STOCKBENCH: CAN LLM AGENTS TRADE STOCKS PROFITABLY IN REAL-WORLD MARKETS?

Yanxu Chen[♠♡*†] **Zijun Yao**[♠*] **Yantao Liu**[♠*] **Jin Ye**[♡] **Jianing Yu**[♡]
Lei Hou[♠] **Juanzi Li**[♠]

[♠]Tsinghua University [♡]Beijing University of Posts and Telecommunications

yaozj20@mails.tsinghua.edu.cn, cyx666@bupt.edu.cn

Project: https://stockbench.github.io/

GitHub: https://github.com/ChenYXxxx/stockbench

ABSTRACT

Large language models (LLMs) have recently demonstrated strong capabilities as autonomous agents, showing promise in reasoning, tool use, and sequential decision-making. While prior benchmarks have evaluated LLM agents in domains such as software engineering and scientific discovery, the finance domain remains underexplored, despite its direct relevance to economic value and high-stakes decision-making. Existing financial benchmarks primarily test static knowledge through question answering, but they fall short of capturing the dynamic and iterative nature of trading. To address this gap, we introduce STOCKBENCH, a contamination-free benchmark designed to evaluate LLM agents in realistic, multi-month stock trading environments. Agents receive daily market signals—including prices, fundamentals, and news—and must make sequential buy, sell, or hold decisions. Performance is assessed using financial metrics such as cumulative return, maximum drawdown, and the Sortino ratio. Our evaluation of state-of-the-art proprietary (*e.g.,* GPT-5, Claude-4) and open-weight (*e.g.,* Qwen3, Kimi-K2, GLM-4.5) models shows that while most LLM agents struggle to outperform the simple buy-and-hold baseline, several models demonstrate the potential to deliver higher returns and manage risk more effectively. These findings highlight both the challenges and opportunities in developing LLM-powered financial agents, showing that excelling at static financial knowledge tasks does not necessarily translate into successful trading strategies. We release STOCKBENCH as an open-source resource to support reproducibility and advance future research in this domain.

1 INTRODUCTION

Large language models (LLMs) have enabled a new wave of autonomous agents, demonstrating strong capabilities in reasoning, tool use, and long-horizon decision making (OpenAI, 2024; Anthropic, 2025a; DeepMind, 2025; Liu et al., 2024; Guo et al., 2025a; Meta-AI, 2025; Yang et al., 2024a; Bai et al., 2025; OpenAI, 2025b). This agentic capability is verified by benchmarks in various different domains, such as software engineering (Jimenez et al., 2024; Yang et al., 2024b), scientific discovery (Mialon et al., 2023), and marketing (Chen et al., 2025; Barres et al., 2025), using the most recent advanced LLMs such as GPT-5 (OpenAI, 2025a) and Claude-4 (Anthropic, 2025b), highlighting their promise for workflow automation and productivity gains. The ever-evolving agent capability of LLMs pushes agent application toward real-world productivity and economic value.

Among various agent application scenarios, the finance domain stands out due to its direct connection to economic value and the high stakes involved in decision making (Wu et al., 2023; Lee et al., 2024; Nie et al., 2024). To holistically evaluate the profitability and risk-management capabilities of LLM agents in finance, an ideal benchmark should adhere to three key principles: **(1) Realistic**

[*]Equal contribution.
[†]Work was completed during an internship at Tsinghua University.

Table 1: Comparison of STOCKBENCH with existing financial benchmarks.

Benchmark	Market Simulation	Multi Month Horizon	Continuous Decision	Contamination Free	Direct Economic Value
FinQA (Chen et al., 2021)	✗	✗	✗	✗	✗
ConvFinQA (Chen et al., 2022)	✗	✗	✗	✗	✗
FLUE (Shah et al., 2022)	✗	✗	✗	✗	✗
FinEval (Guo et al., 2025b)	✗	✗	✗	✗	✗
CPA-QKA (Kuang et al., 2025)	✗	✗	✗	✗	✗
BizFinBench (Lu et al., 2025)	✗	✗	✗	✗	✗
Finance Agent Benchmark (Bigeard et al., 2025)	✓	✗	✓	✗	✗
INVESTORBENCH (Li et al., 2024)	✓	✓	✓	✗	✓
FinSearchComp (Hu et al., 2025)	✗	✓	✓	✗	✓
STOCKBENCH (Ours)	✓	✓	✓	✓	✓

Market Interaction. The agent must operate in a dynamic market environment, responding to real-time price movements and news events. **(2) Continuous Decision Making.** The agent should make sequential trading decisions over an extended horizon, reflecting the iterative nature of investment strategies. **(3) Data Contamination Free.** To ensure fair evaluation, the agent must not have prior exposure to the test data during training, necessitating careful data curation and temporal separation.

However, existing benchmarks for financial agents largely focus on static question-answering tasks (Chen et al., 2021; Zhu et al., 2021; Yin et al., 2023), which are designed to test the financial knowledge coverage of LLMs but fail to reflect practical trading scenarios. Although recent efforts like INVESTORBENCH (Li et al., 2025a) take a step towards simulating trading environments, this thread of works only focuses only on single-stock-trading and is conducted on historical data prior to 2021, raising concerns about potential data contamination.

To mitigate the gap, we propose STOCKBENCH, an evolving benchmark that places LLM agents into realistic stock-trading environments, directly measuring their profitability and risk-management capabilities. Specifically, STOCKBENCH is designed to be: **(1) Realistic.** Agents receive daily market signals including prices, company fundamentals, and news headlines, reflecting real-world trading contexts. **(2) Continuous.** Agents must make sequential daily trading decisions (buy, sell, or hold) over a multi-month horizon, mirroring the iterative nature of investment strategies. **(3) Contamination-Free.** The benchmark is instantiated using recent market data from March 2025 to July 2025 and will be continuously updated to avoid overlap with the training corpora of contemporary LLMs. Performance is evaluated using key financial metrics such as cumulative return, maximum drawdown, and the Sortino ratio, providing a direct and quantitative assessment of trading success.

As a proof of concept, we evaluate a diverse set of LLM agents, including both proprietary models (*e.g.,* GPT-5 (OpenAI, 2025a), Claude-4 (Anthropic, 2025b)) and open-weight models (*e.g.,* Qwen3 (Yang et al., 2025), Kimi-K2 (Team et al., 2025), GLM-4.5 (Zeng et al., 2025)), alongside an equal-weight buy-and-hold baseline. Surprisingly, despite their strong performance on financial QA benchmarks, most LLM agents fail to outperform this simple baseline in terms of both cumulative return and risk-adjusted return. This finding suggests that excelling at static QA does not necessarily translate into effective trading strategies in dynamic market environments, underscoring a key challenge in the development of LLM-powered financial agents.

The main contributions of this work are summarized as follows:

- We introduce STOCKBENCH, a novel benchmark for evaluating LLM agents in realistic stock-trading environment, directly measuring their profitability and risk-management capabilities.
- We design a comprehensive evaluation framework that incorporates realistic market dynamics, diverse input data, and multiple financial metrics to holistically assess agent performance.
- We conduct extensive experiments by implementing various backbone LLM as stock-trading agents, revealing their current limitations in achieving profitable trading strategies and underscoring the need for further advancements in this domain.
- We open-source implementation of STOCKBENCH to facilitate reproducibility and encourage community contributions, fostering further research on LLM-powered financial agents.

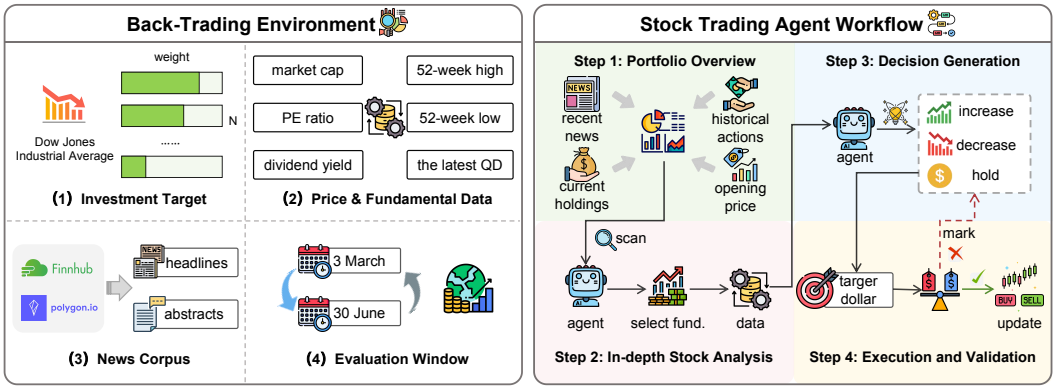

Figure 1: Overview of StockBench. The design of StockBench includes a back-trading benchmark dataset, and an associated workflow that converts backbone LLMs into agents.

2 StockBench

The construction of StockBench consists of two main building blocks. (1) A back-trading environment, which contains historical data necessary for stock-trading decision making. We simulate real-world stock trading using this back-trading setup. (2) An associated stock-trading agent workflow. This workflow allows us to evaluate LLM backbones as agents to engage in the back-trading environment. The overall framework of StockBench is demonstrated in Figure 1.

2.1 Back-Trading Environment

We design the back-trading environment to simulate realistic stock trading, where trading agents are exposed only to data available up to the time of each decision. To set up the environment, we identify three critical sources of information for trading decision making: (1) A bundle of investment targets, which defines the scope of the environment. We pre-define these investment targets to facilitate reproducibility of the evaluation on StockBench. (2) Historical market data, which includes both the prices and fundamental indicators. These enable the evaluated trading agents to perform quantitative analysis. (3) News corpora, which capture events that drive stock price fluctuations. We elaborate on the data collection process below.

Investment Targets. The investment targets are a bundle of stocks that allow the trading agents to perform buy and sell operations. We manually select the investment targets in StockBench to prevent potential outcome fluctuations caused by stock selection—*e.g.*, trading agents might otherwise happen to pick a stock driven by irrational market sentiment—thereby stabilizing the evaluation results.

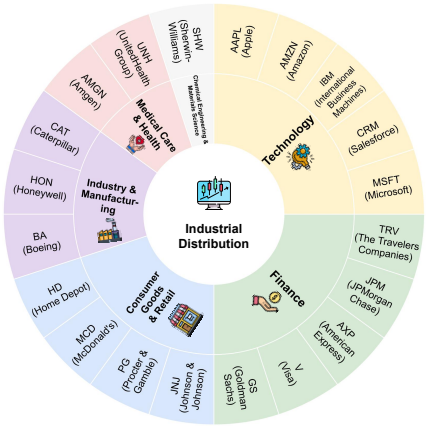

Figure 2: Industry distribution of selected stocks.

To this end, we select 20 stocks from the Dow Jones Industrial Average (DJIA) with the highest weights as our investment targets. In particular, high-weighted DJIA stocks are representative of the global stock market and are less prone to short-term irrational sentiment-driven events. Constraining the trading action space to our selected investment targets mirrors real-world investor attention while keeping the dataset computationally tractable. Moreover, information about these well-known stocks is transparent and easy to collect, being readily accessible through web search engines. We show the distribution of the selected investment targets