Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, et al. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, abs/2503.14476, 2025. URL https://arxiv.org/abs/2503.14476.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, abs/2508.06471, 2025. URL https://arxiv.org/abs/2508.06471.

Yue Zhang, Stefan Zohren, and Stephen Roberts. Deep reinforcement learning for trading. *The Journal of Financial Data Science*, 2(2):25–40, 2020.

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. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3277–3287. Association for Computational Linguistics, August 2021. URL https://aclanthology.org/2021.acl-long.254/.

A ETHICAL STATEMENT

We strictly comply with all applicable financial regulations, data-protection laws, and academic ethical standards during the construction and use of StockBench. All market data (prices, fundamentals, and news) were collected through licensed data vendors or public APIs that explicitly allow research use; no non-public, insider, or personally identifiable information was accessed or stored. The benchmark is provided for academic and non-commercial research purposes only. Users are reminded that StockBench is not intended to offer, or serve as the basis for, any financial advice, trading recommendation, or commercial activity. Any trading strategy tested on StockBench carries inherent market risk; past performance recorded in the benchmark does not guarantee future returns.

B REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work and foster further research in this domain, we plan to open-source our StockBench benchmark, including the dataset and the code for the back-trading workflow. This release will enable other researchers to replicate our experiments, validate our findings, and build upon our methodology.

C PREVENT DATA LEAKAGE

In this study, we minimize the risk of data leakage by carefully planning and evaluating the time frame. When testing large language models (LLMS) in the financial field, a potential concern is that during the training process, the model will learn a lot of past financial knowledge, which may lead to the model's performance being artificially exaggerated. For instance, when asking GPT-5 (without using the search function), we found that the model could accurately predict the stock trend of AAPL in 2021, and the model's response was consistent with the facts.

This discovery indicates that if the evaluation time is relatively early, the model may have obtained future information that could not have been reasonably acquired at the time of evaluation. In view of this, we have decided to limit the data used for evaluation to a more recent time frame, thereby minimizing the possibility of such "data leakage" and ensuring that the model is tested more fairly. By focusing on a narrow evaluation time window, we aim to simulate real-world scenarios where agents can only make trading decisions based on the publicly available information at the time of each decision.

This approach conforms to the best practices of financial model evaluation, ensuring that the evaluation results truly reflect the predictive and decision-making capabilities of LLM agents without being disturbed by the unintentional availability of future data

D Model Return Variance

Table 5: Model Return Variance Across Different Models. This table presents the variance of model returns for various LLMs.

Rank	Model	Var ($\times 10^{-4}$)
1	*DeepSeek-V3*	0.074
2	*DeepSeek-V3.1*	0.203
3	*GPT-5*	0.210
4	*Claude-4-Sonnet*	0.153
5	*GLM-4.5*	0.099
6	*Qwen3-30B-Think*	0.115
7	*Qwen3-235B-Think*	0.321
8	*Qwen3-235B-Ins*	0.281
9	*Qwen3-4B-Ins*	1.382
10	*GPT-OSS-20B*	1.337
11	*Qwen3-Coder*	1.655
12	*Openai-O3*	3.250
13	*Kimi-K2*	1.866
14	*GPT-OSS-120B*	10.19

In this section, we analyze the return variances of different models. Models with higher return variances may exhibit more unpredictable behaviors, which is undesirable in many real-world applications, especially in high-risk environments such as financial decision-making.

We ranked several large language models (LLMS) based on their return variances, as shown in table 5. In the evaluated model, *DeepSeek-V3* exhibited the smallest performance fluctuation, indicating high stability. In contrast, *GPT-OSS-120B* exhibits the highest return variance, indicating a volatility in its performance.

E The Use of Large Language Models

We use LLMs for two purposes. (1) Code implementation. When implementing the code for this paper, including data gathering and experiment implementation, we use LLMs in the form of `copilot` to complete code snippets. The architecture design is conducted by human researchers. (2) Proofreading. To fix grammar issues, we use LLMs as a writing tools to refine the draft.

We would like to highlight that LLMs are not responsible for creativity tasks during conducting the research of this paper, including but not limited to: ideation, experiment design, paper organizing.