INVESTORBENCH: A Benchmark for Financial Decision-Making Tasks with LLM-based Agents

Haohang Li[1,⋆], Yupeng Cao[1,⋆], Yangyang Yu[1,⋆], Shashidhar Reddy Javaji[1], Zhiyang Deng[1], Yueru He[2], Yuechen Jiang[1], Zining Zhu[1], Koduvayur Subbalakshmi[1], Jimin Huang[4], Lingfei Qian[4], Xueqing Peng [4], Qianqian Xie[4,†], Jordan W. Suchow[1]

[1]Stevens Institute of Technology [2]Columbia University
[3]Harvard University [4]The Fin AI
[⋆]Equal Contribution [†] Corresponding author: qianqian.xie@yale.edu

Abstract

Recent advancements have underscored the potential of large language model (LLM)-based agents in financial decision-making. Despite this progress, the field currently encounters two main challenges: (1) the lack of a comprehensive LLM agent framework adaptable to a variety of financial tasks, and (2) the absence of standardized benchmarks and consistent datasets for assessing agent performance. To tackle these issues, we introduce INVESTORBENCH, the first benchmark specifically designed for evaluating LLM-based agents in diverse financial decision-making contexts. INVESTORBENCH enhances the versatility of LLM-enabled agents by providing a comprehensive suite of tasks applicable to different financial products, including single equities like stocks, cryptocurrencies and exchange-traded funds (ETFs). Additionally, we assess the reasoning and decision-making capabilities of our agent framework using thirteen different LLMs as backbone models, across various market environments and tasks. Furthermore, we have curated a diverse collection of open-source, multimodal datasets and developed a comprehensive suite of environments for financial decision-making. This establishes a highly accessible platform for evaluating financial agents' performance across various scenarios. The code is available at Github Repo: https://github.com/felis33/INVESTOR-BENCH

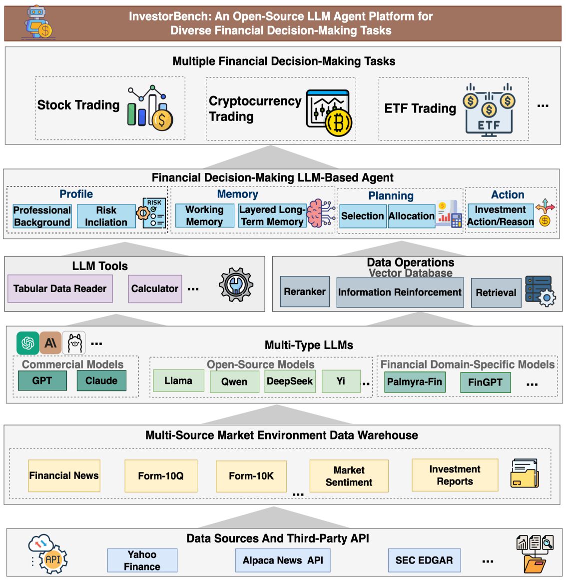


Figure 1: General architecture of INVESTORBENCH.

1 Introduction

The recent studies on large language model (LLM)-based agents have demonstrated impressive performance across a range of decision-making tasks in complex and open-ended environments spanning various domains (Zhang et al., 2024b; Guo et al., 2024; Eigner and Händler, 2024; Wang et al., 2024). However, developing agentic frameworks tailored specifically for financial decision-making remains a significant challenge. This complexity arises from the need for agents to acutely discern and prioritize decisive signals, and then make sequentially high-quality decisions within the volatile and multifaceted financial markets, where information varies in time sensitivity and modality.

Furthermore, the design of financial agents becomes increasingly complex when applied across multiple decision-making tasks, due to the significant variation in key factors influencing financial decisions across different objectives and task types. For instance, single-equity tasks like stock trading require analyzing company-specific and industry-wide data, including market metrics, sector trends, performance reports, and relevant news (Yi et al., 2022). In contrast, cryptocurrency trading is highly sensitive to crypto-specific news and sentiment due to its dynamic nature (Bhatnagar et al., 2023). ETFs, on the other hand, typically follow passive investment strategies, emphasizing long-term growth and cost efficiency (Madhavan, 2016).

The recent emergence of financial LLM-based

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2509–2525
July 27 - August 1, 2025

agent frameworks such as FinMem (Yu et al., 2024a), FinAgent (Zhang et al., 2024a), CryptoTrade (Li et al., 2024), FinRobot (Yang et al., 2024), and FinCon (Yu et al., 2024b) has presented a variety of architectural approaches tailored to specific financial tasks. This diversification has sparked substantial interest across both academic and industrial landscapes. FinRobot is engineered specifically for market analysis, while FinMem and FinAgent are oriented towards trading individual equities like stocks and ETFs. CryptoTrade focuses solely on cryptocurrency trading. FinCon pioneers in addressing portfolio management, although it currently handles only compact portfolios consisting of three stock assets. While these frameworks are effective within their respective niches, they generally focus on addressing only limited types of financial decision-making tasks. This restricts them from further demonstrating the broader applicability of these frameworks and limits the comprehensive, comparative insights that could be drawn from their overall decision-making performance. Furthermore, the frequent reliance on proprietary financial data complicates the evaluation of these tools, obscuring their effectiveness and adaptability in broader contexts. Therefore, there is a pressing need to develop innovative benchmarks specifically designed to evaluate LLM-based agents across a wider spectrum of financial decision-making scenarios. Such benchmarks would enable a more robust assessment of these technologies, facilitating advancements that could cater to various financial applications.

We introduce InvestorBench, an open-source, LLM-based agent benchmark that generalizes across a broad range of financial decision-making tasks. Its detailed structure is illustrated in Figure 1. Further developed upon the foundational framework of FinMem (Yu et al., 2024a), which focuses on single-stock investment decisions, our benchmark extends the scope to encompass an **ensemble of diverse financial market environments** for various financial tasks. InvestorBench's cognitive architecture, similar to FinMem, employs a **layered memory processing** mechanism with distinct decay rates, enabling the agent to store, retrieve, and consolidate insights and reflections more effectively than the pure similarity-based memory retrieval used in FinAgent. This approach ensures that decisions are informed by timely and impactful data, a capability previously shown effective for single-asset trading. These features reflect how human traders draw sequential decisions upon investment signals from multiple sources and varying time sensitivities, allowing the agent to naturally adapt to complex financial tasks. InvestorBench expands its evaluation beyond the original stock trading tasks to encompass three decision tasks significant in the realm of financial investment: **stock trading**, **cryptocurrency trading**, and **ETF investing**.

In summary, we make three key contributions: 1) We establish InvestorBench, an **innovative and comprehensive financial agentic benchmark** designed to evaluate the reasoning and sequential decision-making capabilities of LLM-based agents in complex, open-ended financial scenarios. This benchmark provides a realistic perspective for assessing the design and performance of such agents. 2) We provide a **set of open-source, multi-source market environments** that closely mirror real-world conditions. Furthermore, these environments also serve as a standardized platform for evaluating the decision-making performance of other LLM-based financial agents. 3) We present a **unified, flexible language-agent framework** that allows finance professionals to conveniently customize assess any LLMs serving as the agents reasoning core. In this paper, we conduct a holistic evaluation of 13 LLMs including recent, competitive, and domain-specific fine-tuned models (see Table 1) to provide a broad overview of their reasoning capabilities in sequential decision-making tasks within financial contexts.

2 LLM Trading Agents

In this section, we define a framework of the LLM-based agents in the InvestorBench and formalize the financial decision-making tasks within the context of partially observable Markov decision process (POMDP) (Bertsekas and Shreve, 1996; Liu et al., 2020; Kabbani and Duman, 2022).

2.1 Definition

The LLM-based agent in InvestorBench is structured as a large language model-modulo framework, designed to match or surpass the capabilities of professional human investors. This framework consists of several interconnected modules, each tailored to handle the distinct challenges presented by the financial markets volatility and complexity: **Brain/Backbone (LLM):** This module, which is the LLM itself, serves as the core of the LLM-based agent. It enhances the agent's capabilities by enabling it to understand, process, and generate natural language. This module plays a crucial role in supporting complex decision-making processes, offering interpretations of market-related information, generating predictive analytics, and reflecting on past investment decisions.

Perception: This module serves a critical function by converting raw market data into a structured format that is compatible with the LLM, specifying what the agent perceives and observes, which includes numerical, textual, and visual information.

Profile: This module serves two functions articulated in natural language. Firstly, it describes the agent's role, highlighting its character as an experienced investor with expert-level knowledge and a self-adaptive risk preference. This risk preference dynamically adjusts based on historical market momentum, allowing the agent to optimize its strategies in real time. Secondly, the module provides a detailed background of the decision-making task, specifying the key characteristics and pertinent information about the target assets involved in the trading decisions, such as equity historical performance, price fluctuations, and sector information. This dual-function module supports the agent's decisions with both the current market context and its historical performance.

Memory: This module processes and retains essential market data and historical insights, allowing the agent to draw on a rich repository of knowledge for decision-making. Building upon the pioneering work of Yu et al. (2024a) in FINMEM, the memory architecture comprises two primary components: **Working Memory** and **Layered Long-Term Memory**, as depicted in Figure 2.

Working memory: This component maintains FINMEM's original functionalities: *observation, summarization, and reflection*. It incorporates two reflection mechanisms: *immediate* and *extended*. Immediate reflection produces the agent's reasoning outcomes by integrating current market indicators with the top-K ranked events from each long-term memory layer, which are significant during both warm-up and evaluation stages. In the warm-up stage, the emphasis shifts as the trading direction is predetermined, focusing on understanding market trends and enhancing predictive accuracy. In the evaluation stage, it outputs the trading direction (Buy, Sell, or Hold), the rationale for this decision, identifying the most influential memory events and their respective IDs from each layer.

Layered Long-Term Memory: Inspired by the human cognitive system's varying information decay speeds, Layered Long-Term Memory component structures financial insights across multiple layers. Each layer is represented by a vector database in the Long-Term Memory data warehouse, where information is prioritized and purged based on a specific decay rate. Deeper layers retain information longer with smaller decay rates, while shallower layers, dealing with more transient data, have larger decay rates. This tiered approach is critical as it allows the adaptation of the memory architecture to a broader range of financial tasks beyond single-asset decisions, accommodating an expanded variety of data sources and increasing overall system flexibility. Detailed mechanisms for ranking and decay in each layer are further elaborated in the Appendix A.

Action: This module executes trading and investment decisions based on the analysis provided by other modules. It directly outputs {*"Buy", "Sell", "Hold"*} for traded asset (stock, crypto, or ETF), as instructed by the backbone LLM. Action module synthesizes the operational outcomes from the Profile and Memory modules to facilitate precise and well-informed investment decisions. For its daily trading operations, the agent can choose from three specific actions for the traded asset: "Buy", "Sell", or "Hold". The functionality and input requirements of this module differ significantly between the warm-up and evaluation stages: *during the warm-up stage*, the agent observes daily adjusted price differences between consecutive days, which are critical for identifying potential "Buy" or "Sell" signals. This period allows the agent to calibrate and adjust its decision-making strategies based on near-term market movements; *during the evaluation stage*, access to future price data is restricted, compelling the agent to rely solely on available historical data and its cognitive processing capabilities. In response to trading inquiries, the module integrates historical Profit & Loss (PnL), outcomes from extended reflections, and the top-K retrieved memories. This integration ensures that each trading decision is grounded in a comprehensive analysis of past performance and current market conditions.

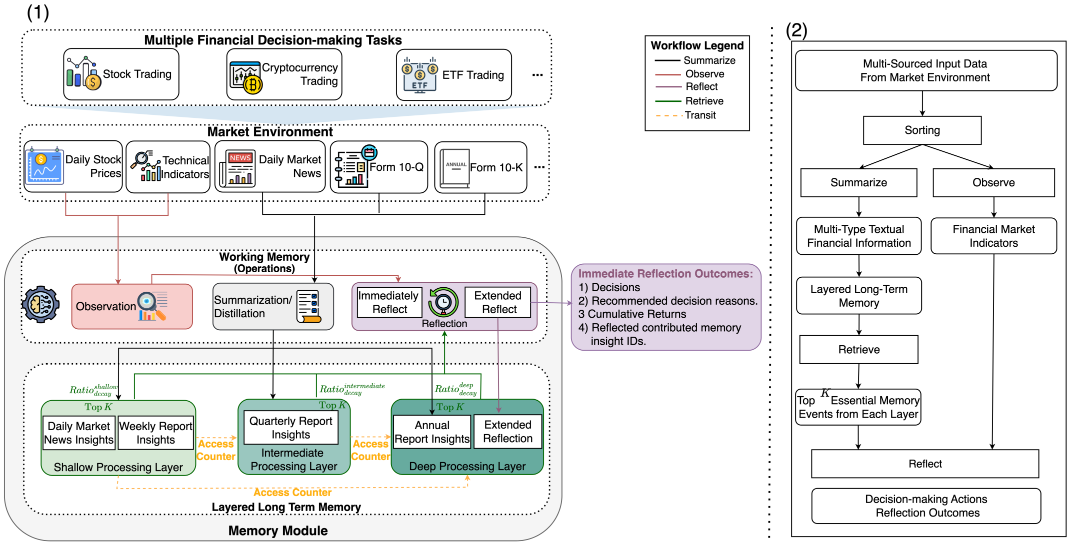


Figure 2: (1) The language agent's memory module is crafted to interact with the market environment to conduct various financial decision-making tasks. It contains two core components – Working Memory and Layered Long-term Memory. (2) The outline of the agent's decision-making workflow for retrieving critical memory events and market observations to inform specific investment decisions.

2.2 Modeling financial decision-making

Formally, we model a financial decision-making process as infinite horizon POMDP with time index $\mathbb{T} = \{0, 1, 2, \cdots\}$ and discount factor $\alpha \in (0, 1]$. This POMDP contains: (1) a state space $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ is the observable component and $\mathcal{Y}$ is unobservable component of the financial market; (2) the action space of the agent is $\mathcal{A}$, which is modeled as {*"Buy", "Sell", "Hold"*}; (3) the reward function $\mathcal{R}(o, b, a) : \mathcal{X} \times \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$ uses daily profit & loss (PnL) as the output; (4) the observation process $\{O_t\}_{t\in\mathbb{T}} \subseteq \mathcal{X}$ is a multi-dimensional process (5) the reflection process $\{B_t\}_{t\in\mathbb{T}} \subseteq \mathcal{Y}$ represents the agent's self-reflection, which is updated from B_t to B_{t+1} on daily basis (Griffiths et al., 2023); (6) the action $A_t \sim \pi(\cdot|\text{prompt})$ represents the way to make investment decision driven by the language conditioned policy π. By denoting daily profit & loss (PnLs) by $R_t^\pi = \mathcal{R}(O_t, B_t, A_t)$ and the set of all admissible language conditioned policies as $\Pi = \{\pi(\cdot|\text{prompt})\}$, the optimization objective of financial decision-making task is then:

$$\max_{\pi\in\Pi} \mathbb{E}\left[\sum_{t\in\mathbb{T}} \alpha^t R_t^\pi\right] \quad (1)$$

3 InvestorBench

He we introduce the detailed architecture of InvestorBench, as illustrated in Figure 1.

3.1 Benchmark Composition

INVESTORBENCH is organized into four main components: (1) **Data Sources and Market Environments**: INVESTORBENCH utilizes a wide range of open-source data and incorporates third-party APIs, such as Yahoo Finance and SEC EDGAR, to create a comprehensive, multi-modal market environment data warehouse. (2) **LLM Agent**: INVESTORBENCH includes an advanced LLM-based agent equipped with modules for Brain, Perception, Profile, Memory, and Action. This agent is enhanced with external tools (such as tabular data readers and API callers) and data operations (including vector database management, information reinforcement, and retrieval). (3) **Financial Decision-Making Tasks**: INVESTORBENCH offers three distinct financial decision-making tasks, differentiated by their asset types. (4) **Evaluation Metrics**: The efficacy of all tasks within INVESTORBENCH is evaluated using a set of standard metrics in the quantitative finance field, providing a thorough evaluation of the decision-making capabilities of the LLM-based agent.

3.2 Trading Environments

We release three datasets, each curated from diverse sources, to construct tailored financial market environments for specific tasks. Our objective is to address the current gap in evaluation environments for financial decision-making agent frameworks and to offer a fully open platform for the comprehensive assessment of agents across various tasks. Below, we introduce each environment, categorized by task type, detailing its scope and the data sources it incorporates.

Stock market environment integrates information from multiple sources, including: 1) Daily stock open, high, low, close, and volume (OHLCV) data acquired from Yahoo Finance. 2) Summarized insights from company quarterly and annual reports (Form 10-Q and 10-K) downloaded from the SEC EDGAR database. 3) News articles for seven stocks collected daily between 2020-07-01, and 2021-05-06. The news data for four of these companiesMicrosoft Corporation (MSFT), Johnson & Johnson (JNJ), UVV Corporation (UVV), and Honeywell International Inc. (HON)-are randomly selected from the pool with the most new records (over five hundred) from the open-access dataset provided by Zhou et al. (Zhou et al., 2021), while the news data for the remaining three companiesTesla, Inc. (TSLA), Apple Inc. (AAPL), and NIO Inc. (NIO)-are obtained from Refinitiv Real-Time News, which primarily contains high-quality news information from Reuters. 4) The sentiment categories (*'positive', 'negative', 'neutral'*) assigned to each news record are generated by gpt-3.5-turbo-0125.

Cryptocurrency market environment encompasses 1) the daily stock open-high-low-close-volume (OHLCV) acquired from CoinMarketCap; 2) the multisource cryptocurrency news data collected from cryptonews, cryptopotato, and cointelegraph(Vanhoucke, 2023); 3) news spanning from 2023-02-13 to 2023-11-05 collected by (Zhou et al., 2021) in daily frequency. 4) The sentiment categories generated by the same means.

ETF market environment is constructed using News-Informed Financial Trend Yield (NIFTY) dataset (Saqur et al., 2024). It contains the processed and curated daily news headlines from 2019-07-29 to 2020-09-21 and generated sentiment categories for each news headline.

In experimental use, we divide the dataset according to the date, with the train set used for the warmup phase to establish the memory database, and the test set used for the test phase to evaluate the model performance.

Table 1: INVESTORBENCH evaluates 13 proprietary or open-source LLMs on financial decision-making tasks.

Model	#Size	Form	Ver.	Model	#Size	Form	Ver.
gpt-4(Achiam et al., 2023)	N/A	api	0613	Qwen2.5-7b(Qwen team, 2024)	7B	open	Instruct
gpt-4o(OpenAI, 2022)	N/A	api	0806	Qwen2.5-32b(Qwen team, 2024)	32B	open	Instruct
gpt-o1-preview	N/A	api	0912	Qwen2.5-72b(Qwen team, 2024)	72B	open	Instruct
DeepSeek-v2(Xin et al., 2024)	15B	open	Lite	llama3.1-8b(Llama team, 2024)	8B	open	Instruct
DeepSeek-llm(Xin et al., 2024)	67B	open	Chat	llama3.1-70b(Llama team, 2024)	70B	open	Instruct
Yi-1.5-9b(Young et al., 2024)	9B	open	Chat	Palmyra-Fin(team, 2024)	70B	open	32K
Yi-1.5-34b(Young et al., 2024)	34B	open	Chat				

3.3 Evaluation metrics

We employ four widely recognized financial metrics to evaluate and compare the investment performance of various LLMs serving as backbones across different tasks: : Cumulative Return (CR) (Hull, 2007), Sharpe Ratio (SR) (Sharpe, 1994), Annualized Volatility (AV) (Cochrane, 1988), and Maximum Drawdown(MDD) (Ang and Chen, 2003). Note that CR and the SR are often considered more essential than AV and MDD in evaluating asset trading performance due to their focus on long-term gains and risk-adjusted returns by their definition. Here, we regard these two metrics as primary metrics when evaluating the experiment outcomes. The detailed explanation is in Appendix B.

4 Experiment and Discussion

To establish a baseline and assess the performance of LLM agents, we standardize experimental settings and evaluation metrics across various financial decision-making tasks. Results are presented on a task-by-task basis. We report the performance of INVESTORBENCH on three single-asset trading tasks: ***stocks***, ***cryptocurrencies***, and ***ETFs*** trading, using closed-source, open-source, and domain-specific LLMs.

4.1 Experiment Setup

Table 1 summarizes the performance of a comprehensive list of trading agents. For single equity tasks, the baseline is set up by Buy and Hold strategy, while for portfolio management task, it is set up by an equal-weight portfolio with the detailed rational explained in Appendix. In our experiments, the temperature parameter of all LLM-based agent systems is set at 0.6 to balance response consistency and reasoning creativity. The performance metrics are reported for the test trajectory with the median CR, SR, AV, and MDD from five repeated epochs. (If the median of these metrics does not belong to the same epoch, the performance is based on the trajectory with the median SR.)

Furthermore, the selection of warm-up and test periods differs across various tasks due to the vary-

Table 2: Performance of stock trading with different LLMs as backbone model across seven stocks.

Model	MSFT				JNJ				UVV				HON			
	CR↑	SR↑	AV↓	MDD↓	CR↑	SR↑	AV↓	MDD↓	CR↑	SR↑	AV↓	MDD↓	CR↑	SR↑	AV↓	MDD↓
Buy & Hold	15.340	0.717	17.236	9.428	13.895	0.927	12.075	9.847	36.583	1.457	20.216	15.406	33.256	1.619	16.537	9.195
Financial Domain Models																
Palmyra-Fin-70B	14.697	0.618	18.987	9.428	5.748	0.311	13.329	9.367	37.875	1.407	21.528	15.967	20.016	1.010	15.852	6.824
Proprietary Models																
GPT-o1-preview	17.184	0.664	20.700	9.428	13.561	0.749	14.396	9.847	41.508	1.481	22.411	9.633	13.162	0.535	19.673	11.558
GPT-4	16.654	0.643	20.715	9.428	13.712	0.761	14.417	9.860	31.791	1.132	22.471	10.434	34.342	1.383	19.858	9.195
GPT-4o	12.461	0.638	15.631	6.647	9.099	0.604	12.055	7.169	8.043	0.342	18.796	14.889	38.540	1.668	18.480	8.979
Open-Source Models																
Qwen2.5-72B-Instruct	7.421	0.406	14.654	6.973	14.353	0.787	14.487	9.812	37.178	1.257	23.614	13.365	34.309	1.380	19.858	9.292
Llama-3.1-70B-Instruct	17.396	0.921	15.105	7.045	13.868	0.773	14.338	9.825	35.981	1.192	24.140	15.406	43.944	1.826	19.253	8.993
DeepSeek-67B-Chat	13.941	0.575	19.376	7.850	14.426	0.818	14.111	9.825	29.940	1.022	23.435	15.407	32.536	1.317	19.753	10.782
Yi-1.5-34B-Chat	22.093	0.865	20.433	9.428	14.004	0.814	13.757	9.847	20.889	0.704	23.748	14.936	30.743	1.258	19.551	9.195
Qwen2.5-32B-Instruct	-0.557	-0.028	15.796	8.946	2.905	0.201	11.540	7.169	-1.623	-0.067	19.301	17.986	26.332	1.366	15.420	5.261
DeepSeek-V2-Lite (15.7B)	11.904	0.479	19.869	16.094	-7.482	-0.462	12.953	17.806	33.560	1.175	22.838	12.984	16.686	0.672	19.852	16.806
Yi-1.5-9B-Chat	19.333	0.755	20.486	9.428	18.606	1.112	13.392	10.986	49.415	1.663	23.768	11.430	29.028	1.173	19.791	12.588
Llama-3.1-8B-Instruct	22.703	0.912	19.910	7.385	13.988	1.025	14.117	9.969	41.108	1.367	24.057	16.429	39.079	1.601	19.526	10.341
Qwen-2.5-Instruct-7B	-10.305	-0.500	16.517	23.371	21.852	0.676	25.823	9.573	11.752	0.588	15.862	15.451	4.291	0.197	17.204	14.156

Model	TSLA				AAPL				NIO				Average			
	CR↑	SR↑	AV↓	MDD↓	CR↑	SR↑	AV↓	MDD↓	CR↑	SR↑	AV↓	MDD↓	CR↑	SR↑	AV↓	MDD↓
Buy & Hold	39.244	0.600	52.339	37.975	10.837	0.324	26.899	19.119	52.216	0.592	73.966	47.766	34.099	0.505	51.068	34.953
Financial Domain Models																
Palmyra-Fin-70B	-6.661	-0.153	34.761	25.820	8.562	0.256	26.835	25.466	-3.261	-0.039	70.181	58.406	-0.453	0.021	43.925	36.564
Proprietary Models																
GPT-o1-preview	34.499	0.549	50.247	35.490	8.238	0.291	22.801	14.412	32.433	0.385	70.704	54.016	25.057	0.408	47.918	34.639
GPT-4	45.246	0.821	44.088	25.031	9.889	0.304	26.226	19.119	75.952	0.887	71.801	37.867	43.696	0.671	47.371	27.339
GPT-4o	45.946	0.930	39.524	21.631	7.405	0.315	18.929	12.824	63.743	0.909	58.795	29.220	39.031	0.718	39.083	21.225
Open-Source Models																
Qwen2.5-72B-Instruct	39.112	0.742	42.184	26.985	11.935	0.395	24.352	19.119	87.412	1.505	48.733	12.464	46.153	0.880	38.423	19.523
Llama-3.1-70B-Instruct	37.545	0.615	48.862	29.813	12.772	0.402	25.569	16.021	66.522	0.771	72.345	46.379	38.946	0.596	48.926	30.738
DeepSeek-67B-Chat	35.647	0.611	46.685	33.359	14.213	0.460	24.921	10.876	30.963	0.413	62.891	45.855	26.941	0.495	44.833	30.030
Yi-1.5-34B-Chat	35.364	0.558	50.757	35.490	14.227	0.430	26.631	19.432	64.307	0.733	73.552	48.042	37.966	0.573	50.313	34.321
Qwen2.5-32B-Instruct	21.336	0.503	33.918	20.704	13.220	0.751	14.160	8.943	28.096	0.450	49.917	37.975	20.884	0.568	32.665	22.541
DeepSeek-V2-Lite (15.7B)	31.458	0.513	47.282	35.404	27.016	0.842	8.183	37.435	27.762	0.327	71.203	48.478	28.745	0.561	42.222	40.439
Yi-1.5-9B-Chat	31.350	0.485	51.677	37.975	3.640	0.112	26.183	17.578	33.748	0.393	72.106	55.284	22.913	0.330	49.989	36.946
Llama-3.1-8B-Instruct	35.622	0.574	49.636	36.383	7.079	0.213	26.742	18.747	33.689	0.386	73.205	56.527	25.463	0.391	49.861	37.219
Qwen-2.5-Instruct-7B	41.203	0.638	51.655	37.975	14.336	0.479	24.098	12.029	33.007	0.377	73.442	53.054	29.515	0.498	49.731	34.353

[1] The Buy & Hold strategy is a passive investment approach commonly used as a baseline strategy, where an investor purchases stocks and holds onto them for an extended period regardless of market fluctuations.

[2] An upward arrow (↑) next to a metric indicates that higher values signify better performance, while a downward arrow (↓) indicates that lower values are preferable.

[3] The numbers highlighted in red indicate the best-performing outcomes for the corresponding metrics.

ing time spans of data collected to construct the agent environment. For the single-asset trading tasks, the warm-up period of stock trading is from 2020-07-01 to 2020-09-30 and the test period is from 2020-10-01 to 2021-05-06. The warm-up period of cryptocurrency trading is from 2023-02-11 to 2023-04-04 and the test period is from 2023-04-05 to 2023-11-05. The warm-up period of ETF trading is from 2019-07-29 to 2019-12-30 and the test period is from 2020-01-02 to 2020-09-21.

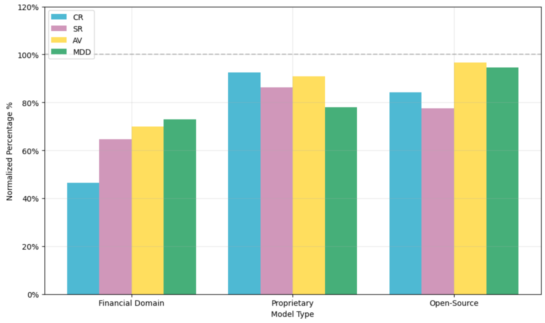
For LLM deployment, we utilize vllm to deploy LLMs. For small-scale LLMs (under 10B parameters), we deploy models on two RTX A6000 GPUs, each with 48GB DRAM. For mid-scale LLMs (10B to 65B parameters), we use four RTX A6000 GPUs. For large-scale LLMs (over 65B parameters), models are deployed on eight A100 GPUs, each equipped with 80GB DRAM.

4.2 Result 1: Stock Trading

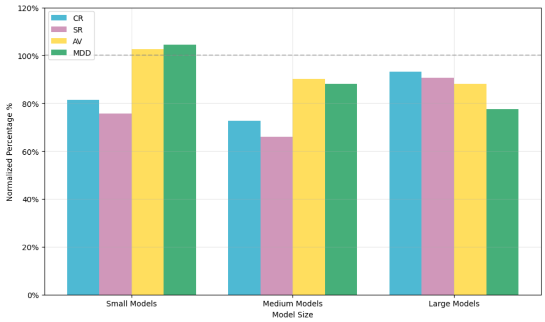
Table 2 presents the performance of thirteen backbone models across seven stocks, accompanied by the average of each metric for all stocks to offer a more comprehensive view of their overall performance. We outline three key insights as follows:

Superior stock trading performance is achieved with proprietary LLMs as agent backbones Compared to agents employing open-source or financial-domain-specific fine-tuned LLMs, those using the three proprietary LLMs demonstrated significantly higher and more consistent average CR and SR, as shown in Figure 3a. Despite being fine-tuned with extensive financial contexts, domain-specific LLMs did not provide a decisive advantage in sequential stock trading decision-making tasks. This may be attributed to their primary training for other functions, such as long financial report analysis exemplified by Palmyra-Fin-70B, rather than decision-making.

Model parameter size increment enhances agent financial decision-making quality and robustness. In the category of open-source LLMs, those exceeding 67B parameters displayed superior CRs and SRs, along with markedly less variance within their category, as illustrated in Figure 3b and Table 2. This underscores the prevailing belief that the reasoning capabilities of LLMs are proportionate to their parameter size, which holds also true in stock



(a) The performance comparison by different model types.



(b) By model parameter sizes across open-source LLMs. *Note: Small-size models refer to models with no more than 10B parameters. Medium-size models refer to models with parameter numbers in the range of (10B, 65B]. Large-size models are those with more than 65B parameters.*

Figure 3: Agent Performance Comparisons from two key perspectives. *The CR, SR, AV, and MDD represent the average values for each model type, expressed as a percentage relative to the Buy & Hold strategy.*

trading, which is a sequential decision-making task in an open-ended, volatile environment by nature.

Proprietary models exhibit significantly stronger decision-making capabilities compared to even the largest open-source LLMs under complex, mixed market conditions, though this advantage is less evident in relatively monotone market environments. During the test phase, primarily influenced by the range of open-source data collected, TSLA and NIO exhibited volatility with mixed upward and downward stock price trends, whereas the other five stocks generally showed bullish trends. The investment signals derived from such complex markets tend to be noisy or delayed, as illustrated in Appendix C. We observed that proprietary models possess a superior ability to manage these challenging conditions and consistently deliver better performance outcomes than large-sized open-source LLMs. Their reasoning capability enables them to effectively utilize other decision-relevant information, such as historical momentum, current holdings, and, most critically, self-reflection outcomes from the agents,

Table 3: Performance of cryptocurrency trading with different LLMs as backbone models across Bitcoin (BTC) and Ethereum (ETH).

Model	BTC				ETH			
	CR↑	SR↑	AV↓	MDD↓	CR↑	SR↑	AV↓	MDD↓
Buy & Hold	21.821	0.989	54.193	20.796	4.528	0.211	60.551	29.889
				Financial Domain Models				
Palmyra-Fin-70B	-20.812	-1.755	29.012	27.782	4.795	0.348	38.986	16.405
				Proprietary Models				
GPT-o1-preview	34.060	1.613	51.905	17.075	2.496	0.123	57.400	27.692
GPT-4	22.396	1.199	45.900	17.206	1.516	0.074	57.648	32.541
GPT-4o	14.330	0.770	45.328	17.278	4.666	0.275	47.858	22.539
Average	23.595	1.195	47.712	17.186	2.893	0.158	54.301	27.591
				Open-Source Models				
Qwen2.5-72B-Instruct	0.549	0.471	2.866	0.897	11.984	0.846	26.866	27.642
Llama-3.1-70B-Instruct	20.440	1.098	45.763	17.813	-11.888	-0.594	56.540	36.416
DeepSeek-67B-Chat	28.307	1.290	53.893	17.944	9.480	0.447	59.902	26.261
Yi-1.5-34B-Chat	13.620	0.628	53.255	22.790	6.325	0.329	54.304	25.707
Qwen2.5-32B-Instruc	11.566	1.258	22.600	7.984	2.823	0.281	28.339	7.883
DeepSeek-V2-Lite (15.7B)	4.804	0.222	53.353	20.562	-9.504	-0.450	59.656	21.270
Yi-1.5-9B-Chat	7.953	0.366	53.285	26.545	-3.684	-0.172	60.552	35.417
Llama-3.1-8B-Instruct	20.521	0.935	53.924	21.104	4.939	0.236	59.264	29.466
Qwen-2.5-Instruct-7B	19.477	0.886	53.994	20.796	-1.339	-0.109	34.932	-16.053
Average	14.137	0.795	43.659	17.382	1.015	0.090	48.928	21.557

thereby facilitating more accurate decisions.

4.3 Result 2 & 3: Cryptocurrency Trading and ETF Trading

In the test phases of both cryptocurrency and ETF trading tasks, market trends are mixed. Notably, the cryptocurrency task shows significantly smaller price fluctuations compared to the ETF task. We outline the key features of using an LLM-agent to make financial decisions across these two distinct markets as follows:

Large-sized open-source models and proprietary models are needed to effectively capture trading signals of cryptocurrency markets, which are highly sensitive to news and financial sentiment. As shown in Table 3, using mid-sized and small-sized open-source models as the decision-making agent backbone generally results in weaker performance than the market baseline with respect to CR and SR.

ETF investment requires proprietary models enriched with extensive pre-trained knowledge to serve as the agents brain and provide robust reasoning support. As shown in Table 4, proprietary models significantly outperform open-source and financial domain-specific models in this task. This advantage arises from the complexity of ETF trading, which necessitates interpreting actionable signals across diverse sectors, demanding more strategic, long-term decisions grounded in deep comprehension and reflection anchored by rich pre-contexts.

4.4 Discussion

Combining all the experimental results, we find that the performance of different LLM varies sig-

Table 4: Performance of ETF trading with different LLMs as backbone models.

ETF	CR↑	SR↑	AV↓	MDD↓
Buy & Hold	2.069	0.06	46.645	35.746
Financial Domain Models				
Palmyra-Fin-70B	24.759	1.152	30.419	8.203
Proprietary Models				
GPT-o1-preview	21.224	0.849	43.766	20.054
GPT-4	2.807	0.110	44.679	37.785
GPT-4o	12.292	0.377	46.150	32.678
Average	12.108	0.445	44.865	30.172
Open-Source Models				
Qwen2.5-72B-Instruct	4.507	0.227	28.090	8.580
Llama-3.1-70B-Instruct	9.895	0.464	30.184	12.759
Yi-1.5-34B-Chat	4.996	0.322	21.986	12.858
Qwen2.5-32B-Instruct	19.617	0.955	29.070	7.496
DeepSeek-V2-Lite (15.7B)	1.389	0.063	31.371	31.831
Yi-1.5-9B-Chat	-4.657	-0.228	28.907	15.545
Llama-3.1-8B-Instruct	11.239	0.475	33.480	15.587
Qwen-2.5-Instruct-7B	-0.384	-0.020	27.596	14.059
Average	5.825	0.282	28.835	14.839

nificantly in stock, cryptocurrency, and ETF trading. This variation not only reflects the inherent complexity of financial markets, but also highlights the importance of model selection or fine-tuning. For instance, proprietary LLM generally exhibit be performance in stock trading due to their strong training on various financial datasets, while open-source models struggle to achieve these results, especially in more volatile environments such as cryptocurrency trading. In addition, the effectiveness of LLM-based agents depends heavily on their ability to adapt to market fluctuations. Agents that incorporate advanced memory systems and dynamic risk assessment capabilities are better able to cope with complex market situations, highlighting the value of the complex architectural features of LLM-based agent framework in financial decision-making tasks.

5 Related Work

5.1 LLM for Financial Domain

The rapid developement of general-domain language models (LMs) has stimulated the exploration of financial LMs, such as pre-trained LMs: FINBERT (Liu et al., 2021; Yang et al., 2020; Araci, 2019; Huang et al., 2023), FINBERT-MRC (Zhang and Zhang, 2023), FLANG (Shah et al., 2022), and several financial LLMs: FINGPT(Liu et al., 2023), FINMA (Xie et al., 2023), INVESTLM (Yang et al., 2023), BloombergGPT (Wu et al., 2023), which leverage extensive training on diverse financial datasets (e.g. stock price data, financial news and analyst reports) and adapt the capabilities of LMs to the unique needs of financial applications. Concurrently, the advancement of LLMs has significantly enhanced the development of language-based agent frameworks in the financial sector, such as FINMEM (Yu et al., 2024a), FINAGENT (Zhang et al., 2024a) and FINROBOT (Yang et al., 2024), characterized by their adaptability and openness. However, variations in framework design, task scope, and data types present challenges in uniformly evaluating the efficacy of LLM agents in financial scenarios.

5.2 Financial LLM Benchmarks

In the realm of financial LLMs, several benchmarks have been developed: FLUE (Shah et al., 2022) introduces the first comprehensive benchmark with five financial NLP tasks, including sentiment analysis, headline classification, named entity recognition, structure boundary detection, and question answering. Pixiu (Xie et al., 2023) expands this benchmark to include financial document understanding and classification tasks, incorporating multimodal datasets. FinBen (Xie et al., 2024) encompasses 36 datasets covering 24 financial tasks. Despite these advancements, there remains a notable gap in benchmarks specifically designed for LLM-based agent applications within the financial sector.

6 Conclusion

INVESTORBENCH offers the community two distinct modes of engagement. The first mode allows participants to integrate their fine-tuned LLMs into the INVESTORBENCH's agent framework to undertake financial decision-making tasks. This setup enables them to benchmark the performance of their models against those previously experimented with by our work. The second mode permits users to directly incorporate the environment and evaluation metrics of INVESTORBENCH into their own designed agents, facilitating a comparative analysis of their agent design's effectiveness. This dual approach provides a flexible framework for testing and enhancing financial decision-making strategies within the INVESTORBENCH ecosystem.

Future research efforts will expand the benchmark by incorporating additional information modalities, such as audio (e.g., earnings call recordings) and graphs (e.g., K-lines, trade charts), to explore whether these data types can enhance decision-making quality. The foundational agent framework of InvestorBench is designed to seamlessly accommodate these modalities, ensuring that the extended benchmark remains easy to use and scalable.

Limitation

First, InvestorBench is currently focusing on single-asset financial decision-making task, without addressing multi-asset tasks such as portfolio management. Second, copyright restrictions on financial domain data may compromise the quality of the datasets we create, potentially limiting the assessment of model performance.

Ethical Statement

The authors take full responsibility for the development of InvestorBench, ensuring that the publicly available part in dataset does not contain personal information, and conform to established ethical guidelines. The data are shared under the MIT license, requiring users to adhere to its terms. InvestorBench 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.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Andrew Ang and Joseph Chen. 2003. Downside risk. *Journal of Portfolio Management*, 29(4):103–112.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Dimitri Bertsekas and Steven E Shreve. 1996. *Stochastic optimal control: the discrete-time case*, volume 5. Athena Scientific.

Mukul Bhatnagar, Sanjay Taneja, and Ramona Rupeika-Apoga. 2023. Demystifying the effect of the news (shocks) on crypto market volatility. *Journal of Risk and Financial Management*, 16(2):136.

John H. Cochrane. 1988. Volatility tests and efficient markets: A review essay. *Journal of Monetary Economics*, 22(3):463–485.

Eva Eigner and Thorsten Händler. 2024. Determinants of llm-assisted decision-making. *arXiv preprint arXiv:2402.17385*.

Thomas L Griffiths, Jian-Qiao Zhu, Erin Grant, and R Thomas McCoy. 2023. Bayes in the age of intelligent machines. *arXiv preprint arXiv:2311.10206*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

Allen H Huang, Hui Wang, and Yi Yang. 2023. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.

John Hull. 2007. *Risk Management and Financial Institutions*. John Wiley & Sons.

Taylan Kabbani and Ekrem Duman. 2022. Deep reinforcement learning approach for trading automation in the stock market. *IEEE Access*, 10:93564–93574.

Yuan Li, Bingqiao Luo, Qian Wang, Nuo Chen, Xu Liu, and Bingsheng He. 2024. 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.

Xiao-Yang Liu, Guoxuan Wang, and Daochen Zha. 2023. Fingpt: Democratizing internet-scale data for financial large language models. *arXiv preprint arXiv:2307.10485*.

Yang Liu, Qi Liu, Hongke Zhao, Zhen Pan, and Chuanren Liu. 2020. Adaptive quantitative trading: An imitative deep reinforcement learning approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 2128–2135.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2021. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence*, pages 4513–4519.

Meta Llama team. 2024. The llama 3 herd of models.

Ananth N Madhavan. 2016. *Exchange-traded funds and the new dynamics of investing*. Oxford University Press.

Jaap MJ Murre and Joeri Dros. 2015. Replication and analysis of ebbinghaus forgetting curve. *PloS one*, 10(7):e0120644.

OpenAI. 2022. Introducing chatgpt.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Alibaba Group Qwen team. 2024. Qwen2.5.

Raeid Saqur, Ken Kato, Nicholas Vinden, and Frank Rudzicz. 2024. Nifty financial news headlines dataset. *arXiv preprint arXiv:2405.09747*.

Raj Sanjay Shah, Kunal Chawla, Dheeraj Eidnani, Agam Shah, Wendi Du, Sudheer Chava, Natraj Raman, Charese Smiley, Jiaao Chen, and Diyi Yang. 2022. When flue meets flang: Benchmarks and large pre-trained language model for financial domain. *arXiv preprint arXiv:2211.00083*.

William F. Sharpe. 1994. The sharpe ratio. *The Journal of Portfolio Management*, 21(1):49–58.

Writer Engineering team. 2024. Palmyra-Fin-70B-32k: a powerful LLM designed for Finance. `https://dev.writer.com`.

Olivier Vanhoucke. 2023. Crypto news dataset. Accessed: 2024-08-20.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. 2024. The finben: An holistic financial benchmark for large language models. *arXiv preprint arXiv:2402.12659*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*.

Huajian Xin, Z. Z. Ren, Junxiao Song, Zhihong Shao, Wanjia Zhao, Haocheng Wang, Bo Liu, Liyue Zhang, Xuan Lu, Qiushi Du, Wenjun Gao, Qihao Zhu, Dejian Yang, Zhibin Gou, Z. F. Wu, Fuli Luo, and Chong Ruan. 2024. Deepseek-prover-v1.5: Harnessing proof assistant feedback for reinforcement learning and monte-carlo tree search.

Hongyang Yang, Boyu Zhang, Neng Wang, Cheng Guo, Xiaoli Zhang, Likun Lin, Junlin Wang, Tianyu Zhou, Mao Guan, Runjia Zhang, et al. 2024. Finrobot: An open-source ai agent platform for financial applications using large language models. *arXiv preprint arXiv:2405.14767*.

Yi Yang, Yixuan Tang, and Kar Yan Tam. 2023. Investlm: A large language model for investment using financial domain instruction tuning. *arXiv preprint arXiv:2309.13064*.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

Ziruo Yi, Ting Xiao, Kaz-Onyeakazi Ijeoma, Ratnam Cheran, Yuvraj Baweja, and Phillip Nelson. 2022. Stock2vec: An embedding to improve predictive models for companies. *arXiv preprint arXiv:2201.11290*.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W Suchow, and Khaldoun Khashanah. 2024a. Finmem: A performance-enhanced llm trading agent with layered memory and character design. In *Proceedings of the AAAI Symposium Series*, volume 3, pages 595–597.

Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yupeng Cao, Zhi Chen, Jordan W Suchow, Rong Liu, Zhenyu Cui, Zhaozhuo Xu, et al. 2024b. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *arXiv preprint arXiv:2407.06567*.

Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, et al. 2024a. Finagent: A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. *arXiv preprint arXiv:2402.18485*.

Yadong Zhang, Shaoguang Mao, Tao Ge, Xun Wang, Adrian de Wynter, Yan Xia, Wenshan Wu, Ting Song, Man Lan, and Furu Wei. 2024b. Llm as a mastermind: A survey of strategic reasoning with large language models. *arXiv preprint arXiv:2404.01230*.

Yuzhe Zhang and Hong Zhang. 2023. Finbert–mrc: financial named entity recognition using bert under the machine reading comprehension paradigm. *Neural Processing Letters*, 55(6):7393–7413.

Zhihan Zhou, Liqian Ma, and Han Liu. 2021. Trade the event: Corporate events detection for news-based event-driven trading. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2114–2124, Online. Association for Computational Linguistics.

Appendices

A Memory Ranking Mechanism of FinMem

Upon receiving an investment inquiry, FinMem retrieves the top-K critical memory events from each layer and channels them to the immediate reflection component of the working memory. These events are selected based on their information retrieval score, γ_l^E, where l represents the layer (shallow, intermediate, or deep), as defined in Equation 2.

$$\gamma_l^E = S^E_{\text{Recency}_l} + S^E_{\text{Relevancy}_l} + S^E_{\text{Importance}_l}, \quad (2)$$

where each memory event is only associated with one score and can only belong to a single layer.

Let E denote a given memory event. The scoring mechanism for E, adapted from Park et al. (Park et al., 2023) but with modified recency and importance computations, is tailored to handle data with various timelines and to achieve layered processing that represents the diverse periodicities of the financial environment. This score encapsulates three metrics: recency (how recently the event occurred), relevancy (the event's pertinence to the current context), and importance (the event's significance). Individual metric scores exceeding 1.0 are scaled to the [0,1] range before being summed, ensuring a balanced contribution from each component and preventing any single metric from dominating the overall score. The resulting composite score provides a comprehensive evaluation of the memory event's significance within the multi-layered, periodically varying financial landscape.

$$S^E_{\text{Recency}_l} = e^{-\frac{\delta^E}{Q_l}}, \quad \delta^E = t_{\text{P}} - t_E, \quad (3)$$

where δ^E represents the time elapsed between a memory event's occurrence and the trading inquiry's arrival. The model utilizes three processing layers, each corresponding to a specific timeframe: shallow ($Q_{\text{shallow}} = 14$ days), intermediate ($Q_{\text{intermediate}} = 90$ days), and deep ($Q_{\text{deep}} = 365$ days). These intervals represent two weeks, a quarter, and a year respectively.

When a trade inquiry P arrives in processing layer l via an LLM prompt, the agent calculates the recency score $S^E_{\text{Recency}_l}$ for a memory event E using Equation 3. This score inversely correlates with the time elapsed between the inquiry and the event's memory timestamp, mapping to Ebbinghaus's forgetting curve (Murre and Dros, 2015). The stability term Q_l in Equation 3 modulates memory decay rates across layers, with higher values in deeper layers indicating longer memory persistence. For instance, in the trading context, company annual reports (e.g., Form 10-Ks) are assigned higher stability values and categorized within deeper processing layers compared to daily financial news, reflecting their extended timeliness, relevance, and impact on financial decision-making.

$$S^E_{\text{Relevancy}_l} = \frac{\mathbf{m_E} \cdot \mathbf{m_P}}{\|\mathbf{m_E}\|_2 \times \|\mathbf{m_P}\|_2} \quad (4)$$

The relevancy score $S^E_{\text{relevancy}l}$ quantifies the semantic similarity between a memory event E and the current query P using cosine similarity of their respective embedding vectors, $\mathbf{mE}$ and $\mathbf{m_P}$, as shown in Equation 4. These embeddings are generated from the event's textual content and the LLM prompt query (which includes trading inquiries and the agent's character setting) using OpenAI's "text-embedding-ada-003" model.

The importance score $S^E_{\text{Importance}l}$ for a memory event E in layer l is calculated as the product of a value v_l^E (derived from a uniform piecewise scoring function, Equation 5) and a degrading ratio θ_l (Equation 6), as shown in Equation 7. This approach, adapted from (Park et al., 2023), is tailored to our stratified long-term memory structure. The likelihood of higher v_l^E values increases from shallow to deep layers, while θ_l measures the diminishing importance of an event over time using layer-specific exponential functions. The base α_l for each layer follows $\alpha shallow < \alpha_{intermediate} < \alpha_{deep}$ (set to 0.9, 0.967, and 0.988 respectively), ensuring θ_l decreases to a threshold of 5 after 30, 90, and 365 days for shallow, intermediate, and deep layers. This layered approach, implemented through three-piece-wise functions for both $S^E_{\text{Importance}l}$ and $S\text{Recency}l^E$, enables FinMem to process long-term memory in a stratified manner. Memory events are purged when $S\text{Recency}l^E$ falls below 0.05 or $S\text{Importance}_l{}^E$ is under 5 (pre-scaling), maintaining the relevance and efficiency of the memory store.

$$v_l^E = \begin{cases} 40 & \text{with probability } p_1 \\ 60 & \text{with probability } p_2 \\ 80 & \text{with probability } p_3 \end{cases} \quad (5)$$

$$\theta_l = (\alpha_l)^{\delta^E}, \quad l = \text{shallow, intermediate, deep}, \quad (6)$$

where $p_1 + p_2 + p_3 = 1$, but their values vary by shallow, intermediate, and deep processing. when shallow processing $p_1, p_2, p_3 = \{0.8, 0.15, 0.05\}$, intermediate processing, $p_1, p_2, p_3 = \{0.05, 0.8, 0.15\}$ and deep processing, $p_1, p_2, p_3 = \{0.05, 0.15, 0.8\}$.

$$S^E_{\text{Importance}_l} = v_l^E * \theta_l, \quad (7)$$

Furthermore, FinMem employs an access counter function to dynamically manage memory events across layers, ensuring that crucial events influencing trading decisions are elevated to deeper layers for extended retention and recurring access. This process, monitored by the LLM validation tool Guardrails AI, tracks critical memory IDs across layers. Events deemed pivotal for investment success receive a 5-point boost to their importance score ($S^E_{\text{Importance}l}$). Upon meeting upgrade criteria for a deeper layer, an event's recency score ($S\text{Recency}_l{}^E$) is reset to 1.0, underscoring its significance and preventing rapid decay. Conversely, less relevant events gradually fade. This mechanism allows FinMem to efficiently identify, prioritize, and retain key events based on their nature and retrieval frequency, while gradually phasing out less impactful information, thereby maintaining a dynamic and relevant memory structure for financial decision-making.

B Details on Evaluation Metrics

Below is a brief overview of these metrics:
Cumulative Return (CR) % measures the total value change of an investment over time by summing daily logarithmic returns, shown in Equation 8. Higher values indicate better strategy effectiveness.

$$\textbf{CR} = \sum_{t=1}^{n} r_i = \sum_{t=1}^{n} \left[\ln \left(\frac{p_{t+1}}{p_t} \right) \cdot \text{action}_t \right] \quad (8)$$

, where r_i is the logarithmic return from day t to $t+1$, p_t and p_{t+1} are the closing prices on days t and $t+1$, respectively, and action_t is the model's trading decision for day t.
Sharpe Ratio (SR) assesses risk-adjusted returns by dividing the average excess return (R_p) over the risk-free rate (R_f) by its volatility (σ_p), detailed in Equation 9. Higher ratios signify better performance.

$$\textbf{SR} = \frac{R_p - R_f}{\sigma_p} \quad (9)$$

Annualized Volatility (AV) % and Daily Volatility (DV) % quantify return fluctuations; AV is derived by scaling DV (*standard deviation of daily logarithmic returns*) by the square root of the annual trading days (252), as in Equation 10. This metric highlights potential return deviations across the year.

$$\textbf{AV} = \textbf{DV} \times \sqrt{252} \quad (10)$$

Max Drawdown (MDD) % calculates the largest portfolio value drop from peak to trough, as given in Equation 11. Lower values indicate lesser risk and higher strategy robustness.

$$\textbf{MDD} = \max\left(\frac{P_{\text{peak}} - P_{\text{trough}}}{P_{\text{peak}}}\right) \quad (11)$$

Note that CR and the SR are often considered more essential than AV and MDD in evaluating asset trading performance due to their focus on long-term gains and risk-adjusted returns by their definition. Here, we regard these two metrics as primary metrics when evaluating the experiment outcomes.

C An example of mixed and lagged market signals: Partial investment insights of TSLA on 2021-03-05

Here are some insights from the agent's memory module for TSLA as of 2021-03-05. A few memory records are omitted; these are either neutral or positive. Despite this, the stock price trend for TSLA is sharply downward, conflicting with the overall positive financial sentiments and market signals. Utilizing proprietary models such as GPT4 and GPT-o1 as backbones, the financial decision-making agent can leverage other investment insights like historical momentum and self-reflection to consistently support a 'Sell' decision. In contrast, the large-sized open-source models like Qwen2.5-72B and DeepSeek-67B-Chat exhibit instability in producing consistent actions across repeated experimental trials.

Short-term Memory

1 **Sentiment: Negative**: The key insights from the news regarding Tesla Inc (NASDAQ: TSLA) losing market share to Ford Motor Company's (NYSE: F) Mustang Mach-E in the United States are as follows:1. **Market Competition**: Tesla is facing increased competition in the electric vehicle (EV) market, particularly from established automakers like Ford. The Mustang Mach-E's success indicates that other companies are effectively entering the EV space, which could impact Tesla's dominance.2. **Market Share Impact**: The loss of market share to Ford suggests that Tesla's growth trajectory in the U.S. may be challenged. Investors should consider the potential implications for Tesla's revenue and profitability if this trend continues...

2 **Sentiment: Positive**: The news about a Banksy artwork being burned and sold as a non-fungible token (NFT) highlights a few key insights relevant to investment decisions, particularly concerning innovative and disruptive technologies: 1. **Emergence of NFTs**: The transformation of physical art into digital assets through NFTs signifies a growing trend in the digital economy. This trend could influence sectors beyond art, including technology and finance, as more industries explore blockchain applications ...

3 **Sentiment: Neutral**: To provide a summary of key insights relevant to making investment decisions about Tesla (TSLA) from the pre-open movers news, I would focus on the following aspects: 1. **Stock Performance**: Look for any significant pre-market price movements for TSLA. If the stock is showing a notable increase or decrease, it could indicate investor sentiment or reaction to recent news.2. **News Catalysts**: Identify any specific news items or announcements related to Tesla that might be influencing its stock price. This could include earnings reports, product launches, regulatory news, or changes in leadership...

4 **Sentiment: positive**: The key insights from the news about Tesla's vehicle registrations in Germany are as follows: 1. **Significant Growth in Registrations**: Tesla experienced a 78% year-over-year increase in vehicle registrations in Germany as of January. This substantial growth indicates a strong demand for Tesla vehicles in one of Europe's largest automotive markets.2. **Market Penetration**: The surge in registrations suggests that Tesla is successfully penetrating the German market, which is traditionally dominated by local automakers. This could imply a growing acceptance and preference for electric vehicles (EVs) in Germany, benefiting Tesla as a leading EV manufacturer...

...

Mid-term Memory

1 **Sentiment: Negative**: The news about U.S. Senator Elizabeth Warren's proposal for a Democratic wealth tax could have several implications for investment decisions regarding Tesla (TSLA):1. **Impact on Wealthy Investors**: The proposed wealth tax targets the wealthiest Americans, which could include major shareholders and investors in Tesla. If implemented, this tax could lead to changes in investment strategies among these individuals, potentially affecting Tesla's stock price due to shifts in large-scale holdings....5. **Long-term Implications**: While the immediate impact might be uncertain, the long-term implications of a wealth tax could lead to a more equitable distribution of wealth, potentially increasing consumer spending power. This could benefit Tesla if it results in a larger market for its products. Investors in Tesla should closely monitor the progress of this proposal and consider its potential effects on market dynamics and investor behavior.

2 **Sentiment: positive**: The key insights from the news information regarding consumer preferences in China for electric cars, particularly focusing on price and driving range, can be relevant for making investment decisions about Tesla (TSLA) in the following ways: 1. **Market Demand Alignment**: Tesla's ability to align its product offerings with the primary concerns of Chinese consumers 2014price and driving rangeˇ2014could significantly impact its market share and sales growth in China. Investors should monitor how Tesla's models compare to competitors in these aspects... Overall, Tesla's success in

addressing the key consumer preferences in China 2014price and driving range 2014will be critical for its growth prospects in the region, making these factors important considerations for investors.

...

Long-term Memory

1 **Sentiment: Negative**: The key insights from the news about Tesla Inc. (NASDAQ: TSLA) facing competition in China are as follows: 1. **Increased Competition**: Tesla's Model 3 is experiencing significant competition from a budget electric vehicle produced through a joint venture involving General Motors (NYSE: GM). This indicates a more competitive landscape in the Chinese electric vehicle (EV) market, which could impact Tesla's market share and sales growth in the region... For investors, these insights suggest a need to closely watch Tesla's strategic responses to competition in China, its pricing strategies, and any potential impacts on its financial performance. Additionally, understanding the broader competitive landscape and Tesla's ability to sustain its growth trajectory will be crucial for making informed investment decisions.

2 **Sentiment: positive**: The news about Bill Gates' concerns regarding Bitcoin's impact on climate change highlights a broader issue of environmental sustainability in the tech and financial sectors. Here are the key insights relevant to making investment decisions about Tesla (TSLA): 1. **Environmental Impact Awareness**: Bill Gates' concerns underscore the growing awareness and scrutiny of the environmental impact of technology and financial products. This is relevant for Tesla, as the company positions itself as a leader in sustainable energy and electric vehicles (EVs)... Overall, the emphasis on environmental impact and sustainability in the tech sector could reinforce Tesla's strategic advantages and appeal to investors prioritizing green investments.

...

D Extended Task: Multi-Asset Portfolio Management Task

Besides the single-asset trading tasks presented in the paper, INVESTORBENCH is able to incorporate the multi-asset portfolio management task, which further illustrates INVESTORBENCHs generalization capabilities. This advanced task involves more sophisticated trading strategies and mathematical inference, allowing the agent to dynamically allocate asset weights at each decision step and rebalance the portfolio based on current market conditions.

The experimental results for a compact portfolio consisting of TSLA, JNJ, and UVV are shown in Table 5, using three LLMs as agent backbones (one from each size category). The INVESTORBENCH agent's performance on this portfolio management task across LLMs aligns with the overall trend with the single-asset trading task presented in Figure 3b in our paper, the agent with the large-size LLM delivers the best performance, while that with small-size and mid-size LLMs performs lower and closely in terms of primary evaluation metrics, Cumulative Returns (CR) and Sharpe Ratio (SR). This illustrates the consistency of our decision-making agent performance across distinct types of LLMs for various task types and trading strategies.

Table 5: Comparison of performance across LLMs of varying sizes in multi-asset trading tasks

Model Type	CR	SR	AV	MDD
Llama-3.1-70B-Instruct	31.473	1.482	17.100	11.205
Yi-1.5-34B-Chat	15.458	0.903	13.769	9.883
Llama-3.1-8B-Instruct	20.255	0.953	17.106	11.219

E Case Study on INVESTORBENCH in Extreme Market Conditions

Table 6: Comparison of performance across LLMs of varying sizes in the extreme market conditions

Model Type	CR	SR	AV	MDD
Buy & Hold	-56.738	-0.936	53.911	52.078
GPT-4	10.163	0.433	20.866	12.672
Llama-3.1-70B-Instruct	-27.596	-0.633	38.708	34.797
Yi-1.5-34B-Chat	-32.723	-0.606	48.012	41.197
Llama-3.1-8B-Instruct	-37.56	-0.736	45.350	47.748

Here, we use a case study to demonstrate INVESTORBENCH's robust performance during periods of extreme market featured by significant market volatility.

Table 7: Performance Comparison of LLMs in Multi-Asset Trading Tasks under Different Market Conditions

Model Type	Bullish				Bearish				Mixed Signal			
	CR	SR	AV	MDD	CR	SR	AV	MDD	CR	SR	AV	MDD
Buy and hold	51.987	3.654	36.322	12.551	-96.218	-2.884	60.645	65.24	39.244	0.600	52.339	37.975
GPT-4	27.219	2.095	33.163	12.551	61.236	1.781	62.513	18.759	45.246	0.821	44.088	25.031
Llama-3.1-70B-Instruct	43.926	3.144	36.349	12.551	20.819	0595	63.612	40.686	37.545	0.615	48.862	29.813
Yi-1.5-34B-Chat	17.571	1.790	25.062	5.617	10.424	0.317	59.668	41.988	35.364	0.558	50.757	35.49
Llama-3.1-8B-Instruct	18.685	1.417	33.655	9.778	-39.44	-1.386	51.737	44.274	35.622	0.574	49.636	36.383

We evaluated the asset, TSLA, by examining Cumulative Returns (CRs) and Sharpe Ratios (SRs), Annualized Volatility (AV) and Maximum Drawdown(MDD). Our dataset spans a training period from 2022-01-17 to 2022-03-31 and a test period from 2022-04-01 to 2022-10-15. These periods are selected because the VIX (CBOE Volatility Index) remained elevatedaveraging above 20indicating heightened market volatility.

We present the INVESTORBENCH agent's performance using four representative LLMs as backbonesone proprietary model (GPT-4) and three open-source models of varying sizes (Llama-3.1-70B-Instruct, Yi-1.5-34B-Chat, and Llama-3.1-8B-Instruct)chosen for their relatively stable performance (See Table 6 below). Overall, the agents performance with each LLM aligns with the results in Table 2 as well as Figure 3a and 3b in our paper. Specifically, 1) The proprietary LLM outperforms the open-source models. 2) Larger model parameter sizes consistently lead to higher decision-making quality and robustness.

F Case Study on INVESTORBENCH across Different Market Scenarios

We introduced an additional set of case studies in bullish, bearish, and mixed-signal market environments to highlight the robustness of the agents performance in INVESTORBENCH.

We assess the INVESTORBENCH agent's performance robustness using three separate time periods with distinct market trends on the same assets, task type, and LLMs. For the bullish market, the training period spanned from 2023-03-06 to 2023-05-01, followed by a test period from 2023-05-02 to 2023-07-10. For the bearish market, the training period ran from 2022-08-08 to 2022-09-26, and the test period from 2022-09-27 to 2023-01-02. We present Bullish and Bearish results below Table 7. For the mixed market condition, we employed the same dates specified in Section 4.1 of the paper.

In the mixed-signal and bearish market conditions, the agents performance trends across various LLMs remain closely aligned with the TSLA trading case shown in Table 2 as well as Figure 3a and 3b of our paper. In the bullish market scenario, while larger LLMs generally maintain stronger performance, the simple Buy-and-Hold strategysupported by a continuously rising asset priceoutperforms strategies provided by the agent that consider multiple investment perspectives. The latter approach can be influenced by noisy market signals, such as occasional neutral or negative news. The proprietary LLM GPT-4, characterized by its comprehensive reasoning, exhibits more caution and thus lower returns compared to the large-scale open-source model Llama-3.1-70B-Instruct. However, both still surpass other LLMs, consistent with several asset trading cases reported in Table 2.

Our case study proves that the INVESTBENCH agent's performance stays robust and consistent with the analysis outcomes in our paper in general.

G Case Study on Single Stock Trading: Forecast for TSLA on 2022-10-25 to Predict Trading Decision on 2022-10-26

Initialize Profile

1. Operations:
- Provide a performance overview of the trading stock based on available data.
- Set up the risk inclination as the key character of the trading agent.

2. Range: Financial information such as the financial sectors, historical performance, and previous stock trends of the trading stock.

3. Prompts: You are an experienced trading manager and investment firm. Your task is to make informed decisions on the given stock based on the provided information.
Under Self-Adaptive Risk Character Setting: When historical momentum is positive, you are a risk-seeking investor. But when historical momentum is negative, you are a risk-averse investor.

4. General background setting:
You have accumulated a lot of information about the following sectors, so you are especially good at trading them: 1)Electric Vehicles (Automotive Sector). 2) Energy Generation and Storage...From year 2021 to 2022 September, Tesla's continued growth and solid financial performance over the defined period ...

Summarize

1. Operations:
- Summarize different types of input information.
- Distribute them to corresponding layers of the long-term memory database.

2. Range: Daily market news, Long Documents such as company 10-K and 10-Q reports

3. Prompts:
- (1). Summarize the contents: Summarize the following documents into 1000 words.
- (2). Comprehend the investment sentiment of news insights: The positive, neutral and negative scores are for understanding the investment sentiments, opinions, or emotions. For example, positive news about a company can lift investor sentiment, encouraging more buying activity, which in turn can push stock prices higher...

4. Outputs:
(1). **To Shallow Memory Layer:**
- [News (ID: 261)] Here's How Much You Would Have Made Owning Tesla Stock In The Last 10 Years Tesla (NASDAQ:TSLA) has outperformed the market over the past 10 years by 50.69% on an annualized basis producing an average annual return of 60.76%. Currently, Tesla has a market capitalization of $683.54 billion.... The sentiment is {positive}.
- [News (ID: 278)] Tesla Q3 Earnings Are Imminent. Can Nio Foreshadow What's To Come? What To Know Before The Print Tesla Inc (NASDAQ: TSLA) shares were trading down slightly Wednesday afternoon ahead of the automakerś third-quarter report, but the stock is up 6% over the last five sessions... The sentiment is {positive}.
- ...

(2). **To Intermediate Memory Layer:**
- [Form 10-Q (ID: 222)] Tesla Q3 2022 revenues were $21.5 billion, up 56% year-over-year. Automotive sales revenue grew 56% to $17.8 billion driven by higher Model 3/Y and Model S/X deliveries. Gross automotive margin declined to 27.9% due to cost inflation and factory ramps. Net income was $3.3 billion, up 102% year-over-year. Positive free cash flow was $6.1 billion...
- [News (ID: 275)] Tesla Q3 Earnings Highlights: Record Revenue, Operating Margin And Free Cash Flow, Tesla Semi Deliveries Coming In December Electric vehicle leader Tesla Inc (NASDAQ: TSLA) reported third-quarter financial results after market close Wednesday...The sentiment is {neutral}.
- [News (ID: 274)] Tesla Preps For 2023 Cybertruck Launch, Will Make Battery Packs In California The Cybertruck is one of Tesla Inc. (NASDAQ: TSLA) most hotly anticipated, but also most delayed, products. - ...The sentiment is {negative}.

(3). **To Deep Memory Layer:**
- [News (ID: 161)] Tesla Whale Trades Spotted A whale with a lot of money to spend has taken a noticeably bearish stance on Tesla. Looking at the options history for Tesla (NASDAQ:TSLA) we detected 477 strange trades. The sentiment is {positive}.
- [Self-reflection (ID: 226)] Given the short-term positive news score in the market for TSLA and a positive cumulative return, there is a high probability of continued growth in the short term. However, investor should be aware of potential threats in the mid-term market with competitors like General Motors, and Nio...

Observe

1. Operations: Access and interpret market indicators such as current stock prices and historical momentum data.
2. Range: Stock's daily adjusted closing price, historical momentum in the past k days ($k = 3$ in this case), etc.
3. Prompts:
- The information below provides a summary of stock price fluctuations over the previous few days, which is the "momentum" of a stock. It reflects the trend of a stock. Momentum is based on the idea that securities that have performed well in the past will continue to perform well, and conversely, securities that have performed poorly will continue to perform poorly.

4. Outputs:
- (1). The daily adjusted closing price of TSLA on {2022-10-25} is {$222.42}.
- (2). **Train:** On {2022-10-25}, the momentum of TSLA, indicated by the price difference between the current and the next trading
day, is {$2.22}.
Test: On {2022-10-25}, the historical momentum of TSLA, as measured by its cumulative logarithmic returns in the past
{3} days, was {7.05%}.

Figure 4: First section of INVESTORBENCH's workflow for perceiving and processing multi-sourced information from market environment.

Reflect

1. Operations:
Train: - Infer the reasoning from the retrieved memories insights for the recommended trading actions the increase or decrease of the trading stock price.
Test: - Generate recommended trading actions. - Infer the reasoning for making such trading action.
Train & Test: - Track the memory insight IDs that contributed to the current trading decision.
2. Range: Interpretation of observed market indicators and top k insights from each layer of the long-term memory database.
3. Prompts:
Train: Given the following information, can you explain why the financial market fluctuation from the current day to the next day behaves like this? Summarize the reason for the decision. You should provide summary information and the IDs of the information to support your summary.
Test: Given the information, can you make an investment decision? Just summarize the reason for the decision. Consider only the available short-term information, the mid-term information, the long-term information, and the reflection-term information. Consider the momentum of the historical stock price. Also, consider how much share of the stock the investor holds now. You should provide exactly one of the following investment decisions: buy or sell. You could go with the 'hold' option if necessary. You also need to provide the IDs of the information to support your decision.
4. Outputs:
Train:
- Recommended Action Reason: The recommended trading action is 'Buy'. This can be attributed to positive market sentiment driven by a range of favorable factors. Notably, Tesla's strong battery supply chain, effective market expansion plans in Germany, and its ability to maintain market strength despite potential competitive threats positively influenced investor confidence, contributing to the stock price rise.
- Reflected Contributed Insights IDs: {261, 278, 275, 274, 161,...}

Test:
- Trading Action: {Buy}
- Recommended Action Reason: Considering the short-term positive revenue growth, higher Model 3/Y and Model S/X deliveries, positive free cash flow, and increased cash and cash equivalents, Tesla appears to have strong financial health. The momentum being positive, it also encourages a risk-seeking strategy. The significant positive signs from news such as the imminent Q3 earnings add to the favorable view of the investment. Despite the mid-term neutral sentiment, long-term information shows only marginal drops and the stock is still looking strong overall. These, along with positive reflection-term information, form the reasoning behind the 'buy' decision...
- Reflected Contributed Insights IDs: {261, 278, 275, 274, 161,...}

Figure 5: Second section of INVESTORBENCH's workflow for generating trading action, reasoning and reflection.