance of representative models (Kimi-K2 and GPT-OSS-120B) across different investment target sizes. Results are reported as mean return (% Mean), standard deviation of returns (% Std), and coefficient of variation (CV).

Stocks	% Mean	% Std	CV
Kimi-K2			
5	−4.6	0.7	0.2
10	3.2	0.6	0.2
20	1.9	1.7	0.9
30	−0.5	1.2	2.2
GPT-OSS-120B			
5	−5.7	0.3	0.1
10	2.5	0.4	0.2
20	−0.4	3.9	10.2
30	−0.9	3.9	4.4

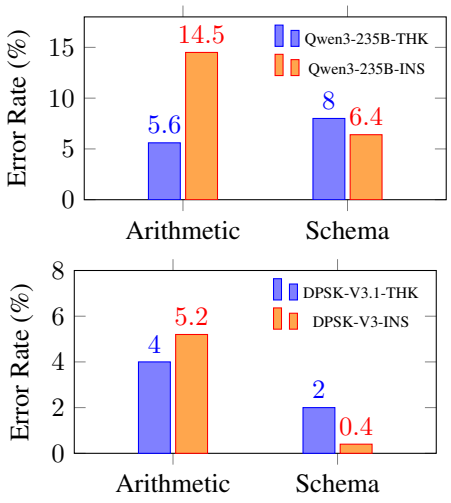

Figure 3: Error distribution (%) by type for Think vs Instruct models.

losses during market downturns. The best-performing agents limit drawdowns to around -11% to -14%, compared to the baseline's -15.2%. **(3) Reasoning model does not guarantee better performance.** Although reasoning-tuned models such as Qwen3-235B-Think and Qwen3-30B-Think exhibit strong performance in tasks requiring complex reasoning, including math and coding (Yang et al., 2025), they do not consistently outperform instruction-tuned counterparts in this trading task. For example, Qwen3-235B-Ins outperforms its reasoning-tuned version with a lower maximum drawdown (-11.2% vs. -14.9%). This suggests there is still a gap between reasoning ability and effective decision-making in dynamic, noisy environments like financial markets.

4 ANALYSIS

4.1 THE INFLUENCE OF INVESTMENT TARGET SIZE

To evaluate the impact of the investment target size on the agent's performance, we conducted the daily trading task with investment targets of 5, 10, 20, and 30 DJIA constituents, repeating the task three times and recording portfolio-weight differences across runs. The results show that variability increases as the investment target expands.

Specifically, as shown in Table 3, **(1) Scalability is inherently challenging.** All evaluated models exhibit performance degradation as the investment portfolio size increases, characterized by declining mean returns and rising return volatility. This indicates that scaling the number of tradable assets poses a non-trivial challenge for LLM agents. **(2) Model scale confers robustness.** The larger-scale model, Kimi-K2, demonstrates greater robustness to portfolio expansion, maintaining relatively stable risk-return profiles and achieving positive expected returns at moderate portfolio sizes (e.g., 10–20 stocks), whereas the smaller GPT-OSS-120B suffers from severe performance deterioration and excessive variability, suggesting that increased model capacity enhances generalization and stability in multi-asset decision-making contexts.

4.2 THE INFLUENCE OF ERROR IN THE TRADING WORKFLOW

During the trading process, various error happened during the agent's interaction with the environment. The most two common errors are: **(1) Arithmetic Error**, where the agent makes mistakes in calculating the number of shares to buy or sell based on the provided budget and stock price. **(2) Schema Error**, where the agent fails to adhere to the specified JSON output format, leading to parsing failures.

Figure 3 illustrates the frequency of these errors across thinking models and instruct models. Specifically, we observe that: Thinking models demonstrate a lower incidence of arithmetic errors compared to instruct models, this observation aligns that thinking models' outstanding performance in reasoning tasks such as math reasoning (Yu et al., 2025; Guo et al., 2025a; Yang et al., 2025). However, as for schema errors, thinking models exhibit a higher frequency of such errors compared to instruct models. This discrepancy aligns with recent findings that reasoning model tend to overthink and produce more complex outputs, which can lead to deviations from the expected format (Fu et al., 2025; Li et al., 2025b).

4.3 Ablation Study on Data Sources

In our workflow, LLM agents rely primarily on two types of information sources: news articles and fundamental financial data. These two modalities provide complementary signals, with news capturing market sentiment and fundamentals grounding the model in key financial indicators. To better understand their respective contributions, we conduct an ablation study by progressively removing these inputs.

Table 4: The cumulative return (CR, %) for Kimi-K2 and GPT-OSS-120B under three input settings: full input (Full), without news articles (w/o News), and without both news and fundamental data (w/o News & Fund.).

Condition	Return (%)
Kimi-K2	1.9
w/o News	1.4
w/o News & Fund.	0.6
GPT-OSS-120B	−1.2
w/o News	−1.2
w/o News & Fund.	−3.4

As shown in Table 4, the cumulative return decreases consistently as we remove news and then fundamental data. This behavior matches our expectation that both information sources play an important role in guiding trading decisions. The Kimi-K2 model remains relatively robust when only news is removed, but its performance deteriorates when both inputs are absent. In contrast, GPT-OSS-120B experiences a sharper decline, indicating that it relies more heavily on explicit signals provided by news and fundamentals. Overall, these findings highlight that LLM-based trading agents are capable of integrating heterogeneous inputs, combining textual information from news with numerical fundamentals to produce more informed and effective trading strategies.

4.4 Impact of Evaluation Window

A good trading model should be able to adapt to changing market conditions over time. To investigate how the choice of evaluation window affects model rankings, we conduct experiments using two different time frames: a downturn period (January to April 2025) and a upturn period (May to August 2025) with Kimi-K2, DeepSeek-series model, GPT-OSS series model and the passive baseline as references. Through this analysis, we aim to understand how models perform under different market regimes and whether their profitability and risk profiles shift accordingly.

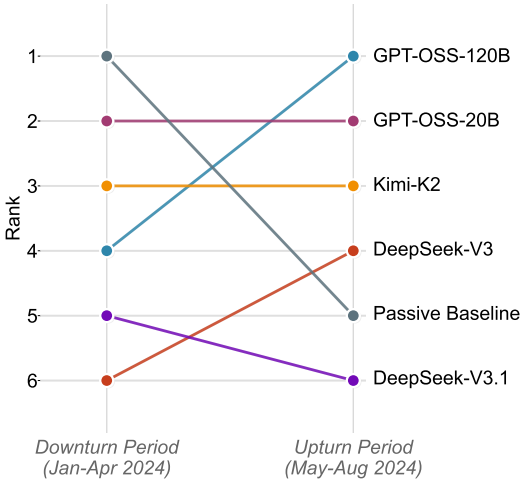

Figure 4: Model performance ranking based on the cumulative return, over two evaluation windows downturn (Jan-Apr 2025) and upturn (May-Aug 2025).

Figure 4 presents the ranking of models based on cumulative return across the two evaluation windows. Notably, we observe significant shifts in model rankings between the downturn and upturn periods. For instance, GPT-OSS-120B, which ranks shift from the bottom during the downturn to the top during the upturn, indicating that it may be better suited to bullish market conditions. While Kimi-K2 maintains a relatively stable ranking across both periods, suggesting its robustness to market fluctuations. This suggests that certain models may be better suited

to specific market conditions, potentially due to their underlying architectures or training data. Besides, we also observe that during the downturn period, all the LLM agents failed to outperform the passive baseline, while in the upturn period, most LLM agents surpass the baseline. This indicates that LLM agents may struggle to navigate bearish markets, highlighting a key area for future improvement.

5 Related Work

5.1 LLM Agents and General Benchmarks

Large language models (LLMs) have rapidly progressed from powerful text completion systems to autonomous agents capable of reasoning, planning, and interacting with external environments (OpenAI, 2024; Anthropic, 2025a; DeepMind, 2025; Liu et al., 2024; Guo et al., 2025a; Meta-AI, 2025; Yang et al., 2024a; Bai et al., 2025; OpenAI, 2025b). There is growing consensus that agentic behavior represents the next stage of LLM development, as it directly connects language understanding with real-world productivity and economic value (OpenAI, 2025a; Anthropic, 2025b). In this paradigm, LLMs are not only evaluated on their static knowledge but also on their ability to continuously perceive, decide, and act.

To capture these emerging capabilities, a variety of benchmarks have been introduced across domains. For example, SWE-Bench (Jimenez et al., 2024) and SWE-Agent (Yang et al., 2024b) target software engineering tasks, GAIA (Mialon et al., 2023) focuses on scientific discovery, and marketing-oriented benchmarks such as XBench (Chen et al., 2025) and Tau2Bench (Barres et al., 2025) examine commercial workflows. These benchmarks highlight the promise of LLM agents for complex, multi-step problem solving and workflow automation. However, despite their breadth, few existing efforts have examined domains where decision-making is directly tied to measurable economic outcomes, such as financial trading.

5.2 Financial Agents and Benchmarks

The financial domain has long been of interest for LLM applications due to its direct link with profitability, risk management, and high-stakes decision making (Wu et al., 2023; Lee et al., 2024; Nie et al., 2024). Most existing benchmarks, however, focus on static question-answering tasks such as FinQA (Chen et al., 2021), TAT-QA (Zhu et al., 2021), and FinBench (Yin et al., 2023). While useful for evaluating financial reasoning and domain knowledge, these tasks do not reflect the iterative, dynamic nature of real-world trading environments.

Recent work has begun to move towards more realistic evaluation settings. For instance, INVESTORBENCH (Li et al., 2025a) introduces an environment for testing trading decisions, marking an important step towards agent-based financial evaluation. However, it primarily considers single-stock-trading and relies on historical data up to 2021, raising concerns about both scope and potential data contamination.

In contrast, our proposed benchmark, StockBench, is the first to embed LLM agents into realistic, multi-stock-trading environments with continuously updated market data. By requiring agents to make sequential trading decisions over extended horizons, StockBench directly evaluates profitability and risk management capabilities. This setting bridges the gap between static financial QA benchmarks and the practical challenges of real-world investment strategies, enabling a more faithful assessment of the readiness of LLM-powered financial agents.

6 Conclusion

In this work, we introduce StockBench, a novel benchmark designed to evaluate the performance of LLM agents in realistic stock-trading scenarios. By simulating dynamic market environments and requiring continuous decision-making over multi-month horizons, StockBench provides a comprehensive framework to assess both profitability and risk management capabilities. Our extensive experiments reveal that while current LLM agents could operate profitably, they still struggle to consistently outperform simple baselines, highlighting the challenges that remain in this domain.