300k/ns team at the Crypto Trading Challenge Task: Enhancing the justification of accurate trading decisions through parameter-efficient fine-tuning of reasoning models

Artem Agarkov and **Misha Kulik** and **Leonid Shmyrkov**
artemagarkov2@gmail.com mishakulik2002@gmail.com beasr228@gmail.com

Abstract

In this paper, we address the Agent-Based Single Cryptocurrency Trading Challenge, focusing on decision-making for trading Bitcoin and Etherium. Our approach utilizes fine-tuning a Mistral AI model on a dataset comprising summarized cryptocurrency news, enabling it to make informed "buy," "sell," or "hold" decisions and articulate its reasoning. The model integrates textual sentiment analysis and contextual reasoning with real-time market trends, demonstrating the potential of Large Language Models (LLMs) in high-stakes financial decision-making. The model achieved a notable accuracy, highlighting its capacity to manage risk while optimizing returns. This work contributes to advancing AI-driven solutions for cryptocurrency markets and offers insights into the practical deployment of LLMs in real-time trading environments. We made our model publicly available. [1]

1 Introduction

Cryptocurrency trading has emerged as one of the most dynamic and volatile sectors in the global financial landscape, attracting considerable attention from investors, researchers, and traders alike. The market is characterized by its sensitivity to a vast array of real-time information—from news and social media trends to regulatory updates and technological advancements (Hu et al., 2019). These data streams vary not only in type but also in their timeliness and impact, creating a complex environment that demands quick and accurate decision-making. Various studies (Vargas et al., 2018; Wang et al., 2024a; Wan et al., 2021) show that news data has a significant impact on cryptocurrency prices, so it should be taken into account when making trading decisions. The rapid fluctuations in cryptocurrency prices, driven by both short-term market sentiment and longer-term economic trends, make the need for sophisticated, automated systems even more urgent. In recent years, advancements in Natural Language Processing (NLP) and, more specifically, Large Language Models (LLMs), have significantly improved the ability of machines to understand and process complex data. LLMs like GPT-3 (Brown et al., 2020), LLaMA (Touvron et al., 2023), FinGPT (Yang et al., 2023) and BloombergGPT (Wu et al., 2023) have shown considerable promise in various financial applications, particularly in making informed decisions based on diverse and multi-timely data. However, cryptocurrency trading, with its unique set of challenges (Lopez-Lira and Tang, 2023), requires further advancements (Inserte et al., 2023) in model capabilities. Some studies (Wang et al., 2024b) show that LLM algorithms can effectively extract textual information such as stock correlations, statistical trends and timestamps directly from these stock prices. We think that price correlations and various trends are more related to financial news. This paper addresses the need for specialized LLMs tailored to the cryptocurrency market, capable of interpreting both immediate and long-term market signals, while making reasoned decisions over sustained trading periods. This research aims to evaluate the performance of LLM-based agents in the context of automated cryptocurrency trading, utilizing the FinMem framework (Yu et al., 2023). FinMem is an integrated agent system designed for financial decision-making, leveraging LLMs to support complex trading strategies.

2 Dataset

The datasets provided for this study included historical price data and daily news related to two prominent cryptocurrencies, Bitcoin and Ethereum. Specifically, over a two-month period, the dataset contained daily price information for each cryptocurrency alongside a collection of news articles

[1] https://huggingface.co/agarkovv/CryptoTrader-LM

Proceedings of the Joint Workshop of the 9th FinNLP, the 6th FNP, and the 1st LLMFinLegal, pages 414–422
January 19–20, 2025. ©2025 Association for Computational Linguistics

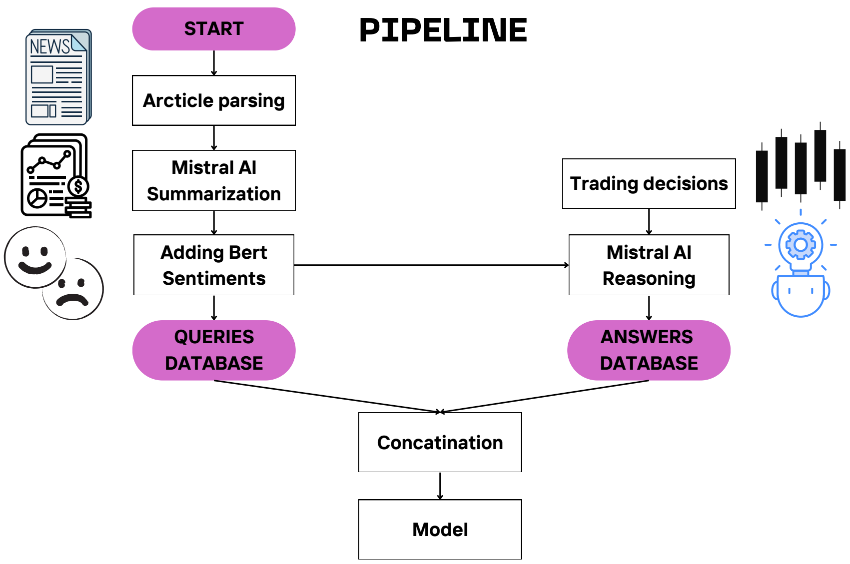

Figure 1: A diagram of the fine-tuning approach

relevant to these assets on each given day. While this data offered a foundational starting point, it was insufficient to train a model capable of capturing the intricate dynamics of cryptocurrency trading. To address this limitation, we extended and enriched the dataset through additional data collection and preprocessing efforts. News from the world of cryptocurrencies was scraped from such sites as The Block, Cnbc, Coindesk, Fortune, Dlnews, Bloomberg, significantly broadening the dataset coverage. By expanding the dataset through these processes, the study ensured that the model had access to a broader context, improving its ability to analyze, reason, and make informed trading decisions in the cryptocurrency market. We compare the suggested and our custom datasets in Figure 2.

3 Approach

The rapid advancements in large language models (LLMs), such as GPT-4 (Achiam et al., 2023) and Mistral (Albert Q. Jiang et al., 2023), highlight the power of integrating diverse data sources during pre-training and fine-tuning. By leveraging multi-faceted datasets, these models achieve remarkable generalization capabilities and domain-specific precision. Using the same approach, we tried to diversify our data as much as possible so that our model knew as much historical information about cryptocurrencies and major players as possible. Our pipeline can be seen in Figure 1.

3.1 Data Pipeline and Preprocessing

Our methodology begins with a robust data acquisition pipeline that collects daily news articles from leading and trusted financial news sources. These articles are parsed and summarized using the Mistral Large model, ensuring the retention of key insights. Each day's summarized content is stored in a query database for systematic access and analysis.

To enrich these textual insights with sentiment analysis, we utilize a pre-trained BERT model (Devlin et al., 2018). BERT processes the summarized articles to classify their sentiment as positive, negative, or neutral. This sentiment tagging is critical for understanding market sentiment trends and their influence on trading behaviors (Nguyen et al., 2015; Bollen et al., 2011).

3.2 Reasoning Through Data Fusion

In parallel with sentiment analysis, trading decisions are computed using a deterministic strategy derived from historical data spanning 2022 to 2024. The strategy involves buying the asset after a decline concludes and before an anticipated rise at the next point, while selling after a growth phase ends and a decline is expected at the following point. These deterministic decisions are then integrated with sentiment-tagged news data to create a rich context for reasoning. To enhance interpretability and decision-making confidence, these trading decisions are explained using human-readable narratives generated by the Mistral Large model.

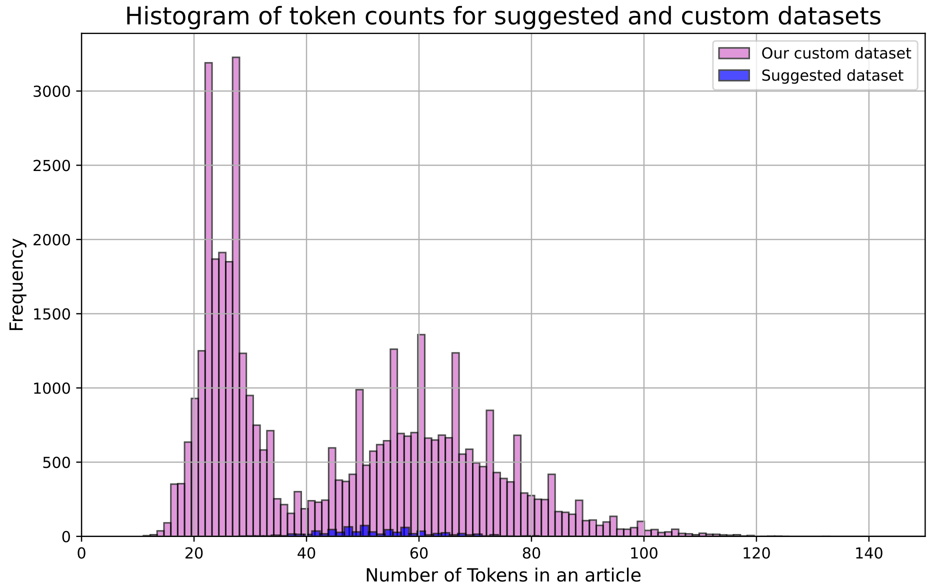

Figure 2: Histogram of token counts for suggested and custom datasets.

This process aligns with recent advances (Wang et al., 2023) in reasoning capabilities for large language models (LLMs), which emphasize the importance of structured reasoning for effective decision-making. The generated explanations are stored in a reasoning database for further use.

3.3 Fine-Tuning for Enhanced Trading Decision Models

To fine-tune our reasoning model, we use trading decisions and their associated reasonings as target outputs, while daily news articles serve as input data. This approach is consistent with strategies aimed at improving stock market forecasting through the integration of sentiment analysis and technical indicators, addressing challenges such as volatility, investor sentiment, and external influences (Shilpa and Shambhavi, 2021; Das et al., 2024). By aligning the training process with these dual targets, we ensure the model captures nuanced relationships between market news and trading outcomes, enhancing its predictive accuracy and interpretability.

4 Experimental Setup

This section provides a detailed explanation of the implementation of the proposed pipeline and the model training process.

4.1 Query Preparation

To enrich the proposed dataset with meaningful data, we scraped significant news articles from January 1, 2022, to September 30, 2024—a span of nearly three years. However, not all articles were fully parsed. For these articles, only their headlines were extracted.

For longer articles, summaries were generated using the Mistral Large model API, ensuring the token distribution matched the overall dataset's news instance characteristics. In total, the dataset consists of 43,553 news instances spanning the specified period. For each instance, both the headlines and parsed article content were annotated with sentiment labels using a BERT model fine-tuned for the financial domain.

4.2 Answer Preparation

We utilized the Mistral Large model API to generate reasoning for daily news batches, given the correct trading decisions for a particular day. The specific prompt used for this task is detailed in Appendix B. The output from the model was stored in an answers database for subsequent use.

4.3 Model Choice

For the base model, we chose the Ministral-8B-Instruct-2410 model, as it shows outstanding performance in the class of small models that do not exceed 8 billion parameters (Chiang et al., 2024).

This model's versatility in instruction-based tasks and its relatively compact size make it ideal for fine-tuning on specific financial datasets like cryptocurrency market news and price data.

4.4 Parameter Efficient Fine-Tuning

To efficiently fine-tune our model, we used low-rank adapters (LoRA) (Hu et al., 2021) for q_{proj} and v_{proj}. In our experiments, we observed that the use of low values of r (e.g. 8) was sufficient to adapt the model to the financial domain, significantly reducing computational overhead.

Furthermore, we performed hyperparameter optimization experiments to fine-tune the rank r and alpha values, leading to the best trade-off between training time and model performance. This allowed us to retain the generalization capabilities of the base model while effectively tuning it for the task of cryptocurrency trading. While techniques like QLoRA (Dettmers et al., 2023) have emerged to further optimize fine-tuning, our approach utilized the default LoRA method for efficiency and simplicity.

4.5 Training Details

We trained the model for both BTC and ETH coins. The dataset was not shuffled to preserve the historical momentum of the data. Training was conducted simultaneously for BTC and ETH, using a maximum context window length of 32,768 tokens. The batch size was set to 1, with 8 gradient accumulation steps.

The learning rate was set to 1×10^{-4} with linear scheduling, and the LoRA parameter weight decay was set to 0.01. The model was trained for 3 epochs for both BTC and ETH. The training process utilized four Nvidia A100 GPUs, each with 40 GB of memory.

4.6 Evaluation Metrics

The evaluation metric for the challenge was **Sharpe Ratio (SR)** (Sharpe, 1994). This is the primary metric for evaluating the risk-adjusted return of the model's trading decisions. A higher Sharpe Ratio indicates that the model is capable of achieving profitable outcomes while managing risk effectively.

5 Results

The results are presented in table 1.

Metric	**Value**
Sharpe Ratio (BTC)	-0.2549
Sharpe Ratio (ETH)	-0.0252
Overall Sharpe Ratio	-0.1401

Table 1: Sharpe ratios for Bitcoin, Ethereum, and the overall model.

We also present the cumulative return graphs for BTC and ETH on the evaluation dataset, covering the period from October 1, 2024, to October 31, 2024. These graphs are provided in Appendix C.

6 Conclusion

This paper presented a novel approach to cryptocurrency trading using fine-tuned large language models (LLMs). By incorporating real-time news sentiment, historical price data, and reasoning capabilities, the CryptoTrader-LM model was able to make informed and reasonable trading decisions for Bitcoin and Ethereum. The use of parameter-efficient fine-tuning techniques, such as LoRA, allowed the model to achieve high accuracy with a relatively small computational footprint.

Our experimental results demonstrated that the model achieved an overall Sharpe ratio of -0.14. The low Sharpe ratios suggest that model's forecasting ability underperformed relative to the risk taken. A likely reason for this is that the model may have struggled to capture a dominant trend or macroeconomic event driving the market. For instance, in November 2024, cryptocurrency growth was largely influenced by macro-political events, such as the election of Donald Trump as president. The model might not have been equipped to fully contextualize or prioritize such long-term news narratives in its decision-making process.

To improve performance in the future, we would like to adjust the model to better interpret and emphasize long-term news trends. By incorporating and analyzing broader contextual information, particularly those tied to significant geopolitical, economic, or technological developments, the model could better align its predictions with prevailing market drivers and enhance its strategic positioning.

References

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

Alexandre Sablayrolles Albert Q. Jiang, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, and Guillaume Lample. 2023. Mistral 7b. *arXiv preprint*, arXiv:2310.06825.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Sentiment analysis on social media for stock movement prediction. *Journal of Computational Science*, Volume 2, Issue 1.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askelland Sandhini Agarwal. 2020. Language models are few-shot learners. *arXiv preprint*, arXiv:2005.14165.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv*, 2403.04132.

Sarmistha Das, R E Zera Marveen Lyngkhoi, Sriparna Saha, and Alka Maurya. 2024. A sophisticated language model solution for financial trading decisions. *ACL Anthology*, Proceedings of the Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning:133–140.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint*, arXiv:2305.14314.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint*, arXiv:1810.04805.

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

Junjie Hu, Wolfgang Karl Härdle, and Weiyu Kuo. 2019. Risk of bitcoin market: Volatility, jumps, and forecasts. *arXiv preprint*, arXiv:1912.05228.

Pau Rodriguez Inserte, Mariam Nakhlé, Raheel Qader, Gaetan Caillaut, and Jingshu Liu. 2023. Large language model adaptation for financial sentiment analysis. *ACL Anthology*, Proceedings of the Sixth Workshop on Financial Technology and Natural Language Processing.

Alejandro Lopez-Lira and Yuehua Tang. 2023. Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint*, arXiv:2304.07619.

Thien Hai Nguyen, Kiyoaki Shirai, and Julien Velcin. 2015. Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications*, 42(24) DOI: 10.1016/j.eswa.2015.07.052.

William F. Sharpe. 1994. The sharpe ratio. *The Journal of Portfolio Management*.

Shilpa and Ravi Shambhavi. 2021. Combined deep learning classifiers for stock market prediction: integrating stock price and news sentiments. *Research gate*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, and Armand Joulin. 2023. Llama: Open and efficient foundation language models. *arXiv preprint*, arXiv:2302.13971.

Manuel Ramon Vargas, Carlos E. M. dos Anjos, Carlos E. M. dos Anjos, and Alexandre Evsukoff. 2018. Deep leaming for stock market prediction using technical indicators and financial news articles. *International Joint Conference on Neural Network (IJCNN)*.

Xingchen Wan, Jie Yang, Slavi Marinov, Jan-Peter Calliess, Stefan Zohren, and Xiaowen Dong. 2021. Sentiment correlation in financial news network and associated market movements. *Scientific Reports*, 3062 (2021).

Meiyun Wang, Kiyoshi Izumi, and Hiroki Sakaji. 2024a. Llmfactor: Extracting profitable factors through prompts for explainable stock movement prediction. *arXiv preprint*, arXiv:2406.10811.

Peiyi Wang, Lei Li, Liang Chen, Feifan Song, Binghuai Lin, Yunbo Cao, Tianyu Liu, and Zhifang Sui. 2023. Making large language models better reasoners with alignment. *arXiv preprint*, arXiv:2309.02144.

Shengkun Wang, Taoran Ji, Linhan Wang, Yanshen Sun, Shang-Ching Liu, Amit Kumar, and Chang-Tien Lu. 2024b. Stocktime: A time series specialized large language model architecture for stock price prediction. *arXiv preprint*, arXiv:2409.082817.

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.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *arXiv preprint*, arXiv:2306.06031.

Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W. Suchow, and Khaldoun Khashanah. 2023. Finmem: A performance-enhanced llm trading agent with layered memory and character design. *arXiv preprint*, arXiv:2311.13743.

A Sample Data

In this appendix, we provide a sample of the dataset used for our cryptocurrency trading model. The data for March 1, 2023, is shown below:


```
 "2023-03-01": {
"prices": 23646.549899145,
"news": [
"the bitcoin market's return to profitability in 2023 is a massive btc bull signal, widely followed on-chain indicator suggests a recent positive shift in the momentum of this key on-chain metric could be a historic buy signal for bitcoin. (sentiment:positive)",
"number of bitcoin wallets with at least 1 btc could soon hit a million... (sentiment:neutral)",
"bitcoin price searches for direction ahead of this week's $710m btc options expiry... (sentiment:neutral)",
"price analysis 3/1: btc, eth, bnb, xrp, ada, doge, matic, sol, dot, ltc... (sentiment:neutral)",
"bitcoin's least volatile month ever? btc price ends february up 0.03    "what is opportunity cost? a definition and examples... (sentiment:neutral)",
"hodlnaut founders propose selling the firm instead of liquidation... (sentiment:neutral)",
"bitcoin 'millionaires' increased 140    "breaking barriers: this protocol brings interoperability and easy swaps across chains... (sentiment:positive)",
"marathon digital bungles crypto impairment sums, will reissue financials... (sentiment:negative)"
]
}
```


Figure 3: Sample Data for 2023-03-01

B Prompt

The following is a multiline text sample illustrating a prompt, given to Mistal model:

```
You are tasked with retrospectively analyzing a correct trading decision (buy, hold, or
sell) for a particular day based on cryptocurrency-related news, historical price
momentum, and investor sentiment. Your goal is to provide clear reasoning for why the
correct decision was made. Just summarize the reason of the decision.

Consider the following:

1. **Short-term Information and Sentiment**: Focus on the short-term crypto news
and market sentiment. Was the news positive or negative? How did it affect market
sentiment in the short term?

2. **Mid-term and Long-term Information**: If mid-term or long-term information is
available (such as regulatory changes, major partnerships, or technological
advancements), consider its relevance. If no such information is available, ignore the
impact of its absence.

3. **Historical Momentum**: Analyze the historical price momentum of the cryptocurrency.
Was the price trend positive or negative in the days leading up to the decision? How did
this momentum influence the decision?
```

```
### Your Task:
Provide reasoning for the **correct** trading decision (buy, hold, or sell) by
analyzingthe following:

- The short-term impact of the news and sentiment.
- The mid-term and long-term information, if available.
- The cryptocurrency's historical momentum and cumulative return.

Your reasoning should clearly explain why **this particular decision (buy, hold, or
sell)** is the most appropriate based on the available information.

Additionally, for each point in your reasoning, provide the **IDs of the information**
that support your decision, but strictly do not just repeat the news, you can only make
judgemets.

News: {news}

Make reasoning for the coin: {coin}

Ticker symbol: {ticker}

Current price: {price}

Price momentum for previous 30 days: {price_momentum}

Correct trading decision: {correct_decision}
```

C Cumulative Return

The charts show that when trading ethereum, the model can qualitatively analyze news and make qualitative predictions. In the case of btc, there were 10% losses in the end. Most likely, this is due to the fact that Ethereum (ETH) has more volatility than Bitcoin (BTC). This is due to the smaller size of the ETH market and its evolving nature. BTC, in turn, has a large market volume and an established reputation, so it experiences less sharp price fluctuations compared to ETH.

ETH turned out to be more suitable for short—term speculative operations, and BTC for hedging or long-term positions. The price of Bitcoin is also very closely tied to the price of the dollar, so analyzing news about the cryptocurrency alone did not allow the model to get the full picture.

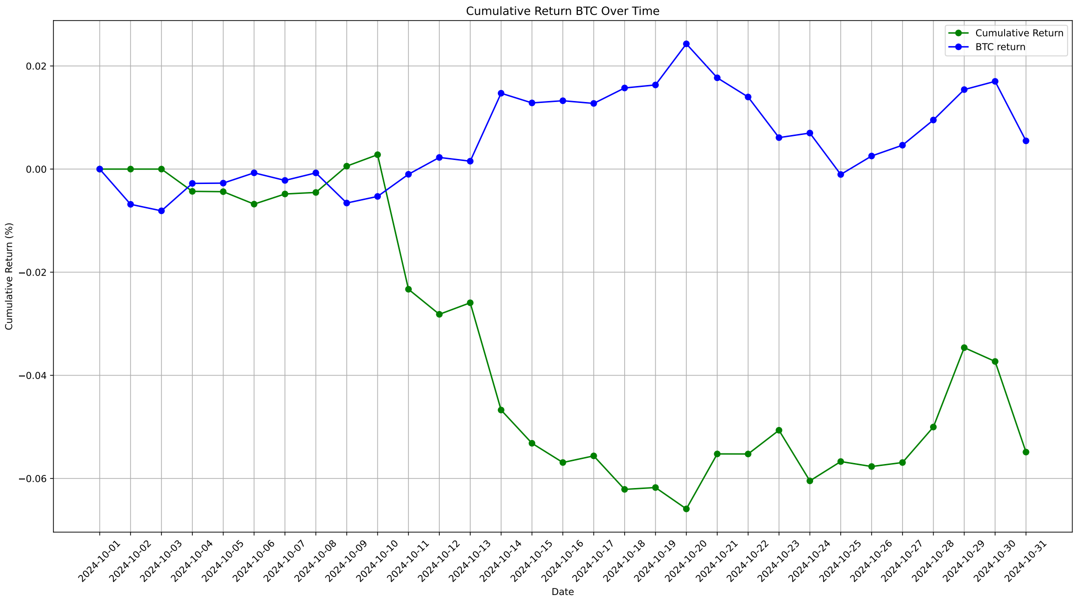

Figure 4: BTC cumulative return

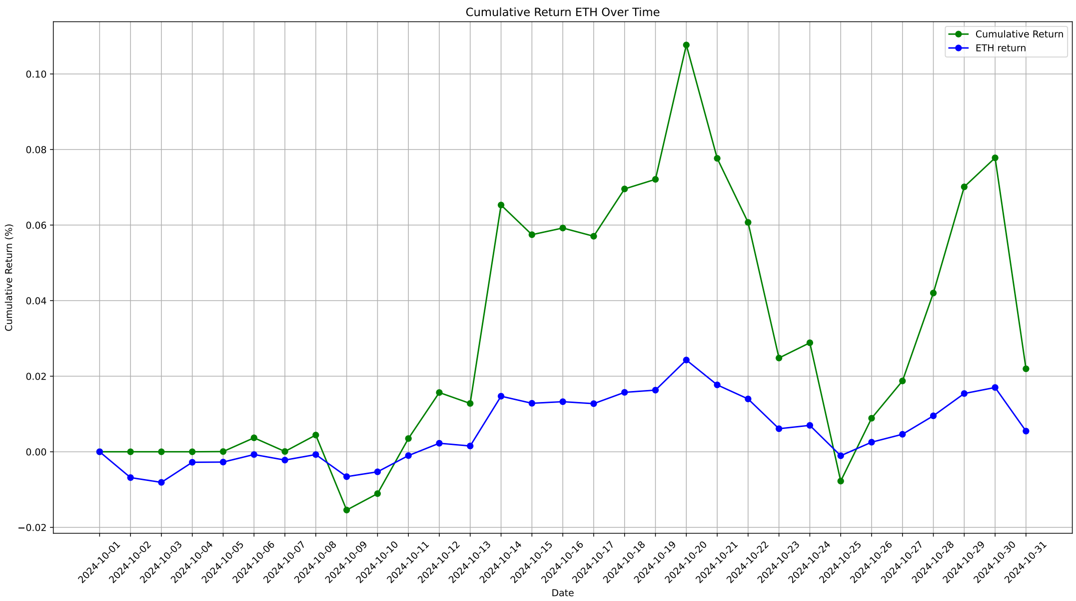

Figure 5: ETH cumulative return