
Time Travel is Cheating: Going Live with DeepFund for Real-Time Fund Investment Benchmarking

Changlun Li^{1,2,*}, Yao Shi^{1,2,*}, Chen Wang, Qiqi Duan, Runke Ruan,
Weijie Huang, Haonan Long, Lijun Huang, Nan Tang^{1,2}, Yuyu Luo^{1,2,†}

¹The Hong Kong University of Science and Technology (Guangzhou)

²Paradox AI Research

Abstract

Large Language Models (LLMs) have demonstrated notable capabilities across financial tasks, including financial report summarization, earnings call transcript analysis, and asset classification. However, their real-world effectiveness in managing complex fund investment remains inadequately assessed. A fundamental limitation of existing benchmarks for evaluating LLM-driven trading strategies is their reliance on historical back-testing, inadvertently enabling LLMs to “time travel” – leveraging future information embedded in their training corpora, thus resulting in possible information leakage and overly optimistic performance estimates. To address this issue, we introduce DeepFund, a live fund benchmark tool designed to rigorously evaluate LLM in real-time market conditions. Utilizing a multi-agent architecture, DeepFund connects directly with real-time stock market data – specifically data published after each model’s pretraining cutoff – to ensure fair and leakage-free evaluations. Empirical tests on nine flagship LLMs from leading global institutions across multiple investment dimensions—including ticker-level analysis, investment decision-making, portfolio management, and risk control—reveal significant practical challenges. Notably, even cutting-edge models such as DeepSeek-V3 and Claude-3.7-Sonnet incur net trading losses within DeepFund real-time evaluation environment, underscoring the present limitations of LLMs for active fund management. Our code is available at <https://github.com/HKUSTDial/DeepFund>.

1 Introduction

The financial industry has witnessed an AI-driven revolution over the past decade [32, 37, 58, 69]. Advanced AI techniques, particularly Large Language Models (LLMs), have transformed practices across multiple domains, including high-frequency trading algorithms [5, 38, 42, 76], risk assessment models [14, 22, 70], investment decisions [31, 33, 63, 66, 68], and data analysis [27, 29, 40, 47, 48, 56], thereby fundamentally reshaping how financial institutions operate and make decisions.

Current financial benchmarks—such as TAT-QA [72], FinTextQA [8], FinBen [64], and Investor-Bench [30]—have made valuable contributions by assessing an LLM’s understanding of financial documents, terminology, and trading performance. For example, these benchmarks typically evaluate LLMs’ effectiveness by simulating trading strategies using historical market data, measuring performance based on metrics such as cumulative returns or risk-adjusted returns. Such evaluations have been widely adopted for both general-purpose foundation models and finance-specific LLMs. **However, a critical gap is that these benchmarks primarily probe static data rather than a model’s ability to make effective investment decisions in real-time market conditions.**

*Both authors contributed equally to this paper

†Yuyu Luo is the corresponding author (yuyuluo@hkust-gz.edu.cn)

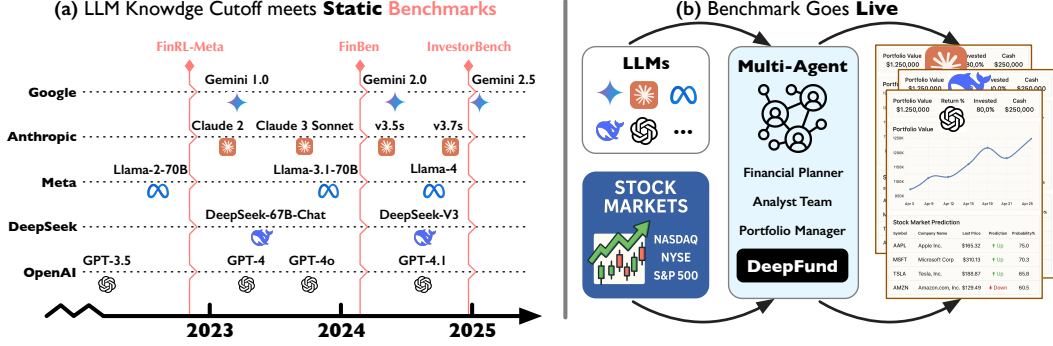


Figure 1: Shifts from static benchmarks to live benchmarks. Particularly for relevant date information in 1(a), we refer to public sources (e.g., model card, arXiv, GitHub) for illustration.

A fundamental limitation in extending these evaluations to **trading performance** lies in their reliance on retrospective back-testing. Back-testing is the standard method for assessing trading strategies [3, 17], but it becomes problematic when applied to LLM-driven strategies because the model may have been pre-trained on the very historical data used for testing, which leads to a severe **information leakage issue** [12, 51, 54, 65]. Undoubtedly, an LLM can appear to perform extraordinarily well on historical market data simply by regurgitating events it has already seen, rather than genuinely predicting outcomes [15, 43, 73, 74]. In other words, the model can effectively “**time travel**” by using future knowledge during evaluation—a form of cheating that inflates its apparent performance. This issue is exacerbated by the varying knowledge cut-off dates of different LLMs.

As shown in Figure 1(a), GPT-4o was trained on data up to October 2023 [19], whereas DeepSeek-V3’s training extends until July 2024 [35]. If we evaluate such a model on a period prior to its knowledge cutoff (e.g., testing DeepSeek-V3 on 2021–2023 data), it will have effectively already seen those market conditions during pre-training, yielding overly optimistic metrics that do not reflect true predictive power.

DeepFund: Fund Benchmark Going Live. To address the critical gap identified above, we introduce DeepFund – a comprehensive framework for real-time fund investment benchmarking, as shown in Figure 1(b). Inspired by previous works [21, 39, 46, 61, 75], DeepFund assesses LLM’s ability to make effective investment decisions in a live-market environment, explicitly preventing any leakage of future data. In particular, our approach offers three key contributions:

(a) **Live Forward Testing:** We introduce a novel benchmarking tool that supports real-time trading conditions to mitigate information leakage. Meanwhile, we provide an interactive web-based interface for performance visualization and comparative analysis on domain-specific financial metrics (e.g., Cumulative Return, Sharpe Ratio) to rigorously assess LLMs’ effectiveness as fund managers.

(b) **Multi-Agent Decision Framework:** We implement a multi-agent architecture in which LLMs assume multiple roles (acting as financial planner, analyst team, and portfolio manager), thereby creating a realistic reproduction of the investment decision-making process. This design mirrors how human analysts and portfolio managers collaborate.

(c) **Empirical Findings:** Through rigorous live environment interaction with various LLMs, we reveal significant performance disparities, highlighting the challenges and possibilities of LLMs in real-time trading, and demystify distinct trading behaviors and personalities exhibited by different LLMs.

2 DeepFund: Multi-Agent Fund Investment Going Live

DeepFund is designed to emulate the dynamics of a real-world fund investment environment, as illustrated in Figure 2. At the top, the **Live Environment** continuously ingests real-time market data, fund asset information, and trading history, ensuring realistic conditions free from information leakage. Below, the **Multi-Agent Workflow** mimics a structured fund management process through three distinct roles: **Financial Planner**, **Analyst Team**, and **Portfolio Manager**. The entire workflow is powered by a single LLM, selected from various providers available in the **LLM Factory** (e.g., Grok as selected in Figure 2), ensuring flexible and consistent backend capabilities across all agents.

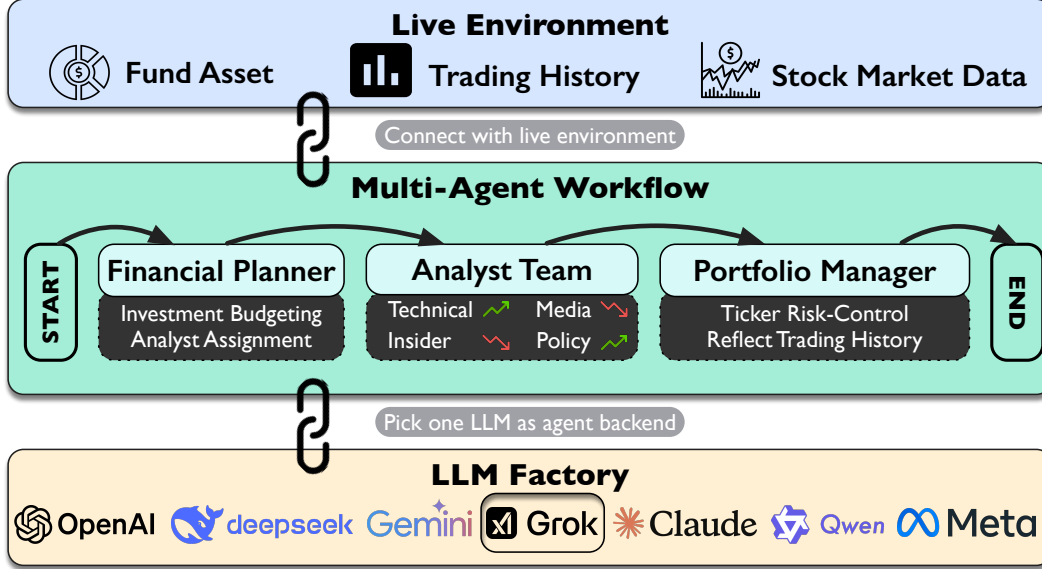


Figure 2: The DeepFund framework.

Live Environment. The live environment serves as the cornerstone of DeepFund, enabling real-time market conditions essential for robust and leakage-free evaluation. Unlike traditional static benchmarks, our environment continuously integrates dynamic financial data streams from three distinct sources: (1) real-time **stock market data**, providing immediate market movements and price fluctuations; (2) up-to-date **fund asset** information, reflecting the current state of investment positions; and (3) detailed **trading history**, recording all activities related to portfolio management.

To facilitate seamless and flexible data ingestion, we implement a **modular API gateway** that interfaces with multiple financial data providers, such as Yahoo Finance and Alpha Vantage. This modularity (refer to Appendix B.5) ensures adaptability to diverse data sources and straightforward integration into varying market contexts. By offering direct feedback from live market conditions, the environment guarantees authenticity in the decision-making process, fundamentally shifting the evaluation paradigm from retrospective back-testing to dynamic, real-time interaction.

Single Agent Design. Similar to previous works [31, 33, 41, 44, 63], each agent in our framework, powered by the selected LLM backend, fulfills a specific role within the investment process:

Financial Planner: Strategically orchestrates the investment analysis by determining analytical priorities and allocating tasks to suitable analysts. It supports two modes: a deterministic mode, allowing predefined analyst selection, and a dynamic mode, leveraging self-reasoning to flexibly select analysts.

Analyst Team: Consists of specialized analyst agents—Fundamental, Technical, Insider, Company News, Macro Economic, and Policy—that analyze domain-specific data and generate standardized signals (Bullish, Bearish, or Neutral), accompanied by detailed justifications. See Table 1 for analyst types and their specialized functions.

Portfolio Manager: Integrates multiple analyst signals to make executive investment decisions (Buy, Sell, Hold), manages risk control (*i.e.*, the portion of holdings and cash), and maintains a dual-memory architecture (see Appendix B.2) to reflect historical transactions and current portfolio states.

Multi-Agent Workflow. The multi-agent workflow [28, 36, 45, 60, 67] adopts an orchestrator-worker paradigm³ to mimic a realistic fund management process. Initially, the **Financial Planner** selects and assigns analysts based on real-time market conditions and portfolio status. Next, the selected **Analyst Team** concurrently processes domain-specific information and generates structured analytical signals. Finally, the **Portfolio Manager** synthesizes these signals, evaluates portfolio risks, decides optimal trading actions, and updates the investment portfolio state. The entire process is rigorously tracked,

³See Anthropic blog <https://www.anthropic.com/engineering/building-effective-agents>

Table 1: Analyst types, their data sources, and specialized features within the Analyst Team.

Analyst Type	Data Source	Feature
Technical	Historical price/volume data	Focuses on price patterns and indicators, such as trends, RSI, volatility, support/resistance.
Fundamental	Financial statements, ratios	Analyzes company financials, such as earnings, margins, valuation metrics.
Insider	Insider transaction reports	Monitors corporate insider activity, such as executive buys/sells, timing patterns.
Company News	News articles, press releases	Assesses company-specific news, such as sentiment, material events.
Macro Economic	Economic indicators	Examines economic conditions, such as GDP, inflation, unemployment, rates.
Policy	Policy news, central bank reports	Analyzes fiscal/monetary policy, such as interest rates, spending, regulation.

Table 2: Evaluated LLMs: The detailed information is sourced from related technical report.

Provider	Model Version	Open Source	Release Date	Knowledge Cutoff
OpenAI	GPT-4.1 [52]	✗	Apr 2025	June 2024
Meta	Llama 4 Scout [50]	✓	Apr 2025	Aug 2024
Google	Gemini 2.5 Flash [16]	✗	Apr 2025	Jan 2025
Anthropic	Claude 3.7 Sonnet [2]	✗	Feb 2025	Oct 2024
xAI	Grok 3 mini Beta [62]	✗	Feb 2025	Nov 2024
DeepSeek	DeepSeek-V3 [11]	✓	Mar 2025	Dec 2024
Alibaba	Qwen2.5-Max [1]	✗	Jan 2025	NA
ByteDance	Doubao-1.5-pro [6]	✗	Jan 2025	NA
Zhipu	GLM-4-Air [71]	✓	Apr 2025	NA

ensuring complete traceability and consistency throughout the decision-making pipeline. We provide more details on how the workflow effectively coordinates information exchange in Appendix B.1.

Evaluation Interface. As shown in Figure 1(b), we provide a web-based interface for presenting the trading performance for each LLM. Inspired by the previous work, ChatBot Arena [9] and Open FinLLM Leaderboard [34], the interface is designed to be comprehensive and fine-grained, allowing for in-depth analysis of the trading behavior. Please refer to Appendix B.3 for more details.

3 Experimental Setting

Financial Data Integration. We integrate upstream data from well-known and trusted financial provider APIs (e.g., Alpha Vantage, Yahoo Finance). The data covers not only granular ticker-level information, such as financial statement, company news, daily trading statistics, and insider transactions, but also the macro indicators and policy news. Parametric settings are presented in Appendix B.4 for more details.

LLMs. We evaluated nine state-of-the-art LLMs from various providers, each with distinct knowledge cutoff dates, to fairly assess their real-time investment performance, as shown in Table 2.

Portfolio Configuration. In a nod to the investment wisdom of Warren Buffett, our experiments target investments in Berkshire Hathaway’s top five holdings as of Q1 2025: Apple (AAPL), American Express (AXP), Bank of America (BAC), Coca-Cola (KO), and Chevron (CVX). Each LLM manages initial cashflow with a total amount of \$100,000. As the *Fundamental* and *Macro-Economic* analysts are designed to provide long-term perspective (i.e., quarterly, half-yearly, or annually), to accommodate the daily trading frequency, each LLM will coordinate the other four analysts: *Technical*, *Company News*, *Policy*, and *Insider* in our experiments.

Trading Period. The trading period is from March 17 to April 17, 2025, covering 24 trading days with a daily trading frequency. Notably, this period captures two significant market events: (1) **FOMC Meeting:** During March 18-19, the Federal Reserve maintained the federal funds rate at 4.5%, marking a second consecutive pause following earlier rate cuts. (2) **Tariff Impact:** During April 2-9, the US government first announced a heavy tariff on global imports, then paused it. This move intensified concerns over inflation and economic slowdown, contributing to fluctuations in major stock indices.

Signal and Decision Validity. The validity of signals and decisions generated by DeepFund was evaluated over 24 trading days. The signal and decision are regarded as valid only if the justification is correctly provided. Failing to do so will result in *No signal provided due to error* or *Just hold due to error*, respectively. Out of a total of 4320 signals and 1080 trading decisions, DeepFund successfully produced **4144 signals (96% validity) and 1059 trading decisions (98% validity)**. Such high validity rates indicate the robustness and reliability in generating timely and actionable outputs. Further detailed statistics are available in Appendix C.

Evaluation Metrics. We employ standard financial metrics to measure performance rigorously, including Cumulative Return (CR) [18], Cumulative Return at Buy & Hold (CR_{bnh}) [13], Sharpe Ratio (SR) [55], Maximum Drawdown (MDD) [49], Win Rate (WR) [7], Beta (β) [23], and Alpha (α) [23]. The detailed definitions and formulas for these metrics can be found in Appendix D.

Implementation Details. We build the agentic workflow via LangChain, a powerful toolkit for building LLM-based applications. All LLM inferences occur through provider-specific APIs, using standardized prompts (see Appendix E) and the same temperature for fairness (refer to Appendix B.4 for more details). Additionally, we utilize Supabase, a PostgreSQL-based cloud database, to store all activities, including historical decisions, portfolio states, and analytical signals. The evaluation incurred approximately \$100 in total costs (*i.e.*, LLM APIs 40%, financial data 40%, and cloud database 20%, correspondingly). To support reproducibility, our code repository is publicly available.

4 Going Live: Revealing the True Trading Power of LLMs

In this section, we delve into the real-world trading performance and behavior of various LLMs deployed in a live market environment. We first present a comprehensive overview of the trading outcomes for all evaluated LLMs. Subsequently, we conduct an in-depth comparative analysis of two representative models: Grok, which uniquely achieved profitability, and DeepSeek, which experienced losses. Through detailed observation of their underlying reasoning processes and decision chains within our trading pipeline, we aim to uncover the critical factors contributing to their divergent performance. Moreover, we provide additional intriguing findings in Appendix A.

This analysis seeks to answer pivotal questions regarding the practical capabilities of LLMs in financial markets: **Q1:** Which LLMs thrive—and which struggle—in the high-stakes arena of live trading? **Q2:** How adeptly can LLMs transform complex, multi-source financial data into precise and actionable trading signals? **Q3:** When the signals are translated into real market actions, do LLMs truly achieve profitable outcomes? **Q4:** What unique trading “personalities” or strategic styles can we uncover among different LLMs in the live market setting?

4.1 Beyond Backtesting: LLMs Face the Live Market Challenge (Q1)

The financial markets are notoriously challenging, often echoing the old adage, “Out of ten gamblers, nine will lose”. Our live trading experiment with LLMs starkly underscores this reality. Table 3 and Figure 3 reveal that while all evaluated LLMs were capable of executing the end-to-end “data-signal-decision” pipeline, their profitability varied dramatically under identical market conditions. A significant majority experienced net trading losses (*i.e.*, Cumulative Return $CR < 0$), highlighting the substantial hurdles of achieving success in real-time fund investment. Strikingly, only the Grok 3 model managed to secure a positive cumulative return.

After the FOMC meeting, all LLMs kept positive net gain. When it came to the tariffs impact, we observe that the DeepSeek incurred the largest drawdown, which is 14.5%. During the bearish market period, Table 3 shows that most US-produced LLMs (except for GPT-4.1) demonstrated lower return

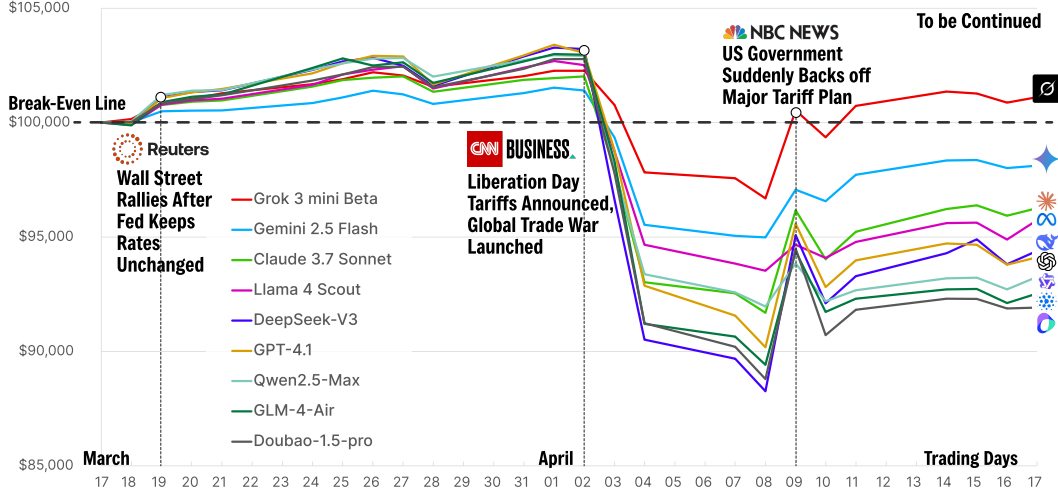


Figure 3: Portfolio asset value for each LLM over time.

Table 3: Overall trading performance of LLMs in DeepFund, sorted by CR ($\downarrow$).

Model Version	$CR(\%)$	$CR_{bnh}(\%)$	SR	MDD (%)	WR (%)	β	α
Grok 3 mini Beta	+1.1	-3.09	0.51	5.5	61	0.42	0.2
Gemini 2.5 Flash	-1.9	-1.58	-1.37	6.4	61	0.35	0.0
Claude 3.7 Sonnet	-3.7	-2.94	-1.45	10.1	70	0.64	0.0
Llama 4 Scout	-4.3	-3.62	-2.42	8.9	61	0.36	-0.1
DeepSeek-V3	-5.7	-5.6	-1.39	14.5	57	0.94	0.0
GPT-4.1	-5.9	-4.41	-1.87	12.8	52	0.77	0.0
Qwen2.5-Max	-6.7	-4.86	-3.12	10.7	65	0.48	-0.2
GLM-4-Air	-7.5	-3.90	-2.31	13.2	57	0.78	-0.1
Doubao-1.5-pro	-8.1	-5.37	-2.35	13.6	65	0.84	-0.1
S&P 500	-6.91	NA	0.3	13.7	NA	1.00	0.0

losses (*i.e.*, CR) than Chinese-produced LLMs. Compared to the CR_{bnh} , we observe that a passive Buy & Hold strategy would have more resilience to such market fluctuations.

The Live Market Gauntlet: Most LLMs Stumble, but Grok Emerges as the Lone Survivor!

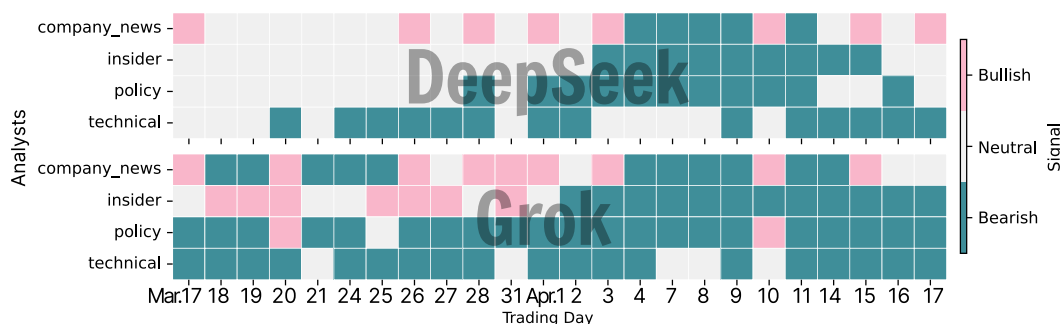
Q1 Takeaway

4.2 Signal or Noise? Decoding LLMs' Analytical Powers (Q2)

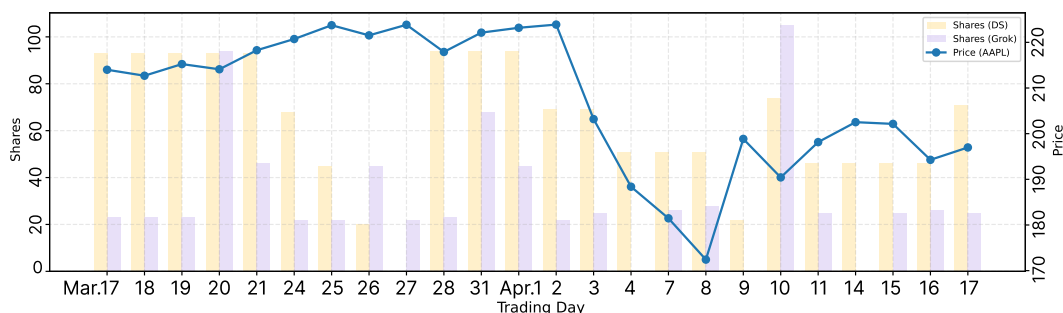
The system employs four functional analysts that generate signals (Bullish, Neutral, or Bearish) based on diverse inputs. The efficacy of an LLM-driven analyst in signal extraction is reflected in the alignment of its aggregated signals with subsequent stock price movements. To assess this, we bind live environment with multi-source data, including company news, insider transactions, policy shifts, and technical indicators. Each analyst is required to generate a signal and justify its reasoning.

Figure 4(a) presents the analyst signal distributions for Apple Inc. (AAPL) powered by Grok and DeepSeek correspondingly. Generally, Grok produced a higher proportion of directional signals (Bullish or Bearish), resulting in greater signal diversity. In contrast, DeepSeek leaned heavily towards Neutral signals given identical conditions.

Specifically, during the period of modest price fluctuation (March 17 to April 2), DeepSeek preferred Neutral signals, suggesting a less sensitive stance. While, Grok preferred Bullish signals. When the tariffs were announced, both turned to pump Bearish signals, indicating a more cautious stance. Notably, both models struggled to predict the significant price surge on April 9 (from 172.42 USD to 198.85 USD), indicating a shared limitation in detecting strong reversal signals.



(a) Daily analyst signal overview by DeepSeek(up) and Grok(down).



(b) Stock price and holdings for DeepSeek (yellow) and Grok (purple) over the trading period.

Figure 4: AAPL trading for DeepSeek and Grok.

Viewing from the perspective of data sources, the behavior revealed further distinctions. Grok-based analysts consistently displayed a bearish perspective in policy and technical analyses, except during tariff-influenced periods. An interesting divergence occurred on April 10 (refer to case study in Appendix F.1 for details), where Grok and DeepSeek held opposing views in policy analysis: both acknowledged short-term uncertainty, but Grok expressed optimism regarding long-term prospects.

Reading the Market's Pulse: Grok Captures Policy Shifts and Technical Trends, DeepSeek Stays Neutral and Misses Key Signals!

Q2 Takeaway

4.3 From Signals to Profits: The Real Test of Trading Decisions (Q3)

In this stage, LLM ingests collective signals, trading history, and holding shares to make trading decisions. We evaluated this signal-to-decision consistency by analyzing how well aggregated signals turned into actions. Intuitively, the "Signal-to-Decision" is regarded as consistent if observed: (i) a Buy followed dominant Bullish signals; (ii) a Sell followed dominant Bearish signals; (iii) a Hold occurred with Neutral or mixed signals without a clear directional bias.

Figure 4(b) shows AAPL's price movements and the trading positions of Grok and DeepSeek. Both models generally integrate signals with the following evidence: (i) Sell decisions are often aligned with Bearish signal from technical and policy analysts; (ii) both models tend to Hold with mixed or predominantly Bearish signals; (iii) insider source has limited impact on Buy and Sell choices.

Particular to the Buy decisions, Grok was more influenced by Bullish signals from company news and policy, while DeepSeek weighted company news more heavily. Although both showed decision-making consistency, Grok demonstrated superior information integration by incorporating policy signals and exhibiting greater decision flexibility, suggesting better market timing capabilities.

Crucially, consistency does not guarantee effectiveness in profitability. We defined an effective decision as a Buy followed by a price increase or a Sell followed by a decrease (excluding the final day). Grok made 11 Buy decisions (7 effective) and 10 Sell decisions (5 effective). Effective decisions are often correlated with Bearish policy/technical or Bullish company news signals. DeepSeek made 3 Buy decisions (1 effective) and 8 Sell decisions (5 effective), showing higher precision in sells but fewer effective buys. Overall, Grok was more effective in leveraging diverse

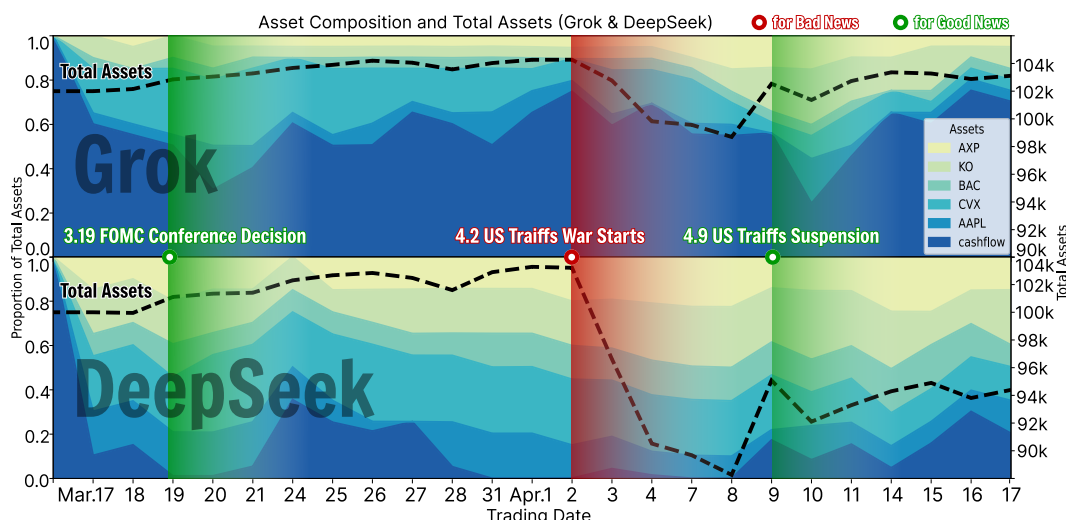


Figure 5: Composition portfolio value for DeepSeek and Grok during the trading period. It shows the holding value of each ticker in the portfolio and the remaining cash flow.

signals, while DeepSeek cautious stance, though precise in sells, limited its ability to capitalize on market opportunities and realize profits.

Turning Signals into Dollars: Grok Masters the Art, While DeepSeek Hesitates at Crucial Moments!

Q3 Takeaway

4.4 Trading Personalities Unveiled: Profiling the LLM Investor (Q4)

Figure 5 illustrates the distinct trading profiles of Grok and DeepSeek through their portfolio compositions and asset evolution during the trading period. Their trading styles diverged significantly since Day One. Grok initially allocated about 40% cash to establish positions, maintaining a relatively 60% high reserve and gradually increasing equity holdings. In stark contrast, DeepSeek aggressively invested nearly 90% of its initial cash, keeping cash levels consistently below 40%, indicating high capital utilization. Subsequently, the dovish remarks made by the FOMC on March 19 brought substantial gains to Deepseek, and this upward trend continued until April 2. However, this low cash reserve severely hampered DeepSeek's flexibility during the market downturn starting April 3, when the US launched a tariff war, impeding timely loss mitigation (refer to case study in Appendix F.2 for details). Grok, with its higher cash position, demonstrated better risk diversification and adaptability. A sufficient cash reserve enables Grok to seize genuine opportunities, significantly increase positions after a sharp decline, and achieve substantial profits during the subsequent rebound after April 9, when US government announced tariffs suspension for most countries.

The models also showed different preferences in sector exposure. Before the period of policy-induced volatility, Grok favored energy (CVX) and consumer staples (KO). DeepSeek, however, concentrated heavily on energy and financial stocks (CVX, BAC, AXP). This lack of sectoral diversification left DeepSeek highly vulnerable to policy shocks without adequate hedging, exacerbating losses during tariff-driven market declines. While Grok mitigated losses by reducing exposure to high-risk assets.

Grok pursued a low-frequency trading strategy with minimal portfolio churn, preferring long-term holdings in what it identified as undervalued blue-chip stocks like KO and CVX. Its maximum drawdown was a mere 3%, reflecting effective risk management via diversification and dynamic rebalancing. DeepSeek, conversely, adopted a high-frequency, momentum-driven approach, frequently adjusting its portfolio to chase short-term fluctuations. While this initially allowed DeepSeek to profit from selling AAPL at a peak, its portfolio concentration and aggressive cash utilization proved detrimental during the later downturn. Operating with low reserves, DeepSeek was forced into unfavorable Buy or Sell positions, preventing loss recovery. Grok's more measured approach, though slower in early profit growth, ensured greater stability and loss minimization.

In essence, Grok embodied a prudent, long-term oriented strategy akin to professional fund management, characterized by risk control and diversification. DeepSeek, conversely, exhibited traits of

a high-frequency retail speculator—concentrated, momentum-driven, and ultimately vulnerable to market shifts, mirroring the common challenges faced by individual traders.

Battle of the Trading Styles: Grok’s Steady Precision vs. DeepSeek’s Bold Gambles—Cash is King in a Bearish Market!

Q4 Takeaway

5 Related Work

Benchmarking LLMs in Financial Domain. With the blossoming research on LLMs in finance [31, 33, 63], numerous benchmarks have been developed to evaluate their capabilities in financial contexts. Financial LLM benchmarks have evolved from document-understanding frameworks like TAT-QA [72], FinanceBench [20], FinTextQA [8], and CFBenchmark [26] to investment decision-making evaluations such as FinRL-Meta [39], FinBen [64] and InvestorBench [30]. These benchmarks share a common focus on assessing LLMs’ financial knowledge and reasoning capabilities, but still face a fundamental challenge that the temporal mismatch between model pre-training data and the evaluation window leads to either information leakage when models are tested on historical data they’ve been trained on, or incomplete evaluation when tested on periods beyond their knowledge boundary. Our work shifts from static evaluation to dynamic evaluation to tackle this challenge.

Live Benchmarking. There has been a growing interest in developing live benchmarks for AI systems to eliminate the “time-travel” problem in recent years. Several works have explored contamination-free evaluation approaches in general domains, such as LiveCodeBench [21], ForecastBench [24], and LiveBench [61]. These benchmarks are designed to evaluate LLMs with in-context learning, code generation, and domain-specific tasks, with regular updates to ensure the benchmark is always up-to-date. Particularly in the financial domain, FinRL-Meta [39] builds benchmarks for reinforcement learning approaches, and ForecastBench [24] explores the evaluation of LLMs’ forecasting capabilities through question-answering that covers market-related questions. With the advancement of multi-agent systems, several works [57, 59] have developed a self-evolving multi-agent framework for dynamic evaluation on LLMs. However, in the specific domain of stock trading and fund investment, existing benchmarks have largely remained static, relying on historical data and back-testing approaches. Our work represents the first work that achieves true live benchmarking for fund investment.

6 Conclusion

DeepFund is a novel benchmarking tool for evaluating and comparing the performance of various LLMs in the context of real-time fund investment. It provides a standardized multi-agent trading workflow with a connection to a live environment and LLM factory. Besides, we have conducted empirical studies to show the effectiveness of our framework and reveal the notable potential of LLMs in fund investment. Overall, DeepFund creates a new paradigm for evaluating LLMs in fund investment, which could contribute to the development of reliable and effective financial AI tools.

7 Limitation

The current implementation highly simplifies the trading context (*i.e.*, US stock market only) and does not account for many practical considerations such as transaction fees, market trading restrictions, and hybrid trading strategies. These details could potentially impact trading performance, but have not been evaluated in our current framework. Therefore, we aim to incorporate index-aligned universes, realistic execution frictions, and broader market conditions in future iterations. Meanwhile, the analysis depth can be further improved by adding more index-based (*e.g.*, relative performance to market indices) and LLM-based (*e.g.*, reasoning cost, consistency, explainability) evaluation metrics. Lastly, the evaluation period was short and occurred during a volatile market, which could skew our results toward specific trading approaches. Conducting longer tests across various circumstances (*e.g.*, bullish, bearish and volatile markets) would manifest more reliable and applicable insights into LLMs’ investment abilities.

8 Broader Impacts

Our work contributes to the understanding and evaluation of LLMs in financial applications, potentially leading to more robust and effective AI-driven investment strategies for research and educational purposes. However, the application of AI in finance carries inherent risks, including the potential for exacerbating market volatility, introducing biases leading to unfair outcomes, or generating financial losses if misused outside of a controlled research setting. We emphasize that DeepFund is intended for academic study and benchmarking, and responsible use is paramount to mitigate these risks.

9 Ethical Statement

The authors take full responsibility for the development of DeepFund, ensuring that the code repository is publicly available and shared under the MIT license, requiring users to adhere to its terms. DeepFund is intended for academic and educational purposes only and is not a substitute for professional advice. While efforts have been made to ensure its accuracy, the authors and their institutions disclaim liability for any outcomes arising from its use. Users agree to take responsibility for ethical and lawful use and to indemnify the authors and their affiliates against any claims or damages resulting from reliance on this material.

10 Acknowledgements

This paper was supported by Young Talent Support Project of Guangzhou Association for Science and Technology (QT-2025-001); the NSF of China (62402409); Guangdong Basic and Applied Basic Research Foundation (2023A1515110545); Guangzhou Basic and Applied Basic Research Foundation (2025A04J3935); Guangzhou-HKUST(GZ) Joint Funding Program (2025A03J3714); and Guangdong Provincial Project (2023CX10X008). Lastly, we genuinely thank the reviewers for their valuable comments and suggestions during the rebuttal period.

References

- [1] Alibaba. Qwen2.5-max: Exploring the intelligence of large-scale moe model. <https://qwenlm.github.io/blog/qwen2.5-max/>, 2025. Accessed: 2025-05-07.
- [2] Anthropic. Claude 3.7 sonnet and claude code. <https://www.anthropic.com/news/claude-3-7-sonnet>, 2025. Accessed: 2025-05-07.
- [3] Robert D. Arnott, Campbell R. Harvey, and Harry M. Markowitz. A backtesting protocol in the era of machine learning. In *The Journal of Financial Data Science*, 2018.
- [4] Richard C Atkinson and Richard M Shiffrin. Human memory: A proposed system and its control processes. In *Psychology of learning and motivation*, volume 2, pages 89–195. Elsevier, 1968.
- [5] Antonio Briola, Jeremy D. Turiel, Riccardo Marcaccioli, and Tomaso Aste. Deep reinforcement learning for active high frequency trading. *ArXiv*, abs/2101.07107, 2021.
- [6] ByteDance. Doubao-1.5 pro. https://seed.bytedance.com/en/special/doubao_1_5_pro, 2025. Accessed: 2025-05-07.
- [7] Ernest P. Chan. *Quantitative Trading: How to Build Your Own Algorithmic Trading Business*. Wiley, Hoboken, NJ, 2 edition, 2021. ISBN 9781119800064.
- [8] Jian Chen, Peilin Zhou, Yining Hua, Loh Xin, Kehui Chen, Ziyuan Li, Bing Zhu, and Junwei Liang. Fintextqa: A dataset for long-form financial question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6025–6047, 2024.
- [9] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

- [10] Nelson Cowan. What are the differences between long-term, short-term, and working memory? *Progress in brain research*, 169:323–338, 2008.
- [11] DeepSeek. Deepseek-v3-0324 release. <https://api-docs.deepseek.com/news/news250325>, 2025. Accessed: 2025-05-07.
- [12] Jasper Dekoninck, Mark Müller, and Martin Vechev. Constat: Performance-based contamination detection in large language models. *Advances in Neural Information Processing Systems*, 37: 92420–92464, 2024.
- [13] Eugene F Fama and Kenneth R French. Permanent and temporary components of stock prices. *Journal of Political Economy*, 96(2):246–273, 1988.
- [14] Paolo Giudici. Fintech risk management: A research challenge for artificial intelligence in finance. *Frontiers in Artificial Intelligence*, 1:1, 2018.
- [15] Paul Glasserman and Caden Lin. Assessing look-ahead bias in stock return predictions generated by gpt sentiment analysis. In *Social Science Research Network*, 2023.
- [16] Google. Start building with gemini 2.5 flash. <https://developers.googleblog.com/en/start-building-with-gemini-25-flash/>, 2025. Accessed: 2025-05-07.
- [17] Campbell R. Harvey and Yan Liu. Backtesting. *The Journal of Portfolio Management*, 2015.
- [18] John Hull. *Risk management and financial institutions*, + *Web Site*, volume 733. John Wiley & Sons, 2012.
- [19] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- [20] Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*, 2023.
- [21] Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*, 2024.
- [22] Haider Ali Javaid. Ai-driven predictive analytics in finance: Transforming risk assessment and decision-making. *Advances in Computer Sciences*, 7(1), 2024.
- [23] Michael C Jensen. The performance of mutual funds in the period 1945–1964. *The Journal of Finance*, 23(2):389–416, 1968.
- [24] Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E Tetlock. Forecastbench: A dynamic benchmark of ai forecasting capabilities. *arXiv preprint arXiv:2409.19839*, 2024.
- [25] Taewoon Kim, Michael Cochez, Vincent François-Lavet, Mark Neerincx, and Piek Vossen. A machine with short-term, episodic, and semantic memory systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 48–56, 2023.
- [26] Yang Lei, Jiangtong Li, Dawei Cheng, Zhijun Ding, and Changjun Jiang. Cfbenchmark: Chinese financial assistant benchmark for large language model. *arXiv preprint arXiv:2311.05812*, 2023.
- [27] Boyan Li, Yuyu Luo, Chengliang Chai, Guoliang Li, and Nan Tang. The dawn of natural language to SQL: are we fully ready? *Proc. VLDB Endow.*, 17(11):3318–3331, 2024.
- [28] Boyan Li, Jiayi Zhang, Ju Fan, Yanwei Xu, Chong Chen, Nan Tang, and Yuyu Luo. Alpha-sql: Zero-shot text-to-sql using monte carlo tree search. *CoRR*, abs/2502.17248, 2025.
- [29] Changlun Li, Chenyu Yang, Yuyu Luo, Ju Fan, and Nan Tang. Weak-to-strong prompts with lightweight-to-powerful llms for high-accuracy, low-cost, and explainable data transformation. *Proc. VLDB Endow.*, 18(8):2371–2384, 2025.

- [30] Haohang Li, Yupeng Cao, Yangyang Yu, Shashidhar Reddy Javaji, Zhiyang Deng, Yueru He, Yuechen Jiang, Zining Zhu, Koduvayur Subbalakshmi, Guojun Xiong, et al. Investor-bench: A benchmark for financial decision-making tasks with llm-based agent. *arXiv preprint arXiv:2412.18174*, 2024.
- [31] Xiangyu Li, Yawen Zeng, Xiaofen Xing, Jin Xu, and Xiangmin Xu. Hedgeagents: A balanced-aware multi-agent financial trading system. *arXiv preprint arXiv:2502.13165*, 2025.
- [32] Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. Large language models in finance: A survey. In *Proceedings of the fourth ACM international conference on AI in finance*, pages 374–382, 2023.
- [33] Yuan Li, Bingqiao Luo, Qian Wang, Nuo Chen, Xu Liu, and Bingsheng He. Cryptotrade: A reflective llm-based agent to guide zero-shot cryptocurrency trading. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1094–1106, 2024.
- [34] Shengyuan Colin Lin, Felix Tian, Keyi Wang, Xingjian Zhao, Jimin Huang, Qianqian Xie, Luca Borella, Matt White, Christina Dan Wang, Kairong Xiao, et al. Open finllm leaderboard: Towards financial ai readiness. *arXiv preprint arXiv:2501.10963*, 2025.
- [35] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.
- [36] Bang Liu, Xinfeng Li, Jiayi Zhang, Jinlin Wang, Tanjin He, Sirui Hong, Hongzhang Liu, Shaokun Zhang, Kaitao Song, Kunlun Zhu, et al. Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems. *arXiv preprint arXiv:2504.01990*, 2025.
- [37] Junhua Liu. A survey of financial ai: Architectures, advances and open challenges. *ArXiv*, abs/2411.12747, 2024.
- [38] Xiao-Yang Liu, Hongyang Yang, Qian Chen, Runjia Zhang, Liuqing Yang, Bowen Xiao, and Christina Dan Wang. Finrl: A deep reinforcement learning library for automated stock trading in quantitative finance. *arXiv preprint arXiv:2011.09607*, 2020.
- [39] Xiao-Yang Liu, Ziyi Xia, Jingyang Rui, Jiechao Gao, Hongyang Yang, Ming Zhu, Christina Wang, Zhaoran Wang, and Jian Guo. Finrl-meta: Market environments and benchmarks for data-driven financial reinforcement learning. *Advances in Neural Information Processing Systems*, 35:1835–1849, 2022.
- [40] Xinyu Liu, Shuyu Shen, Boyan Li, Peixian Ma, Runzhi Jiang, Yuxin Zhang, Ju Fan, Guoliang Li, Nan Tang, and Yuyu Luo. A survey of text-to-sql in the era of llms: Where are we, and where are we going? *IEEE Trans. Knowl. Data Eng.*, 37(10):5735–5754, 2025.
- [41] Xinyu Liu, Shuyu Shen, Boyan Li, Nan Tang, and Yuyu Luo. Nl2sql-bugs: A benchmark for detecting semantic errors in nl2sql translation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, KDD '25*, page 5662–5673, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3737427.
- [42] Yang Liu, Qi Liu, Hongke Zhao, Zhen Pan, and Chuanren Liu. Adaptive quantitative trading: An imitative deep reinforcement learning approach. In *AAAI Conference on Artificial Intelligence*, 2020.
- [43] Alejandro Lopez-Lira, Yuehua Tang, and Mingyin Zhu. The memorization problem: Can we trust llms’ economic forecasts? *arXiv preprint arXiv:2504.14765*, 2025.
- [44] Tianqi Luo, Chuhan Huang, Leixian Shen, Boyan Li, Shuyu Shen, Wei Zeng, Nan Tang, and Yuyu Luo. nvbench 2.0: Resolving ambiguity in text-to-visualization through stepwise reasoning. In *NeurIPS*, 2025.
- [45] Yuyu Luo, Xuedi Qin, Nan Tang, and Guoliang Li. Deepeye: Towards automatic data visualization. In *ICDE*, pages 101–112. IEEE Computer Society, 2018.

- [46] Yuyu Luo, Nan Tang, Guoliang Li, Chengliang Chai, Wenbo Li, and Xuedi Qin. Synthesizing natural language to visualization (NL2VIS) benchmarks from NL2SQL benchmarks. In *SIGMOD Conference*, pages 1235–1247. ACM, 2021.
- [47] Yuyu Luo, Nan Tang, Guoliang Li, Jiawei Tang, Chengliang Chai, and Xuedi Qin. Natural language to visualization by neural machine translation. *IEEE Trans. Vis. Comput. Graph.*, 28(1):217–226, 2022.
- [48] Yuyu Luo, Guoliang Li, Ju Fan, Chengliang Chai, and Nan Tang. Natural language to SQL: state of the art and open problems. *Proc. VLDB Endow.*, 18(12):5466–5471, 2025.
- [49] Malik Magdon-Ismael and Amir F Atiya. Maximum drawdown. *Risk Magazine*, 17(10):99–102, 2004.
- [50] Meta AI. Llama 4. <https://www.llama.com/models/llama-4/>, 2025. Accessed: 2025-05-07.
- [51] Shiwen Ni, Xiangtao Kong, Chengming Li, Xiping Hu, Ruifeng Xu, Jia Zhu, and Min Yang. Training on the benchmark is not all you need. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24948–24956, 2025.
- [52] OpenAI. Introducing gpt-4.1 in the api. <https://openai.com/index/gpt-4-1/>, 2025. Accessed: 2025-05-08.
- [53] Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023.
- [54] Mathieu Ravaut, Bosheng Ding, Fangkai Jiao, Hailin Chen, Xingxuan Li, Ruochen Zhao, Chengwei Qin, Caiming Xiong, and Shafiq Joty. How much are large language models contaminated? a comprehensive survey and the llmsanitize library. *arXiv preprint arXiv:2404.00699*, 2024.
- [55] William F Sharpe. The sharpe ratio. *Journal of portfolio management*, 21(1):49–58, 1994.
- [56] Nan Tang, Chenyu Yang, Ju Fan, Lei Cao, Yuyu Luo, and Alon Y. Halevy. Verifai: Verified generative AI. In *CIDR*. www.cidrdb.org, 2024.
- [57] Xin Tong, Bo Jin, Jingya Wang, Wenpeng Xing, Tian Xia, and Meng Han. Ide: A multi-agent-driven iterative framework for dynamic evaluation of llms. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10890123.
- [58] Saizhuo Wang, Hao Kong, Jiadong Guo, Fengrui Hua, Yiyang Qi, Wanyun Zhou, Jiahao Zheng, Xinyu Wang, Lionel M Ni, and Jian Guo. Quantbench: Benchmarking ai methods for quantitative investment. *arXiv preprint arXiv:2504.18600*, 2025.
- [59] Siyuan Wang, Zhuohan Long, Zhihao Fan, Zhongyu Wei, and Xuanjing Huang. Benchmark self-evolving: A multi-agent framework for dynamic llm evaluation. *arXiv preprint arXiv:2402.11443*, 2024.
- [60] Zora Zhiruo Wang, Jiayuan Mao, Daniel Fried, and Graham Neubig. Agent workflow memory. *arXiv preprint arXiv:2409.07429*, 2024.
- [61] Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddhartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-free LLM benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [62] xAI. Grok 3 beta — the age of reasoning agents. <https://x.ai/news/grok-3>, 2025. Accessed: 2025-05-07.

- [63] Yijia Xiao, Edward Sun, Di Luo, and Wei Wang. Tradingagents: Multi-agents llm financial trading framework. *arXiv preprint arXiv:2412.20138*, 2024.
- [64] 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*, 37:95716–95743, 2024.
- [65] Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. Benchmarking benchmark leakage in large language models. *ArXiv*, abs/2404.18824, 2024.
- [66] Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W Suchow, and Khaldoun Khashanah. Finmem: A performance-enhanced llm trading agent with layered memory and character design. In *Proceedings of the AAAI Symposium Series*, pages 595–597, 2024.
- [67] Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xiong-Hui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. AFlow: Automating agentic workflow generation. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [68] Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, et al. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4314–4325, 2024.
- [69] Huaqin Zhao, Zhengliang Liu, Zihao Wu, Yiwei Li, Tianze Yang, Peng Shu, Shaochen Xu, Haixing Dai, Lin Zhao, Gengchen Mai, et al. Revolutionizing finance with llms: An overview of applications and insights. *arXiv preprint arXiv:2401.11641*, 2024.
- [70] Xiao-lin Zheng, Meng-ying Zhu, Qi-bing Li, Chao-chao Chen, and Yan-chao Tan. Finbrain: when finance meets ai 2.0. *Frontiers of Information Technology & Electronic Engineering*, 20 (7):914–924, 2019.
- [71] Zhipu. Glm-4-air. <https://open.bigmodel.cn/dev/releasenotes/Release-notes>, 2025. Accessed: 2025-05-07.
- [72] Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat seng Chua. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *ArXiv*, abs/2105.07624, 2021.
- [73] Yizhang Zhu, Shiyin Du, Boyan Li, Yuyu Luo, and Nan Tang. Are large language models good statisticians? In *NeurIPS*, 2024.
- [74] Yizhang Zhu, Runzhi Jiang, Boyan Li, Nan Tang, and Yuyu Luo. Elliesql: Cost-efficient text-to-sql with complexity-aware routing, 2025.
- [75] Zhenzhen Zhuang, Jiandong Chen, Hongfeng Xu, Yuwen Jiang, and Jialiang Lin. Large language models for automated scholarly paper review: A survey. *Information Fusion*, page 103332, 2025.
- [76] Chuqiao Zong, Chaojie Wang, Molei Qin, Lei Feng, Xinrun Wang, and Bo An. Macrohft: Memory augmented context-aware reinforcement learning on high frequency trading. In *Knowledge Discovery and Data Mining*, 2024.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [\[Yes\]](#)

Justification: Please refer to the abstract and introduction of the paper.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [\[Yes\]](#)

Justification: Please refer to the Section 7.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [\[NA\]](#)

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [\[Yes\]](#)

Justification: We have released the code repository with detailed instructions to ensure the reproducibility of the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We have released the code repository with detailed documentation.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Please refer to the Section 3.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Our experiments involve forward-testing on real-time stock trading of which represents a single path realization, making traditional statistical significance tests and error bars non-standard and difficult to define. We report performance metrics and findings in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer “Yes” if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: Our experiment is not involved with huge computing resources.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [Yes]

Justification: We have reviewed the NeurIPS Code of Ethics and stated in Section 9.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Please refer to Section 8.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not include models that have a high risk for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have properly credited and mentioned the license and terms of use in code repository.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [\[Yes\]](#)

Justification: We have released the code repository with detailed documentation.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [\[NA\]](#)

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [\[NA\]](#)

Justification: The paper does not involve research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are a core component of our research methodology, which focuses on benchmarking and evaluating their performance in real-time fund investment.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

A Additional Experiments

A.1 Trading Performance by Market Sector

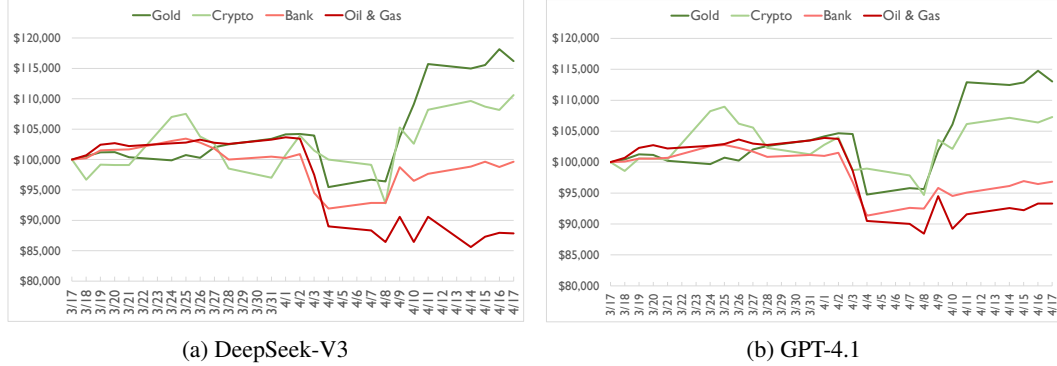


Figure 6: Trading performance by market sector. The green and red line indicate a profit gain and loss, respectively.

In this experiment, we evaluate the investment performance of GPT-4.1 and DeepSeek-V3 in the following four sectors with relevant tickers: **Gold** (NEM, GLD, AEM, GFI, KGC), **Oil & Gas** (XOM, CVX, NFG, CRGY), **Crypto** (COIN, MSTR, MARA), **Banking** (JPM, BAC, WFC, C, RY).

The results are presented in Figure 6 with the same trading period in main experiments. Overall, GPT-4.1 exhibits stable growth with low volatility, suitable for conservative strategies. In contrast, DeepSeek-V3 shows high return potential but greater fluctuations, suitable for more aggressive investors. Specifically, the Gold and Crypto markets are profitable sectors, while the Oil & Gas and Banking markets suffer losses. In the profitable sectors, DeepSeek-V3 shows higher returns and a stronger growth potential than GPT-4.1. On the other hand, GPT-4.1 has fewer losses in the Oil & Gas industry, demonstrating better risk control, while DeepSeek-V3 manifests fewer losses in the Banking sector, indicating better resilience against market downturns.

A.2 The Cost Efficiency on OpenAI Family

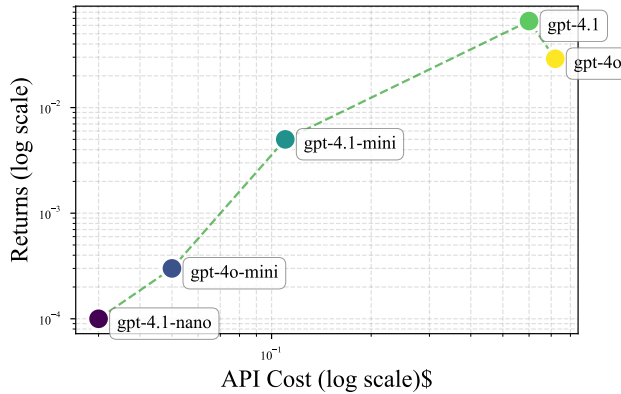


Figure 7: Cost-performance tradeoff analysis for OpenAI family models.

In the mid April 2025, OpenAI released a new series of GPT-4.1 models [52]. In this experiment, we explore the cost-efficiency of the OpenAI family models in our framework. As shown in Figure 7, we evaluate it using two key metrics: LLM API inference cost and trading cumulative return, respectively in x-axis and y-axis. We adopt a shorter trading period from 2025-04-07 to 2025-04-21 on ticker portfolio of AAPL, AXP, BAC, KO, AMT. We recognize gpt-4.1-mini as the best cost-efficiency (*i.e.*, moderate return per dollar spent) choice for most fund applications. For high-frequency trading, we recommend gpt-4.1 as it provides the highest return per dollar spent.

A.3 Extended Trading Performance to entire Q2 2025

To address the concern on the narrow evaluation window, we expanded our live trading window to cover the entire Q2 2025 (see Table 4). Each LLM continues to actively manage its portfolio and surf in the market, with Grok-3 still maintaining its leading profitability. Some LLMs such as GPT 4.1, Claude 3.7, and DeepSeek V3 achieved net profits eventually, while other LLMs such as GLM 4, Qwen Max, and Gemini 2.5 still incur losses. Results from such a longer period remain consistent with our original findings, which further supports the robustness of our conclusions.

Table 4: Averaged Weekly Asset Total Value by LLM in Q2 2025. Dates are the end of each week.

Date	GPT 4.1	Claude 3.7	Grok 3	Llama 4	Gemini 2.5	DeepSeek V3	Qwen Max	Doubao 1.5	GLM 4
2025-04-06	100207.56	99370.47	101031.59	100200.65	99819.65	99305.63	99973.54	99387.25	99629.81
2025-04-13	92824.93	93933.56	98978.32	94181.13	96280.95	91682.35	92652.86	91204.98	91695.78
<i>Below are new results</i>									
2025-04-20	94322.82	96202.42	101156.52	95466.44	98212.7	94351.44	93099.94	92098.54	92523.98
2025-04-27	94993.84	97186.15	101527.47	96768.25	98205.38	96133.87	94084.57	93030.82	93893.19
2025-05-04	96627.35	98242.37	102151.96	98289.12	98349.17	97987.88	94765.69	94182.54	95209.34
2025-05-11	96247.15	98319.67	102227.45	97770.66	97374.24	98223.28	94376.73	93827.1	95528.87
2025-05-18	98815.25	100726.09	103425.42	102377.7	98323.0	101482.47	95853.75	97527.84	98038.4
2025-05-25	97701.93	99340.52	103385.97	100680.71	97822.53	99767.74	94142.15	96939.64	96392.79
2025-06-01	98589.52	99419.7	103625.78	100703.5	98473.83	99969.97	94540.49	97865.15	96501.95
2025-06-08	99292.85	99914.3	103972.52	101252.86	98746.37	100703.84	94868.62	98584.24	96893.33
2025-06-15	99942.06	100360.74	104263.5	101935.87	99192.8	101416.86	95380.01	99262.71	97060.48
2025-06-22	99470.03	99668.47	104700.82	101698.64	98919.57	100873.6	95216.37	98985.76	96356.1
2025-06-29	101442.76	101234.08	106262.44	102859.48	99762.68	102851.0	96383.33	100865.84	97670.72

B Technical Details

B.1 Data Models and Operational Schemas

To facilitate the message communication and information flow throughout the system, we implement hierarchical data models and operational schemas that standardize agent interactions. These elements ensure consistent information processing while maintaining semantic richness across the multi-agent framework.

Our core data models encapsulate domain-specific financial information, not limited to the following examples: **MediaNews** are normalized containers for company-specific news, press releases, and policy updates; **Insiders** are formalized tracking of insider transactions, executive changes, and corporate governance events; **Fundamentals** are standardized models of financial statements, valuation metrics, and growth indicators; **OHLCV** metrics represent the daily trading statistics, contributing to the calculation of technical indicators. All data models are designed to be the information upstream of the specialised agents.

Apart from data models, we also define a set of operational schemas that govern system behavior: **Signal**: Structured output from analyst agents containing direction (Bullish, Bearish, Neutral) and detailed justification; **Decision**: Formalized investment actions (Buy, Sell, Hold), number of shares, price that day, and reasonings; **Portfolio**: Comprehensive state representation tracking cashflow, and holding positions with corresponding shares and risk exposure. Notably, all of the above schemas are encapsulated into a unified object, **FundState**, which is a system-wise message container that persists the current state of the fund.

If the Policy and Fundamental analysts are selected, they will receive upstream **MediaNews** and **Fundamentals** as input, and output corresponding **Signal** object. Consequently, the portfolio manager will analyse based on those signals and current holding positions, and output a **Decision** object. Finally, the **Portfolio** object will be updated according to the **Decision** object. Eventually, this model-schema governance enables both specialized analysis and integrated decision-making while maintaining strict data consistency and provenance tracking throughout the system.

B.2 Memory Management

Memory management is crucial for maintaining context and learning in our multi-agent system [36, 60]. We implement a dual-memory architecture that combines short-term operational memory with long-term historical memory to enable both immediate decision-making and continuous learning.

Short-term Memory. The primary short-term memory in our system is implemented through the **FundState** object, which serves as a thread-scoped memory container [25, 53]. This stateful object maintains the current operational context of the fund, encompassing current portfolio positions and cash balance, recent trading decisions and their rationales, active signals from analyst agents, and the latest market data and news context. The **FundState** is updated in real-time as the system processes new information and makes decisions, ensuring all agents have access to the most recent operational context. This short-term memory is essential for maintaining consistency across agent interactions and enabling coherent decision-making within a single trading session.

Long-term Memory. Our system maintains long-term memory through comprehensive trading history records from live market interactions [4, 10]. This historical memory serves as a foundation for performance tracking and analysis of trading strategies, enabling the system to learn from past decisions and their outcomes. Through pattern recognition across different market conditions, the system continuously improves its decision-making capabilities. This persistent memory layer helps the system adapt and improve its performance over time by incorporating lessons from previous trading sessions.

The combination of short-term operational memory through **FundState** and long-term historical memory creates a robust memory architecture that supports both immediate decision-making and continuous system improvement. This dual-memory approach enables our multi-agent system to maintain context awareness while learning from past experiences.

B.3 Evaluation Interface

The interface serves as the first AI Live Investment Arena, designed to evaluate the trading and investment capabilities of various Large Language Models (LLMs) across financial markets. It provides a comprehensive environment for assessing how LLMs can ingest financial information, drive multi-agent systems, and make trading decisions in real-world market scenarios. You can explore its features further by visiting <https://deepfund.paradoox.ai/>. The main features are:

Table 5: Key features of the DeepFund valuation interface.

Feature Name	Description
Performance Leaderboard	A competitive ranking system comparing LLM models based on investment metrics (total returns, daily returns, portfolio values).
Interactive Data Visualization	Charts displaying cumulative returns over time with adjustable periods for detailed performance analysis.
Portfolio Analysis	Detailed breakdown of each LLM agent’s portfolio, including holdings, asset allocation, and value distribution.
Market Comparison	Direct comparison between LLM performance and major market indices (NASDAQ, S&P 500, DOW JONES).
Agent Lab	An environment for users to customize and develop their own LLM trading agents to compete.
Reports Section	Provides analytical reports on performance and market trends.

B.4 Parameter Settings

Table 6: DeepFund parameter settings.

Parameter	Default Value	Usage
LLM temperature	0.5	Control the randomness of the LLM inference.
Retry times	3	Number of retries for LLM inference.
Technical window	100	OHLCV covered days for technical analysis.
Insider count	10	Insider transactions for insider analysis.
Number of news	10	Company news and policy analysis.
Decision memory size	5	Number of past recent actions for decision-making.

B.5 Systematic Scalability

To ensure long-term viability and extensibility, DeepFund implements a modular architecture that decouples core functionalities into components. This design enables seamless integration of diverse LLMs, data sources, and agent composition without architectural modifications. **LLM module** provides a unified interface that abstracts provider-specific implementations, enabling fair comparison and rapid integration of new LLMs. **Data source module** implements a similar abstraction for financial information, standardizing diverse sources into consistent internal data models. **Agent composition module** enables community extension through a well-defined protocol for adding specialized analytical methodologies. Thus, researchers can contribute novel analytical approaches that seamlessly integrate with the existing system. This interface architecture transforms DeepFund from a fixed evaluation platform into an evolving research ecosystem, providing standardized benchmarking while supporting continuous incorporation of advances in LLM technology, financial data analysis, and agent system design.

C Signal and Decision Statistics

As shown in Figures 8 and 9, the various LLMs exhibit distinct decision-making behaviors and analytical signal processing patterns. The correlation between overt trading behaviors (i.e., buy, sell, and hold decisions) and the corresponding signal data offers valuable dimensions for a more nuanced analysis and characterization of the distinct trading styles inherent to each model.

Decision Distribution. We observe that distinct models employ varied signal processing mechanisms and decision-making frameworks, resulting in a spectrum of observable trading strategies. For instance, models characterized by a high proportion of hold recommendations exhibit discernible differences in their underlying signal information profiles when contrasted with models demonstrating more frequent buy or sell activities.

Signal Distribution. These models display heterogeneous sensitivities to diverse market signals, with certain signal categories evidently assuming a predominant role in the strategic outputs of particular models. This variance suggests the implementation of sophisticated and potentially model-specific internal weighting systems for signal aggregation and interpretation.

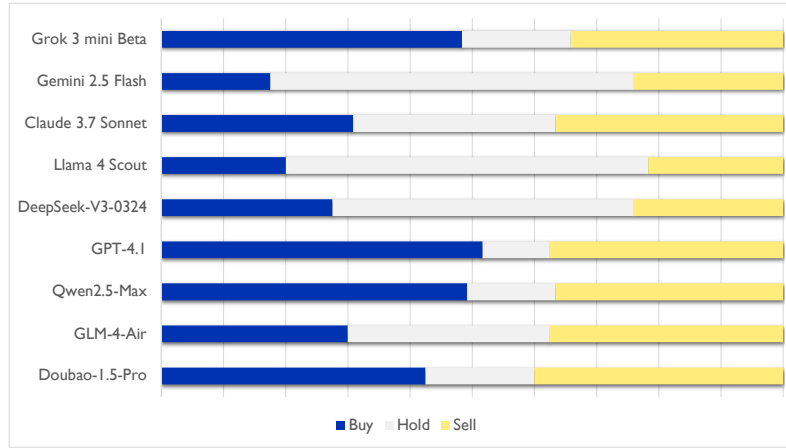


Figure 8: Decision statistics by LLMs across all tickers.

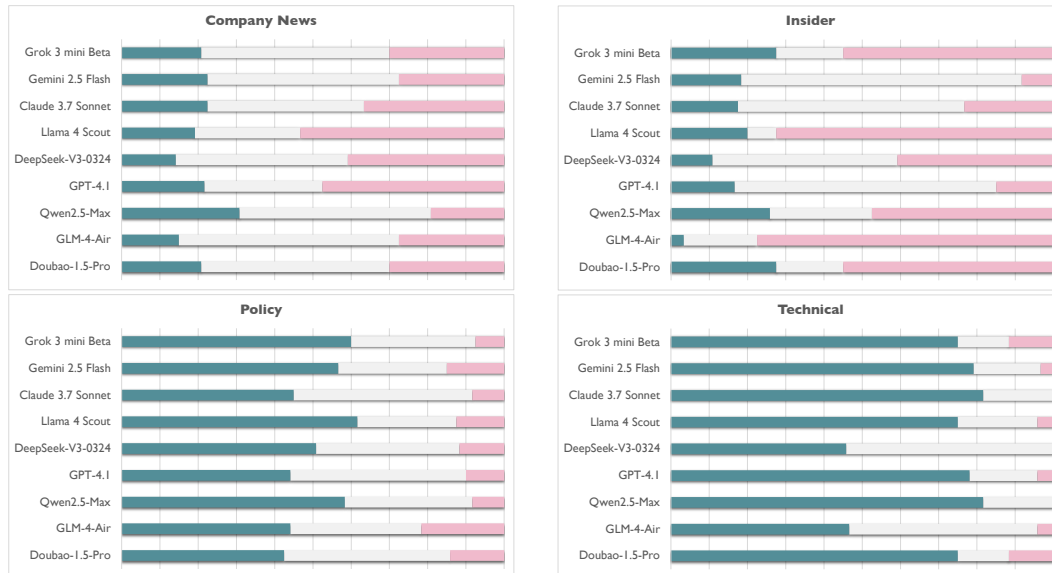


Figure 9: Analytical signal statistics by LLMs across all tickers.

D Evaluation Metrics

In this section, we provide detailed definitions and formulas for the evaluation metrics used in our experiments.

Cumulative Return (CR). [18] Total percentage gain or loss from the initial investment:

$$CR(\%) = \left(\frac{P_{\text{final}}}{P_{\text{initial}}} - 1 \right) \times 100 \quad (1)$$

where P_{initial} and P_{final} denote the initial and final portfolio values, respectively.

Cumulative Return at Buy & Hold ($CR_{b\&h}$). [13] Buy & Hold is a passive investment approach, where an investor purchases stocks and holds onto them for an extended period regardless of market fluctuations. We harness this strategy as an alternative investigation to evaluate the performance of LLMs. The portfolio is initialized by LLM since day 1 and held until the end of the test period.

Sharpe Ratio (SR) [55] Excess return divided by its standard deviation, using the risk-free rate based on the 1-month US Treasury bill (4.29% as of 2025-04-17):

$$SR = \frac{\bar{r}_e}{\sigma_{r_e}} \times \sqrt{252} \quad (2)$$

where $\bar{r}_e$ is the average daily excess return ($r_s - r_f$), σ_{r_e} is the standard deviation of excess returns, and 252 is the number of trading days in a year.

Maximum Drawdown (MDD) [49] Largest percentage decline from peak portfolio value, indicating downside risk:

$$MDD(\%) = \max_{t \in [0, T]} \left(\max_{s \in [0, t]} \frac{P_s - P_t}{P_s} \right) \times 100 \quad (3)$$

where P_t is the portfolio value at time t , and T is the total investment horizon.

Win Rate (WR) [7] Percentage of profitable trades executed:

$$WR(\%) = \frac{\sum_{i=1}^n \mathbf{1}_{r_i > 0}}{n} \times 100 \quad (4)$$

where r_i is the return of the i -th trade, $\mathbf{1}_{r_i > 0}$ is the indicator function that returns 1 if the trade was profitable, and n is the total number of trades.

Beta (β) [23] Portfolio volatility relative to the S&P 500:

$$\beta = \frac{\text{Cov}(r_s, r_m)}{\text{Var}(r_m)} \quad (5)$$

where r_s is the return of the strategy, r_m is the return of the market (S&P 500), $\text{Cov}(\cdot)$ denotes covariance, and $\text{Var}(\cdot)$ denotes variance.

Alpha (α) [23] Excess return compared to the market benchmark (S&P 500):

$$\alpha = r_s - [r_f + \beta(r_m - r_f)] \quad (6)$$

where r_s is the strategy return, r_f is the risk-free rate (4.29% as of 2025-04-17), r_m is the market return, and β is the strategy's beta as defined above.

E Prompt Template

As we committed to open source our code repository, the full details are documented in `src/llm/prompt.py` under our project repository <https://github.com/HKUSTDial/DeepFund>. Here we provide the prompt template of technical analyst and portfolio manager for illustration.

E.1 Technical Analyst

You are a technical analyst evaluating ticker using multiple technical analysis strategies. The following signals have been generated from our analysis:

Price Trend Analysis: {trend}
Mean Reversion: {mean_reversion}
RSI: {rsi}
Volatility: {volatility}
Volume Analysis: {volume}
Support and Resistance Levels: {price_levels}

You must provide your analysis as a structured output with the following fields:

- signal: One of ["Bullish", "Bearish", "Neutral"]
- justification: A brief explanation of your analysis

Your response should be well-reasoned and consider all aspects of the analysis.

Specifically, the technical analyst process the OHLCV data to analyse price patterns, momentum indicators, and statistical trends.

E.2 Portfolio Manager

You are a portfolio manager making final trading decisions based on decision memory, and the provided optimal position ratio.

Here is the decision memory: {decision memory}

Current Price: {current price}
Holding Shares: {current shares}
Tradable Shares: {tradable shares}

If the value of tradable shares is positive, you can buy more shares.
If the value of tradable shares is negative, you can sell some shares.
If the value of tradable shares is close to 0, you can hold.

You must provide your decision as a structured output with the following fields:

- action: One of ["Buy", "Sell", "Hold"]
- shares: Number of shares to buy or sell, set 0 for hold
- price: The current price of the ticker
- justification: A brief explanation of your decision

Your response should be well-reasoned and consider all aspects of the analysis.

Specifically, the decision memory is retrieved from trading history. The current price represents the latest price of the ticker, holding shares represent the shares currently held based on the current fund state. The tradable shares represent the shares that can be bought or sold, which is derived from the risk control module.

F Case Studies

F.1 Divergent Signals from Policy News on April 10th

Fiscal View:

- Beijing white paper outlines China trade stance before retaliatory strike on US.
- ‘Brace for impact’: Chinese economist warns the gloves are off in US trade war.
- United Airlines, Microchip Technology lead the stocks posting 20%+ reversals on tariff pause.

Monetary View:

- Bond Market Twists Tariff Arm to 90 Day Pause.
- Citigroup Vs Wells Fargo: Which Bank Stock is a Smarter Investment?
- Here’s How to Play Goldman Stock Ahead of Its Q1 Earnings Release.

Grok:

Signal: Bullish

Justification: The news highlights a tariff pause leading to stock surges, increased investor confidence, and positive market reactions in stocks and crypto, outweighing short-term uncertainties from trade tensions, suggesting potential economic growth over the next 6 months.

DeepSeek:

Signal: Bearish

Justification: The fiscal policy news highlights escalating trade tensions between China and the US, including retaliatory measures and warnings from Chinese economists. This uncertainty could negatively impact global trade and economic stability in the short term. Additionally, while some positive developments like tariff pauses and bank stock surges are noted, the overall tone suggests heightened risks and volatility.

F.2 Decision Discrepancy on April 3rd

Trading Facts:

- Ticker: Chevron Corp (NYSE: CVX)
- Price of the Day: 156.12
- Historical Prices In The Past Four Days: [166.47, 168.51, 167.29, 166.06]

Grok:

Decisions In The Past Four Days: [Hold, Hold, Hold, Hold]

Holding Shares: 184, Cashflow: 4252.06, Total Asset: 96622.39

Action: Sell 27 shares

Justification: The tradable shares value is -27.0, indicating an opportunity to sell shares to align the portfolio with the optimal position ratio. The current price of 156.12 is lower than recent historical prices, but selling the excess shares will help rebalance the portfolio.

DeepSeek:

Decisions In The Past Four Days: [Buy 1 share, Sell 61 shares, Sell 1 share, Buy 30 shares]

Holding Shares: 184, Cashflow: 59941.02, Total Asset: 100768.53

Action: Buy 68 shares

Justification: Tradable shares are positive at 68.0, indicating an opportunity to buy more shares based on the optimal position ratio.