# FinNLP-FNP-LLMFinLegal @ COLING 2025 Shared Task: Agent-Based Single Cryptocurrency Trading Challenge

**Yangyang Yu[1], Haohang Li[1], Yupeng Cao[1], Keyi Wang[2], Zhiyang Deng[1], Zhiyuan Yao[1], Yuechen Jiang[1], Dong Li[3], Ruey-Ling Weng[4], Jordan W. Suchow[1,*],**
[1]Stevens Institute of Technology, [2]Northwestern University,
[3]The FinAI, [4]Yale University
[*]Corresponding author
{yyu44,hli113,ycao33}@stevens.edu

## Abstract

Despite the growing potential of large language model (LLM)-based agent frameworks in stock trading, their applicability to comprehensive analysis and multi-asset financial trading, particularly in cryptocurrency markets, remains underexplored. To bridge this gap, we introduce the *Agent-Based Single Cryptocurrency Trading Challenge*, a shared financial task featured at the COLING 2025 FinNLP-FNP-LLMFinLegal workshop. This challenge focuses on two prominent cryptocurrencies: Bitcoin and Ethereum. In this paper, we present an overview of the task and associated datasets, summarize the methodologies employed by participants, and evaluate their experimental results. Our findings highlight the effectiveness of LLMs in addressing the unique challenges of cryptocurrency trading, offering valuable insights into their capabilities and limitations in this domain. To the best of our knowledge, this challenge is among the first to systematically assess LLM-based agents in cryptocurrency trading. We conclude by providing detailed observations and actionable takeaways to guide future research and development in this emerging area.

## 1 Introduction

Large Language Models (LLMs) have showcased remarkable capabilities in text generation (Achiam et al., 2023; Dubey et al., 2024) and reasoning (Wei et al., 2022; Huang and Chang, 2022; Jin et al., 2024b) across various domains, including healthcare (Peng et al., 2023; Jin et al., 2024a) and education (Jia et al., 2024; Liu et al., 2024). These advancements have sparked growing interest within the financial sector. Recent research (Xie et al., 2024b) has highlighted the substantial potential of cutting-edge LLMs in financial Q&A (Islam et al., 2023), financial text analysis (Yang et al., 2024), financial risk prediction (Cao et al., 2024a,b), and financial forecasting (Xie et al., 2024a).

Furthermore, significant research has begun to explore the utilization of LLMs as backbone models for developing agent frameworks to address complex financial decision-making tasks, such as asset trading (Mirete-Ferrer et al., 2022) and market simulation (Li and Yang, 2022; Yao et al., 2024). For instance, FinMem (Yu et al., 2024a) introduces a single-agent framework leveraging an LLM to enhance trading performance by establishing a memory database to store historical trading experiences. Similarly, StockAgent (Zhang et al., 2024) simulates market dynamics by facilitating interactions among multiple agents. FinCon (Yu et al., 2024b) incorporates a reflection mechanism through verbal reinforcement, improving risk management and extending its applicability to multi-asset trading tasks. Despite the notable achievements of LLM-based agent frameworks in stock trading, several critical aspects remain underexplored: 1) The predictive performance across diverse financial assets, such as cryptocurrencies, warrants further investigation; 2) The reliance on closed-source models in existing frameworks necessitates additional validation of open-source models to assess their effectiveness in these contexts.

To further investigate the potential of LLM-based agent frameworks for cryptocurrency trading under an open-source large language model setting, we introduce the *Agent-Based Single Cryptocurrency Trading Challenge* at COLING 2025. This challenge focuses on two leading cryptocurrencies: Bitcoin (BTC) and Ethereum (ETH). For this task, we have curated open-source news datasets for BTC and ETH, enabling participants to evaluate the performance of various open-source models within the FinMem (Yu et al., 2024a) framework. Participants are also permitted to incorporate additional private datasets for pre-training or fine-tuning the backbone open-source LLMs. The goal is to optimize the generation of “buy", “sell", or “hold" decisions by the LLMs and achieve the possibly



*Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal*, pages 401–406
January 19–20, 2025. 

highest trading profits during the designated test period.

This paper provides an overview of the performance of LLM-based agent on cryptocurrency trading, as well as the datasets featured in the *Agent-Based Single Cryptocurrency Trading Challenge*. It summarizes participants' methodologies and evaluates their experimental results to investigate the capabilities of LLMs. Our comprehensive evaluation highlights both the strengths and limitations of current approaches, demonstrating the effectiveness of LLM-based agent frameworks in cryptocurrency trading.

## 2 Challenge Description

### 2.1 Challenge Definition

In this task, participants are required to submit a pre-trained or fine-tuned LLM as the backbone model for conducting daily trading with single cryptocurrencies within the agent framework. We selected FINMEM as the evaluation framework due to its single-agent architecture, which combines comprehensive functionality with precise control over different LLMs. This setup enables clear observations of the trading performance of various LLMs serving as the backbone, thereby facilitating an effective evaluation of open-source models and datasets in cryptocurrency trading.

Participants are allowed to incorporate private datasets and are encouraged to utilize the FINMEM repository on GitHub[1] for model evaluation and selection of optimal training checkpoints. After pre-training or fine-tuning their models, participants can assess their model's performance using FINMEM on the training data. Once satisfactory results are achieved, participants may upload their models to Hugging Face for further testing. Submitted models will undergo final evaluation on a separate test set to ensure robust performance assessment. To support participants, we provide a tutorial code inspired by CATMEMO (Cao et al., 2024c), demonstrating how to efficiently perform fine-tuning, enabling participants to get started more easily. The steps for the challenge are summarized as follows:

- **Pre-training/Fine-tuning Customized Models**: Participants are expected to pre-train or fine-tune their chosen LLMs for cryptocurrency trading. A specific example for fine-tuning is provided in the challenge repository to guide participants.[2]
- **Uploading Models to Hugging Face**: Once participants have finalized their models, they are required to upload them to Hugging Face. Detailed documentation on the uploading process is available.[3]
- **Validation and Leaderboard Updates**: All submitted models will be validated under the FINMEM framework, and performance metrics will be used to rank the models. The leaderboard will be released and updated on the challenge website for participants to track their standings.[4]

### 2.2 Dataset

The dataset for this challenge consists of three elements for each cryptocurrency (BTC and ETH): 1) **Date Information**; 2) **Cryptocurrency to USD Exchange Rates** (floating-point values); 3) **News Articles** (textual data, including sentiment classification). Each data point strictly adheres to the following JSON format:

```
{
 "datetime.date(2022, 11, 29)":{
 "prices": 16444.9832700291,
 "news": ["News Content_1 and
           Sentiment",
              ...
          "News Content_n and
           Sentiment"]}
}
```

Here, the primary key is the date, formatted in DateTime, while the corresponding price and associated news are stored together as a dictionary. The news data is sourced from the *Crypto News Recent* data source, ensuring it is free from copyright restrictions and suitable for academic use. Each day's news includes multiple entries, which are summarized and categorized by sentiment as *positive, negative, or neutral*. The datasets for both BTC and ETH cover the same time intervals:

- **Practice Data Period**: from 2022-11-29 to 2023-01-02.

[1] https://github.com/felis33/coling-cryptocurrency-trading-challenge-evaluation

[2] https://github.com/felis33/coling-cryptocurrency-trading-challenge/blob/main/examples/finetuning_example.ipynb

[3] https://huggingface.co/docs/hub/models-uploading

[4] https://coling2025cryptotrading.thefin.ai/

| Ranking | Team Name | Sharpe Ratio (BTC) | Sharpe Ratio (ETH) | Sharpe Ratio (Overall) |
|---|---|---|---|---|
| 1st | `Sams'Fans` | 2.0694 | 0.8373 | 1.4534 |
| 2nd | `Capybara` | 0.6898 | -0.5752 | 0.0573 |
| 3rd | `300k/ns` | -0.2549 | -0.0252 | -0.1401 |
| Baseline | `B & H` | 1.4403 | 0.9381 | 1.1892 |

Table 1: Team performance based on Sharp Ratio

- **Training Data Period**: from 2023-02-13 to 2023-04-02.

We published Practice Set[5] and Training Set[6] for academic purposes. However, **Testing Set** is reserved for internal assessment to ensure unbiased evaluation of submitted models.

### 2.3 Evaluation Pipeline and Metrics

To evaluate the fine-tuned LLMs, participants can use the FINMEM framework to assess their models' performance on the Training Set. The final competition rankings will be determined by the trading performance of the fine-tuned models on Testing Set, evaluated by the performance metrics in FINMEM.

We provide a comprehensive evaluation of profitability, risk management, and decision-making prowess using a series of metrics. One of the primary metrics is the **Sharpe Ratio (SR)**, which assesses risk-adjusted returns. The SR is mathematically expressed by Equation 1:

$$\textbf{SR} = \frac{R_p - R_f}{\sigma_p} \tag{1}$$

Note that $(R_p - R_f)$ denotes the excess expected return, where $R_p$ is the portfolio's return, $R_f$ is the risk-free rate, and $(\sigma_p)$ is the portfolio's volatility. Higher SR indicate better performance, as they reflect greater returns relative to the risk taken. This metric, along with others, will be used to comprehensively evaluate the fine-tuned models' effectiveness in cryptocurrency trading.

## 3 Participants and Results

A total of 28 teams registered for the *Agent-Based Single Cryptocurrency Trading Challenge*, out of which 5 teams successfully submitted their models for evaluation. Following the release of the leaderboard, three teams managed to outperform the Buy-and-Hold (`B&H`) baseline results, while two teams submitted detailed solution description papers. The rankings and performance of the participating teams are summarized in Table 1. The `Sam's Fans` team secured first place, outperforming the baseline in BTC but failing to do so in ETH. The `Capybara` team finished second, coming close to the baseline in BTC but underperforming in ETH. The `300k/ns` team ranked third, failing to beat the baseline in both BTC and ETH. In this section, we provide a detailed overview of the technical approaches employed by the two teams that submitted solution description papers: `Sam's Fans` and `300k/ns`.

### 3.1 Sam's Fans Team

The `Sam's Fans` team explored the application of fine-tuned LLMs for cryptocurrency trading. The team fine-tuned two state-of-the-art LLMs, LLAMA3.1-8B (Dubey et al., 2024) and QWEN2.5-7B (Qwen Team, 2024), within the FINMEM framework, within the FinMem framework to improve the models' ability to process temporal market data and make effective trading decisions. Motivated by the complexity and volatility of cryptocurrency markets, the team sought to enhance LLM predictive capabilities by integrating domain-specific knowledge and employing a threshold-based decision-making approach. Their methodology involved curating a dataset of domain-specific questions and answers to refine market understanding, followed by fine-tuning the models to make trading decisions based on FinMem-processed data. Their experimental results indicated varying success: the fine-tuned models outperformed the baseline in BTC trading but failed to do so in ETH trading. The authors attributed the improved performance in BTC trading to the models' enhanced ability to analyze market conditions and make informed decisions across different time periods. Their paper concludes by recommending future work on larger models and more advanced

[5] https://drive.google.com/drive/u/1/folders/1Hg_Ee-5NwSy8vdA5eMsTqEAE02w92-zs

[6] https://drive.google.com/drive/u/1/folders/1fr0nBUhpJ0BIo_rukGPWa9skX4Fj_FeY

decision strategies to better integrate static knowledge with dynamic market conditions, aiming to further improve trading performance.

### 3.2 300k/ns Team

The `300k/ns`'s approach integrates sentiment analysis using a pre-trained BERT model (Devlin, 2018), combining textual sentiment with real-time market trends to inform trading decisions. This demonstrates the potential of LLMs in financial decision-making under high-stakes conditions, highlighting significant accuracy and risk management capabilities. The experimental setup features a robust data acquisition and preprocessing pipeline that incorporates sentiment analysis and a deterministic trading strategy based on historical data. Fine-tuning is performed using LoRA (Hu et al., 2021) for efficient adaptation to the financial domain, optimizing computational efficiency while capturing market dynamics. Despite these advancements, the results reveal underperformance in SR, indicating areas for future improvement. The authors suggest enhancing the model's ability to interpret and integrate long-term news trends and broader contextual data to better align predictions with market drivers. This research contributes to the growing application of AI-driven solutions in cryptocurrency trading, offering insights into deploying LLMs in trading scenarios while identifying pathways for improving the reliability and accuracy of automated trading systems.

## 4 Discussion

### 4.1 BTC Performance

The BTC performance in the *Agent-Based Single Cryptocurrency Trading Challenge* varied significantly among the participating teams, as detailed in Table 1. The top-performing team, `Sam's Fans`, achieved a SR of 2.0694, substantially outperforming the `B&H` baseline, which had a SR of 1.4403. This result demonstrates superior risk-adjusted returns, highlighting the effectiveness of their model in navigating BTC's volatility and market dynamics during the challenge period. The second-place team, `Capybara`, achieved a SR of 0.6898, falling short of the `B&H` baseline, indicating that their strategy was less effective at managing risk and leveraging BTC's market trends. The third-place team, `300k/ns`, recorded a negative SR of -0.2549, reflecting underperformance compared to a risk-free investment and suggesting deficiencies in their trading strategy or their model's responsiveness to market conditions. These results underscore the challenge's complexity and the critical importance of advanced model tuning and strategic decision-making in cryptocurrency trading. The wide dispersion in performance illustrates the varying capabilities of LLM-based agent frameworks to adapt to BTC's unique market characteristics.

### 4.2 ETH Performance

The ETH performance presented more challenging conditions for participants. The highest SR, 0.9381, was achieved by the `B&H` baseline, indicating that none of the teams outperformed the baseline in ETH trading. The top-performing team, `Sam's Fans`, achieved a SR of 0.8373, coming close to the baseline but still falling short. `Capybara` and `300k/ns` faced significant difficulties, recording SRs of -0.5752 and -0.0252, respectively. These results may reflect the distinct market dynamics of ETH, characterized by potentially higher volatility and unpredictability compared to BTC, which could have reduced the effectiveness of the deployed models. The findings emphasize the need for enhanced predictive accuracy and more robust risk management strategies to address the volatilities specific to ETH and other cryptocurrencies. The variation in performance underscores the importance of tailoring model development and strategy formulation to align with the unique behaviors of individual cryptocurrency markets.

## 5 Conclusions

In this paper, the *Agent-Based Single Cryptocurrency Trading Challenge* has highlighted the efficacy and potential of LLMs in cryptocurrency trading. By providing a structured framework and extensive resources, the challenge has significantly contributed to advancing research in this domain. Participants leveraged these resources to develop innovative strategies and models, leading to notable improvements in performance across various tasks. The experimental results from BTC and ETH underscore the considerable value of LLM-based approaches, demonstrating their ability to navigate complex market dynamics effectively. A clear trend emerged, indicating that performance improves with increasing model size, as well as advancements in fine-tuning techniques and prompt engineering. These findings provide valuable insights for future research on financial tasks utilizing

LLMs. Moreover, the success of this challenge underscores the importance of collaborative efforts in driving forward the boundaries of AI applications in decentralized finance, offering promising directions for future innovations in the field.

## Acknowledgments

We extend our gratitude to the anonymous reviewers, area chairs, and organizers for their invaluable comments and support throughout this work.

## 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*.

Yupeng Cao, Zhi Chen, Qingyun Pei, Fabrizio Dimino, Lorenzo Ausiello, Prashant Kumar, KP Subbalakshmi, and Papa Momar Ndiaye. 2024a. Risklabs: Predicting financial risk using large language model based on multi-sources data. *arXiv preprint arXiv:2404.07452*.

Yupeng Cao, Zhi Chen, Qingyun Pei, Nathan Lee, KP Subbalakshmi, and Papa Momar Ndiaye. 2024b. Ecc analyzer: Extracting trading signal from earnings conference calls using large language model for stock volatility prediction. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 257–265.

Yupeng Cao, Zhiyuan Yao, Zhi Chen, and Zhiyang Deng. 2024c. Catmemo@ ijcai 2024 finllm challenge: Fine-tuning large language models using data fusion in financial applications. In *Proceedings of the Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning*, pages 174–178.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.

Qinjin Jia, Jialin Cui, Haoze Du, Parvez Rashid, Ruijie Xi, Ruochi Li, and Edward Gehringer. 2024. Llm-generated feedback in real classes and beyond: Perspectives from students and instructors. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 862–867.

Mingyu Jin, Qinkai Yu, Dong Shu, Chong Zhang, Lizhou Fan, Wenyue Hua, Suiyuan Zhu, Yanda Meng, Zhenting Wang, Mengnan Du, et al. 2024a. Health-llm: Personalized retrieval-augmented disease prediction system. *arXiv preprint arXiv:2402.00746*.

Mingyu Jin, Qinkai Yu, Dong Shu, Haiyan Zhao, Wenyue Hua, Yanda Meng, Yongfeng Zhang, and Mengnan Du. 2024b. The impact of reasoning step length on large language models. *arXiv preprint arXiv:2401.04925*.

HaoHang Li and Steve Y Yang. 2022. Impact of false information from spoofing strategies: An abm model of market dynamics. In *2022 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr)*, pages 1–10. IEEE.

Chengyuan Liu, Jialin Cui, Ruixuan Shang, Qinjin Jia, Parvez Rashid, and Edward Gehringer. 2024. Generative ai for peer assessment helpfulness evaluation. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 412–419.

Pedro M Mirete-Ferrer, Alberto Garcia-Garcia, Juan Samuel Baixauli-Soler, and Maria A Prats. 2022. A review on machine learning for asset management. *Risks*, 10(4):84.

Cheng Peng, Xi Yang, Aokun Chen, Kaleb E Smith, Nima PourNejatian, Anthony B Costa, Cheryl Martin, Mona G Flores, Ying Zhang, Tanja Magoc, et al. 2023. A study of generative large language model for medical research and healthcare. *NPJ digital medicine*, 6(1):210.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. 2024a. 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. 2024b. Pixiu: A comprehensive benchmark, instruction dataset and large language model for finance. *Advances in Neural Information Processing Systems*, 36.

Yuzhe Yang, Yifei Zhang, Yan Hu, Yilin Guo, Ruoli Gan, Yueru He, Mingcong Lei, Xiao Zhang, Haining Wang, Qianqian Xie, et al. 2024. Ucfe: A user-centric financial expertise benchmark for large language models. *arXiv preprint arXiv:2410.14059*.

Zhiyuan Yao, Zheng Li, Matthew Thomas, and Ionut Florescu. 2024. Reinforcement learning in agent-based market simulation: Unveiling realistic stylized facts and behavior. *arXiv preprint arXiv:2403.19781*.

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*.

Chong Zhang, Xinyi Liu, Mingyu Jin, Zhongmou Zhang, Lingyao Li, Zhengting Wang, Wenyue Hua, Dong Shu, Suiyuan Zhu, Xiaobo Jin, et al. 2024. When ai meets finance (stockagent): Large language model-based stock trading in simulated real-world environments. *arXiv preprint arXiv:2407.18957*.