Enhancing LLM Trading Performance with Fact-Subjectivity Aware Reasoning

"The strongest soldiers do not always win the battle." – Ecclesiastes 9:11

Anonymous ACL submission

Abstract

While many studies prove more advanced LLMs perform better on tasks such as math and trading, we notice that in cryptocurrency trading, stronger LLMs work worse than weaker LLMs often. To study how this counter-intuitive phenomenon occurs, we examine the LLM reasoning processes on making trading decisions. We find that separating the reasoning process into factual and subjective components can lead to higher profits. Building on this insight, we introduce a multi-agent framework, FS-ReasoningAgent, which enables LLMs to recognize and learn from both factual and subjective reasoning. Extensive experiments demonstrate that this framework enhances LLM trading performance in cryptocurrency markets. Additionally, an ablation study reveals that relying on subjective news tends to generate higher returns in bull markets, whereas focusing on factual information yields better results in bear markets. Our code and data are available at https://anonymous.4open.science/r/FS-ReasoningAgent-B55F/.

1 Introduction

Large Language Models (LLMs) demonstrate excellent reasoning abilities (Chang et al., 2024) and achieve outstanding performance in fields that require high-level reasoning, such as coding and mathematics (Guo et al., 2023). Recent research also highlights their ability to interpret financial time series and improve cross-sequence reasoning (Wei et al., 2022; Yu et al., 2023; Zhang et al., 2023; Zhao et al., 2023; Yang et al., 2024). Furthermore, the development of LLM-based trading strategies such as Sociodojo (Cheng and Chin, 2024) and CryptoTrade (Li et al., 2024b) underscore the LLM's abilities to make high-return trading decisions based on market statistics and related news.

However, we observe that stronger LLMs have not performed as well as expected in certain trading scenarios (Li et al., 2024b), with their trading strategies that fail to align with weaker LLMs. To validate this observation, we re-implement the experiments in CryptoTrade (Li et al., 2024b) to assess the trading performance of advanced LLMs, including GPT-3.5-turbo[1], GPT-4[2], and GPT-4o[3] on Bitcoin (BTC), Ethereum (ETH), and Solana (SOL). These cryptocurrencies are selected due to their popularity and significant influence in the market. In addition, we select o1-mini[4] as another backbone, as o1 models demonstrate excellent abilities in tasks that require deep reasoning (OpenAI, 2024). The results for Ethereum are presented in Table 1, indicating that not only stronger LLMs (o1-mini, GPT-4o) generate lower overall profits (return), but they also exhibit a less favorable risk-adjusted performance (sharpe ratio) compared to the weaker LLMs (GPT-4, GPT-3.5-turbo) under both bullish and bearish market conditions. Similar trends are observed in the results for BTC and SOL, with all relevant data and experimental details provided in Appendix B. This counter-intuitive results motivate the following research questions:

Strategy	Total Return (%)		Daily Return (%)		Sharpe Ratio	
	Bull	Bear	Bull	Bear	Bull	Bear
Buy and Hold	22.59	-12.24	$0.36_{\pm 2.62}$	$-0.17_{\pm 2.39}$	0.14	-0.07
SMA	10.17	-10.12	$0.18_{\pm 2.29}$	$-0.15_{\pm 1.64}$	0.08	-0.09
SLMA	5.20	-15.90	$0.11_{\pm 2.37}$	$-0.24_{\pm 1.86}$	0.05	-0.13
MACD	7.72	-12.15	$0.13_{\pm 1.22}$	$-0.18_{\pm 1.56}$	0.10	-0.12
Bollinger Bands	2.59	-0.41	$0.04_{\pm 0.40}$	$0.00_{\pm 0.58}$	0.11	-0.01
GPT-3.5-turbo	15.21	-14.40	$0.30_{\pm 2.01}$	$-0.22_{\pm 2.08}$	0.11	-0.10
GPT-4	25.72	-13.72	$0.41_{\pm 2.45}$	$-0.21_{\pm 2.02}$	0.17	-0.10
GPT-4o	25.47	-15.35	$0.40_{\pm 2.25}$	$-0.23_{\pm 2.16}$	0.18	-0.11
o1-mini	20.78	-17.05	$0.34_{\pm 2.49}$	$-0.25_{\pm 2.31}$	0.14	-0.12

Table 1: Performance comparison of CryptoTrade and baseline trading strategies on ETH during both Bull and Bear market conditions.

> *Why stronger LLMs with advanced reasoning ability fail to outperform weaker ones in trading? How to better exploit their advanced reasoning ability?*

To address the above questions, we conduct an in-depth investigation into the reasoning process. Given that news is a significant factor influencing cryptocurrency prices (Aggarwal et al., 2019; Anamika and Subramaniam, 2022; Lee and Jeong, 2023), we examine how

[1] https://platform.openai.com/docs/models/gpt-3-5-turbo

[2] https://platform.openai.com/docs/models/gpt-4

[3] https://platform.openai.com/docs/models/gpt-4o

[4] https://openai.com/index/introducing-openai-o1-preview/

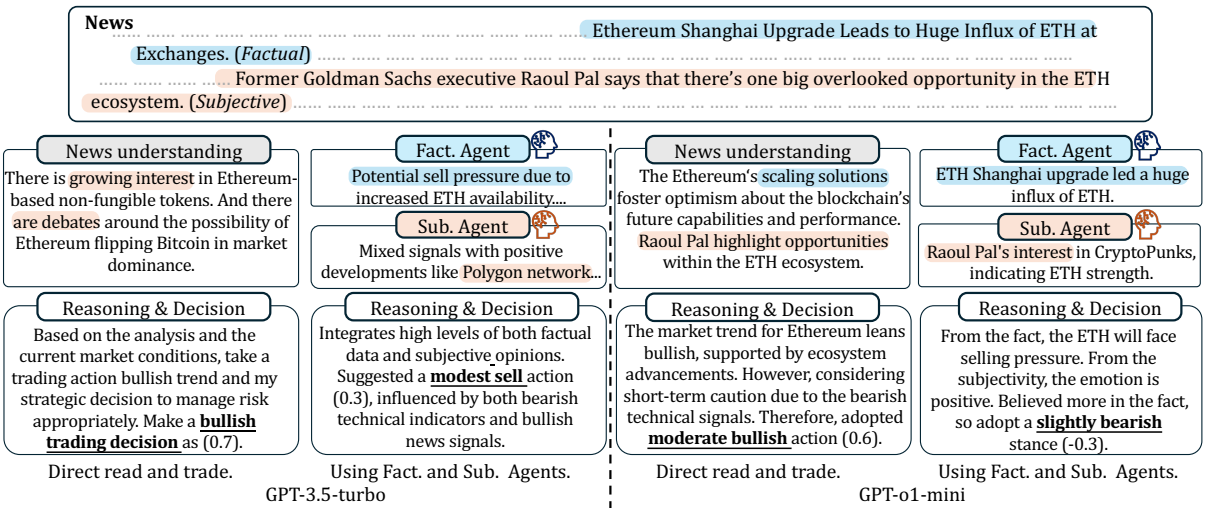

Figure 1: Comparison of Reasoning Processes: With and Without Fact and Subjectivity Agents in Trading Decisions on April 18, 2023, in the ETH Market between GPT-3.5-turbo and o1-mini. The floating numbers represent buy/sell actions: for example, 0.7 indicates using 70% of available cash to buy ETH, while -0.3 indicates selling 30% of held ETH.

LLMs interpret news and make trading decisions. We introduce two agents to separate the news into factual and subjective components. Subsequently, we have the weaker LLM, GPT-3.5-turbo, and the stronger LLM, o1-mini, make decisions based on statistical price data and the refined news. A comparison of their reasoning processes is presented in Figure 1, demonstrating that by separating reasoning into factual and subjective components, the LLMs are able to make more profitable trading decisions. This is particularly evident in this example as the price of ETH on April 18, 2023, was the highest in the following three months.

This comparison also reveals that the stronger LLM tends to focus more on the facts while the weaker LLM give more weight to subjective news. However, as shown in Table 1, the increased reasoning ability does not bring a higher return in the cryptocurrency trading. This outcome aligns with economic theory, which suggests that typical market participants are only partially rational, with investors driven by emotional and psychological factors that push asset prices far beyond stock's intrinsic value (Rubinstein, 2001; Meltzer, 2002).

Motivated by the above investigation and traders can be often motivated by market sentiments (Economou et al., 2023), we propose a novel multi-agent framework, Fact-Subjectivity-ReasoningAgent (FS-ReasoningAgent), which makes trading decisions by reasoning on both factual data and subjectivity. The framework is illustrated in Figure 2. FS-ReasoningAgent splits the reasoning process into a hierarchical structure through multiple agents: (1) dividing raw input data as statistics, factual and subjective news; (2) summarizing and reasoning according to factual or subjective information; (3) trading based on the processed information and reflection; (4) reflecting based on market returns, trading decisions and reasoning processes. FS-ReasoningAgent differentiates itself from previous LLM-based trading agents through its fine-grained reasoning mechanism, which balances both facts and subjectivity, and its adaptive reflection mechanism that dynamically adjusts trading decisions based on past performance. FS-ReasoningAgent allows for more flexible, context-aware decision-making, enhancing the model's adaptability to market conditions.

To evaluate the performance of FS-ReasoningAgent in cryptocurrency trading, we conducted experiments on BTC, ETH, and SOL under both bull and bear market conditions between November 2023 and July 2024. The results show that our approach significantly outperforms CryptoTrade across all three cryptocurrencies in both bull and bear markets, achieving substantial increases in both returns and sharpe ratios. Moreover, FS-ReasoningAgent achieves results comparable to the traditional trading strategy - Buy and Hold, representing a clear improvement not achieved by CryptoTrade. Furthermore, our ablation study of the FS-ReasoningAgent provides two interesting insights: relying on subjective information leads to higher returns in bull markets, while focusing on factual data can result in better performance in bear markets.

Our contribution can be summarized as the following three points:

- We are the first to highlight that stronger LLMs may not necessarily achieve better performance than weaker LLMs on trading tasks. This observation motivates the community to focus more on effectively leveraging the multi-step reasoning capabilities of advanced LLMs.

- We propose a novel multi-agent framework, FS-ReasoningAgent, which separates factual and subjective information and their reasoning pro-

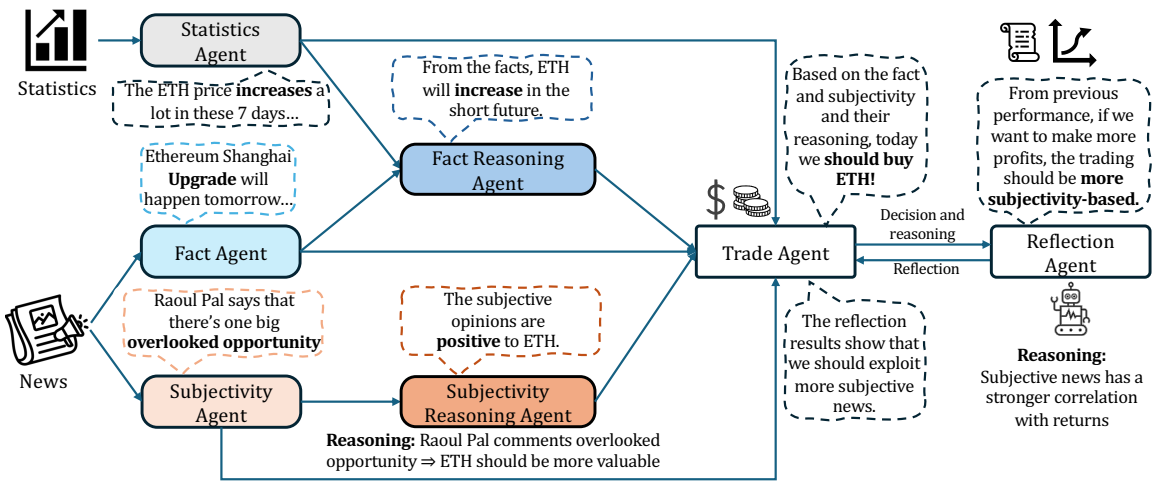

Figure 2: Fact-Subjectivity Reasoning Agent Framework. This framework contains the following agents: Statistics Agent, Fact Agent, Subjectivity Agent, Fact Reasoning Agent, Subjectivity Agent, Trade Agent, and Reflection Agent. We provide an example of each agent's analysis displayed besides the corresponding agent.

cesses. By introducing a reflection agent to guide the weighting of facts and subjectivity, FS-ReasoningAgent automatically adjusts its reasoning preferences based on market returns, enhancing trading strategies in both bull and bear markets.

- Through rigorous experiments, we provide empirical evidence demonstrating the efficacy of FS-ReasoningAgent compared to other baselines. Our ablation study highlights the different performance of factual and subjective information in bull and bear markets, providing valuable insights for future researchers aiming to make trading decisions in the market.

2 FS-ReasoningAgent Framework

In this section, we first provide the data collection process in our experiments and the FS-ReasoningAgent framework. Then, based on the experiment results, we analyze why stronger LLMs with advanced reasoning ability fail to outperform weaker ones. Then, built upon our analysis, we design the FS-ReasoningAgent framework as shown in Figure 2.

2.1 Data Collection

We collect data from various open-source websites. The licensing information for each source is detailed in Appendix C, and the ethical requirements are explained in Appendix D. The specific details of the data are as follows:

- **Statistics:** We collect historical data from CoinMarketCap[5], which provides daily insights into prices, trading volumes, and market capitalization of BTC, ETH, and SOL. For each day, we collect the opening price, closing price, transaction volume, average gas fees, the number of unique addresses, and the total value transferred on the cryptocurrency.

- **News:** We employ the Gnews API[6] to collect the news. The news dataset includes articles related to the cryptocurrencies, including Bitcoin, Ethereum, and Solana, to ensure comprehensive and diverse coverage. The process starts from defining daily intervals within the date range we specified. For each day, relevant news articles are searched using the name of each cryptocurrency as a keyword, ensuring the news language is in English. A particular focus is placed on filtering news from reputable financial and cryptocurrency-specific news sources such as Bloomberg, Yahoo Finance, and crypto.news to guarantee the reliability and relevance of the information. Finally, we get a high-quality, well-organized dataset that supports in-depth analysis of cryptocurrency news trends and developments.

2.2 Demystifying LLM Reasoning in Trading

In the context of financial markets, the news plays a pivotal role in influencing the prices of various assets (Goldstein, 2023; Dhingra et al., 2024). Due to the huge message length like news, the input messages are actually very long contexts for LLMs. Stronger LLMs normally have better in-context learning abilities. However, due to their current limitations of the long-context understanding problems (Li et al., 2024a; Jin et al., 2024), it is difficult for LLMs to simultaneously understand all news and make precise trading decisions. Thus, even for stronger LLMs, some important information that hide in the news sea cannot be well captured like the Needle-In-A-Haystack experiments (Kuratov et al., 2024; Wang et al., 2024; Laban et al., 2024).

As shown in Figure 1, LLMs understand news content differently. Not all valuable information is captured by

[5] https://coinmarketcap.com

[6] https://pypi.org/project/gnews/

LLMs, leading to sub-optimal trading decisions. Built upon this observation, we introduce two separate agents that are responsible to extract factual and subjective components from the news, respectively. Thus, each individual agent now has more capacity to focus on extract information. Based on the processed information from the factual and subjective agents, the trading agent has better summarized information to make trading decisions. We will illustrate the detailed agent design in subsequent sections.

You are an eth statistics agent. The recent price and auxiliary information is given in chronological order below:
{Open price: 2241.75, unique addresses: 577757, ...;
Open price: 2317.97, unique addresses: 576510, ...}
Write one concise paragraph to analyze the recent information and estimate the statistical trend accordingly.

Responses: Over the recent period, Ethereum's open price has demonstrated a notable upward trend...towards continued positive momentum in Ethereum's market performance.

Figure 3: An example of Statistics Agent.

You are an eth fact agent. You are required to analyze only the factual news, not the subjective news such as someone's comments from following news articles:
{'title': 'XRP, Bitcoin and Ethereum Eye Unusual Transfer Activity Worth Millions...}
Write one concise paragraph to analyze and summarize the factual news and estimate the market trend accordingly.

Responses: Recent significant movements in the Ethereum ecosystem highlight sustained investor activity...positive trend of ETH.

Figure 4: An example of Fact Agent.

You are an eth trading analyst. You are required to analyze only the subjective news such as someone's comments from following news articles:
{'title': 'Analyst projects Ethereum potentially to 2k...}
Write one concise paragraph to analyze the subjective news and estimate the market trend accordingly.

Responses: FxPro senior analyst Alex Kuptsikevich forecasts a potential decline of ETH to 2k...ETH will appear downward.

Figure 5: An example of Subjectivity Agent.

2.3 Component Design of FS-ReasoningAgent

After data collection and analyzing LLM reasoning in trading, we introduce each component of the FS-ReasoningAgent, demonstrating how the framework makes its trading decisions, as illustrated in Figure 2.

You are an eth fact reasoning agent. Based on the following factual news analysis and statistics analysis, provide your reasoning for the trading suggestions. This reasoning will be used for the final trading action.
Factual News Analysis: {Fact Agent Responses}
Statistics Analysis: {Statistics Agent Responses}

Responses: The following factors: Liquidity Influx, Technological Advancements, ...ETH exhibits positive growth trajectory.

Figure 6: An example of Fact Reasoning Agent.

You are an eth subjectivity reasoning agent. Based on the following subjective news summary and analysis, provide your reasoning for the trading suggestions. This reasoning will be used for the final trading action.
Subjective News Analysis: {Subjectivity Agent Responses.}

Responses: Given influencers highlighting ETH vulnerability...immediate market conditions warrant a risk-managed approach.

Figure 7: An example of Subjectivity Reasoning Agent.

You are an experienced eth trader and you are trying to maximize your overall profit by trading eth. In each day, you must make an action to buy or sell eth. You are assisted by a few agents below and need to decide the final action.
STATISTICS AGENT REPORT: "{REPORT.}"
...
REFLECTION AGENT REPORT: "{REPORT.}"
Now, provide your response in the following format:
1. Reasoning: Briefly analyze the given reports, considering both factual and subjective elements. Pay attention to the reflection report's guidance.
2. Factual vs Subjective Weighting: If there's a conflict between factual and subjective information, explain which you favor and why. Assign weights to factual and subjective information (e.g., 0.7 factual, 0.3 subjective) that sum to 1. Justify your weighting on the reflection report's recommendation to maximize returns.
3. Risk Management: Describe how you're managing risk in light of the current market conditions and the reflection report's insights.
4. Action: Indicate your trading action as a 1-decimal float in the range of [-1,1]. This should reflect your confidence in the market trend, your strategic decision to manage risk, and the balance between factual and subjective reasoning.

Responses: Action: -0.4...Slight sell to reduce exposure while acknowledging underlying network strength and current bearish sentiment.

Figure 8: An example of Trade Agent.

You are an eth reflection agent. Reflect on your recent trading performance and provide guidance for future trades: ...

Responses: To maximize trading performance in the current Ethereum market conditions, maintain a balanced approach with approximately 60% weighting on factual information and 40% on subjectivity...

Figure 9: An example of Reflection Agent.

2.3.1 Statistics Agent

Statistics Agent is responsible for extracting, analyzing, and summarizing key market data related to cryptocurrencies. It reads various quantitative metrics such as the opening price, total transaction volume, average gas fees, unique addresses, and total value transferred on the cryptocurrency. Based on this data, the Statistics Agent identifies short-term market trends and provides an essential foundation for the overall trading strategy. This agent plays a vital role in ensuring that trading decisions are grounded in up-to-date, quantifiable market conditions. An example of Statistics Agent is shown in Figure 3.

2.3.2 Fact Agent

Fact Agent focuses on gathering and analyzing factual news related to the cryptocurrency market. It filters out subjective commentary, relying instead on concrete events such as regulatory updates, technological advancements, and major market shifts. The Fact Agent improves the trading decision process by identifying impactful facts, such as Ethereum's technological progress and regulatory updates related to Ethereum ETFs shown in the news. This information is crucial for generating rational trading strategies, as it provides context on real-world factors that can influence the market. An example of Fact Agent is shown in Figure 4.

2.3.3 Subjectivity Agent

Subjectivity Agent is tasked with extracting and analyzing subjective opinions, sentiments, and expert commentary from news sources. This agent gathers market rumors, social media trends, and expert predictions that reflect emotions and biases in the market. By capturing market sentiment, this agent helps identify how emotional responses may drive short-term price volatility and offers insights into potential market shifts driven by psychological factors. An example of a subjectivity agent is shown in Figure 5.

2.3.4 Fact Reasoning Agent

Fact Reasoning Agent synthesizes information from both the Statistics Agent and the Fact Agent to provide logical, data-driven reasoning about current market conditions. This agent ensures that the trading strategy is grounded in objective data. This agent's role is to ensure that the factual information is processed to generate actionable insights for trading decisions. An example of Fact Reasoning Agent is shown in Figure 6.

2.3.5 Subjectivity Reasoning Agent

Subjectivity Reasoning Agent interprets the subjective insights gathered by the Subjectivity Agent, offering a more fine-grained analysis on market trends. This agent considers how emotions, biases, and opinions may influence market movements and price volatility. By reasoning in these subjective elements, this agent provides a complementary layer of reasoning to fact-based analysis, enriching the overall decision-making process. An example of Subjectivity Reasoning Agent is shown in Figure 7.

2.3.6 Trade Agent

Trade Agent serves as the decision-making core of the FS-ReasoningAgent framework, synthesizing inputs from the Statistics Agent, Fact Agent, Fact Reasoning Agent, and Subjectivity Reasoning Agent to make final trading decisions. Converts the collective analysis into an actionable decision, represented on a continuous scale from $[-1, 1]$, where -1 means a full sell action, 0 represents a hold, and 1 indicates a full buy action. Trade Agent carefully balances factual data and subjective sentiment to optimize trades for profit while managing risk. Upon executing a trade, a proportional transaction fee is applied based on the value traded. An example of Trade Agent is in Figure 8.

2.3.7 Reflection Agent

Reflection Agent plays a critical role in learning and adapting the FS-ReasoningAgent's trading strategy over time. It reviews past trading actions and outcomes, analyzing the effectiveness of the reasoning process and the information used in decision-making. By examining recent prompts, decisions, and market returns, the Reflection Agent identifies which types of information—factual data or subjective opinions—had the most significant impact on trading success. This feedback loop allows the system to adjust future strategies, improving performance by focusing on the most influential factors. An example of this reflective process is illustrated in Figure 9.

3 Experiments

In this section, we detail the experiments designed to evaluate the performance of FS-ReasoningAgent in comparison to established baseline strategies in the cryptocurrency trading domain.

3.1 Experimental Setup

Datasets. To ensure our experiments are robust across different cryptocurrencies and market conditions, we base our study on a dataset covering several months, detailed in Table 2. This dataset captures the recent market performance of BTC, ETH, and SOL, highlighting challenges in identifying market trends and volatility. We divide the dataset into validation and test sets, using the validation set to fine-tune model hyperparameters and prompts, and the test set to evaluate model performance. The data period, spanning from November 2023 to July 2024, is carefully chosen to prevent data leakage, as all GPT models have a knowledge cutoff prior to November 2023[7]. The dataset covers both bull and bear markets, allowing us to assess the effectiveness of both the baseline models and our proposed model (Baroiu et al., 2023; Cagan, 2024; Li et al., 2024b), ensuring

[7] https://openai.com/api/pricing/

Type	Split	Start	End	Open	Close	Trend
BTC	Validation	2023-11-16	2024-01-15	37879.97	42511.96	12.23%
	Test Bullish	2024-01-24	2024-03-13	39877.59	71631.35	79.63%
	Test Bearish	2024-05-21	2024-07-13	71443.06	59231.95	-17.09%
ETH	Validation	2023-11-10	2024-01-08	2121.06	2333.03	9.99%
	Test Bullish	2024-01-24	2024-03-13	2241.74	4006.45	78.72%
	Test Bearish	2024-05-27	2024-07-08	3826.13	2929.86	-23.42%
SOL	Validation	2023-11-16	2024-01-08	65.53	97.79	49.18%
	Test Bullish	2024-01-24	2024-03-13	84.28	151.02	77.35%
	Test Bearish	2024-05-21	2024-07-11	186.51	127.61	-15.53%

Table 2: Dataset splits with prices in US dollars. Each split includes the start date but excludes the end date for transaction days. The total profit is evaluated on the end date.

reliable and robust experimental results across varying market conditions.

Evaluation Metrics. We initialize the FS-ReasoningAgent with a starting capital of one million US dollars, evenly split between cash and BTC/ETH/SOL, allowing it to capitalize on both buying and selling opportunities in the cryptocurrency market. At the end of the trading session, we assess performance using the following commonly accepted metrics: **Return**, **Sharpe Ratio**, **Daily Return Mean**, and **Daily Return Std**. This evaluation approach ensures a thorough and unbiased comparison between FS-ReasoningAgent and baseline strategies.

(1) Return measures the overall performance of the trading strategy, calculated as $\frac{w^{end}-w^{start}}{w^{start}}$, where w^{start} and w^{end} denote the initial and final net worth, respectively.

(2) Sharpe Ratio evaluates the risk-adjusted return, computed as $\frac{\bar{r}-r_f}{\sigma}$, where $\bar{r}$ is the average daily return, σ is the standard deviation of daily returns, and r_f is the risk-free return, which we set to 0, following SocioDojo (Cheng and Chin, 2024).

(3) Daily Return Mean reflects the average daily performance of the trading strategy over the trading period.

(4) Daily Return Std represents the standard deviation of daily returns, indicating the volatility and risk associated with the strategy's daily performance.

Baseline Strategies. To benchmark FS-ReasoningAgent's performance, we compare it against widely recognized baseline trading strategies. The baselines are detailed in Appendix E.

3.2 Experimental Results

The performance comparison is presented in Table 3, Table 4, Table 5 between various trading strategies and FS-ReasoningAgent. We highlight the findings from the results in the following two key aspects.

FS-ReasoningAgent's Superior Performance under Different Market Conditions. The results across multiple experiments show that the FS-ReasoningAgent consistently outperforms CryptoTrade in both bull and bear markets for BTC, ETH, and SOL. In the case of BTC, during a bull market, FS-ReasoningAgent (GPT-4) achieves a total return of 77.47%, surpassing CryptoTrade's best-performing strategy (GPT-3.5-turbo) by approximately 7%. A similar trend is observed in the bear market, where FS-ReasoningAgent (GPT-4o) achieves a loss of only -13.94%, outperforming CryptoTrade's GPT-4o strategy, which shows a larger decline of -20.21%. This pattern is also evident in the Sharpe ratio, where FS-ReasoningAgent maintains a relatively stable performance in both bull (0.46) and bear (-0.23) conditions compared to CryptoTrade.

In SOL's bull market, FS-ReasoningAgent's superiority is even more striking. The best-performing model, FS-ReasoningAgent (o1-mini), delivers a total return of 76.71%, which is more than 10% higher than CryptoTrade's top model (GPT-3.5-turbo) at 66.64%. Additionally, FS-ReasoningAgent also provides a more favorable sharpe ratio (0.36) compared to CryptoTrade's highest value of 0.34. This suggests that while the Buy and Hold strategy remains slightly ahead in total return during bull markets, FS-ReasoningAgent provides a robust alternative, staying within 2% of the best traditional strategy while significantly outperforming CryptoTrade. These findings emphasize the necessity and effectiveness of FS-ReasoningAgent's design to leveraging both factual and subjective data and reasoning.

FS-ReasoningAgent Makes Stronger LLMs Great Again. The performance of FS-ReasoningAgent not only shows an advantage in overall returns but also highlights how it effectively utilizes stronger LLMs, such as GPT-4 and o1-mini, which CryptoTrade fails to fully harness. For example, in BTC's bear market, CryptoTrade's GPT-3.5-turbo produces the best results within the CryptoTrade framework, achieving a loss of -18.08%. However, FS-ReasoningAgent's GPT-4o reduces this loss significantly to -13.94%, demonstrating a clear advantage. Similarly, in SOL's bear market, FS-ReasoningAgent (GPT-4o) achieves a much smaller loss of -14.52% compared to CryptoTrade's best-performing model (GPT-3.5-turbo) at -22.35%. This suggests that the CryptoTrade framework may underutilize the advanced reasoning capabilities of stronger LLMs, while FS-ReasoningAgent maximizes their potential by systematically processing factual and subjective inputs. As

Strategy	Total Return		Daily Return		Sharpe Ratio	
	Bull	Bear	Bull	Bear	Bull	Bear
Buy and Hold	**79.63**	-19.15	1.18±2.21	-0.38±1.79	**0.53**	-0.21
SMA	69.51	-9.80	1.09±2.57	-0.19±0.76	0.43	-0.25
SLMA	53.09	-8.30	0.89±2.49	-0.16±0.97	0.36	-0.16
MACD	22.01	-15.26	0.41±1.28	-0.29±1.66	0.32	-0.18
Bolling Bands	8.28	**-6.10**	0.16±0.51	**-0.11**±1.01	0.32	**-0.11**
CryptoTrade(GPT-3.5-turbo)	70.25	-18.08	1.12±2.53	-0.36±1.75	0.44	-0.21
CryptoTrade(GPT-4)	66.83	-21.11	1.08±2.21	-0.43±1.66	0.39	-0.26
CryptoTrade(GPT-4o)	68.35	-20.21	1.10±2.57	-0.41±1.68	0.43	-0.24
CryptoTrade(o1-mini)	70.83	-19.89	1.13±2.58	-0.40±1.61	0.44	-0.25
Ours(GPT-3.5-turbo)	73.55	-19.15	1.16±2.61	-0.39±1.71	0.23	-0.23
Ours(GPT-4)	77.47	-15.23	**1.21**±2.63	-0.30±1.20	0.46	-0.25
Ours(GPT-4o)	74.27	-13.94	1.17±2.60	-0.28±0.85	0.45	-0.33
Ours(o1-mini)	76.19	-15.91	1.20±2.62	-0.32±0.93	0.46	-0.35

Table 3: Performance of each strategy on BTC under both bull and bear market conditions. For each market condition and metric, the best result is highlighted in bold, the runner-up is indicated with an underline, and the best result among each families of LLM-based strategies is highlighted in green.

Strategy	Total Return (%)		Daily Return (%)		Sharpe Ratio	
	Bull	Bear	Bull	Bear	Bull	Bear
Buy and Hold	**78.72**	-23.63	1.18±2.21	-0.60±2.13	0.53	-0.28
SMA	59.60	-19.13	0.96±2.11	-0.49±1.00	0.45	-0.49
SLMA	60.31	**-9.01**	0.97±2.07	**-0.21**±1.34	0.47	**-0.16**
MACD	12.93	-20.10	0.25±0.78	-0.50±2.00	0.32	-0.25
Bollinger Bands	77.24	-23.68	1.17±2.20	-0.60±2.12	0.53	-0.29
CryptoTrade(GPT-3.5-turbo)	74.83	-22.35	1.17±2.20	-0.58±1.94	0.53	-0.30
CryptoTrade(GPT-4)	74.41	-23.06	1.17±2.20	-0.60±2.08	0.53	-0.29
CryptoTrade(GPT-4o)	74.23	-24.13	1.16±2.18	-0.63±2.11	0.53	-0.30
CryptoTrade(o1-mini)	75.01	-23.68	1.17±2.19	-0.62±2.15	0.53	-0.29
Ours(GPT-3.5-turbo)	71.09	-22.33	1.12±2.15	-0.58±1.92	0.52	-0.30
Ours(GPT-4)	76.67	-23.41	1.19±2.21	-0.61±1.97	**0.54**	-0.31
Ours(GPT-4o)	76.74	-21.64	1.19±2.21	-0.56±1.82	**0.54**	-0.31
Ours(o1-mini)	77.28	-21.88	**1.20**±2.22	-0.57±1.79	**0.54**	-0.32

Table 4: Performance of each strategy on ETH under bull and bear market conditions.

Strategy	Total Return (%)		Daily Return (%)		Sharpe Ratio	
	Bull	Bear	Bull	Bear	Bull	Bear
Buy and Hold	**77.35**	-24.08	**1.23**±3.39	-0.45±3.97	**0.36**	-0.11
SMA	42.09	-27.17	0.74±2.65	-0.58±2.37	0.28	-0.24
SLMA	47.84	-18.92	0.83±2.93	-0.39±1.74	0.28	-0.22
MACD	34.63	-15.44	0.62±2.17	-0.29±2.58	0.29	-0.11
Bollinger Bands	22.97	**-8.94**	0.42±1.23	**-0.13**±3.15	0.34	**-0.04**
CryptoTrade(GPT-3.5-turbo)	66.64	-23.56	1.10±3.25	-0.45±3.77	0.34	-0.12
CryptoTrade(GPT-4)	32.59	-21.51	0.61±2.65	-0.41±3.65	0.23	-0.11
CryptoTrade(GPT-4o)	48.41	-24.63	0.84±2.52	-0.48±3.83	0.33	-0.13
CryptoTrade(o1-mini)	42.48	-21.95	0.76±2.60	-0.43±3.40	0.29	-0.13
Ours(GPT-3.5-turbo)	68.03	-24.67	1.12±3.27	-0.49±3.55	0.34	-0.14
Ours(GPT-4)	64.35	-25.33	1.07±3.25	-0.52±3.07	0.33	-0.16
Ours(GPT-4o)	69.67	-14.52	1.14±3.30	-0.26±3.05	0.35	-0.09
Ours(o1-mini)	76.71	-19.40	1.22±3.38	-0.36±3.40	**0.36**	-0.11

Table 5: Performance of each strategy on SOL under bull and bear market conditions.

a result, stronger LLMs, which initially lagged in CryptoTrade, demonstrate superior performance when employed in the FS-ReasoningAgent framework, thus unlocking their full capabilities.

3.3 Ablation Study

To evaluate the contribution of each agent in the FS-ReasoningAgent, we conduct an ablation study using

Components	Return (%)		Sharpe Ratio	
	Bull	Bear	Bull	Bear
Full	76.19	-15.91	0.46	-0.35
w/o Reflection Agent	71.77	-17.85	0.44	-0.40
w/o Fact Reasoning Agent	72.23	-19.21	0.43	-0.39
w/o Sub. Reasoning Agent	66.04	-16.83	0.42	-0.36
w/o Statistics Agent	74.25	-20.40	0.45	-0.36

Table 6: Ablation study of each agent's performance under BTC bull and bear market conditions.

the o1-mini backbone on BTC in both bull and bear markets. In each iteration, we remove one component from the full framework and observe the resulting changes in return and sharpe ratio. The results are presented in Table 6. From this ablation study, we have two interesting insights.

Insight 1: Subjectivity is More Important in the Bull Market. The performance in the bull market, reflected by both returns and the sharpe ratio, suggests that subjective reasoning plays a crucial role in capturing the market's positive sentiment. Removing the Subjective Reasoning Agent results in a notable drop in returns from 76.19% to 66.04%, along with the largest decline in the sharpe ratio from 0.46 to 0.42. This indicates that in bullish markets, understanding and interpreting market sentiment—such as reactions to news, emotions, and trends—is essential for maximizing profits.

Insight 2: Facts are More Important in the Bear Market. In bear markets, factual reasoning plays a critical role in minimizing losses. The study shows that removing the Fact Reasoning Agent leads to a deeper negative return of -19.21%, compared to -15.91% for the full framework. Similarly, the sharpe ratio drops from -0.35 to -0.39 without the factual component. A similar pattern is observed when the Statistics Agent is removed, causing the largest decrease in returns from -15.91% to -20.40%, as statistical data also represent factual insights. This highlights the importance of relying on clear data and objective analysis during bearish periods, when fear and pessimism dominate. In these scenarios, decisions based on emotions or subjective reasoning can amplify losses, while focusing on facts helps mitigate risks.

4 Related Work

LLMs for Economics and Financial Decisions Recent progress in LLMs has had a notable impact on economics and financial decision-making. Models specifically designed for finance, such as FinGPT, BloombergGPT, and FinMA, FinAgent (Liu et al., 2023; Wu et al., 2023; Xie et al., 2023; Zhang et al., 2024a), have been applied to tasks like sentiment analysis, entity recognition, and making trading decisions. LLM-driven agents for financial trading have also drawn considerable attention. The Sociodojo framework (Cheng and Chin, 2024), for instance, developed analytical agents for managing stock portfolios, demonstrating the potential for creating "hyper-portfolios." Although there are studies focused on cryptocurrency markets, many models still struggle to outperform a basic buy-and-hold strategy, revealing the need for further improvements (Li et al., 2024b; Yu et al., 2024).

Reasoning Process of LLM Agents A common method for examining the reasoning process of LLMs involves generating intermediate reasoning steps using techniques such as chain-of-thought reasoning (Wei et al., 2022; Kojima et al., 2022) and question decomposition (Zhou et al., 2022). However, a key challenge remains the absence of a standardized benchmark for evaluating LLM agents' reasoning abilities (Zhang et al., 2024b). Additionally, the reasoning process behind LLMs' trading decisions has been largely unexplored (Ding et al., 2024). To address these gaps, we propose a FS-ReasoningAgent designed to evaluate LLM agents' reasoning, focusing on how they incorporate both fact and subjectivity when making decisions in cryptocurrency markets. This framework aims to clarify how LLMs reason through trading decisions, providing valuable insights that can guide future research in this field.

Reflection in LLM Agents LLMs are able to reflect upon their own CoT and produce guidance that can significantly improve problem-solving performance (Renze and Guven, 2024). Recent advancements in automatic prompt refinement (Pryzant et al., 2023; Ye et al., 2024) and the use of automated feedback to enhance reasoning (Paul et al., 2023) have further pushed the boundaries in this area. A key development is the "Reflexion" framework by (Shinn et al., 2024; Cheng and Chin, 2024), which transforms the reinforcement of language agents by incorporating linguistic feedback and reflective text stored in an episodic memory buffer, rather than relying on traditional weight updates. These innovations emphasize the ability of LLMs to learn from mistakes and improve through self-reflection.

5 Conclusion

We introduce FS-ReasoningAgent, a novel multi-agent framework that enhances LLM-based cryptocurrency trading by integrating both factual data and subjective market sentiment. Our findings challenge the assumption that stronger LLMs always outperform weaker ones, demonstrating that advanced reasoning alone does not guarantee consistently better trading outcomes. FS-ReasoningAgent improves decision-making by dynamically adjusting its reliance on factual and subjective information based on market conditions, enabling it to adapt effectively to both bull and bear markets. Experimental results show that FS-ReasoningAgent outperforms the previous LLM-based framework, CryptoTrade, achieving higher returns and sharpe ratios across various market scenarios. Moreover, this research highlights the importance of fully leveraging the multi-step reasoning capabilities of advanced LLMs, rather than assuming their superiority will automatically lead to better results.

Limitations

One limitation of the FS-ReasoningAgent framework is its current focus on a limited number of cryptocurrencies, as it has been tested on individual assets. In the future, we plan to expand the framework to handle a diversified portfolio of cryptocurrencies, as well as explore its applicability to traditional financial markets, including stocks in the S&P 500.

Broader Impacts

Our research has several potential broader impacts beyond the scope of cryptocurrency trading. One important consideration is the risk that individuals might try to apply the trading strategies we discuss, leading to possible financial losses. We stress that the strategies presented are intended for academic research and experimental purposes only, and FS-ReasoningAgent is not designed or intended to offer investment advice.

Beyond the financial implications, our work encourages the broader research community to rethink the assumption that more powerful models always deliver better results in all contexts. By demonstrating that stronger LLMs may not outperform simpler models in certain tasks, we emphasize the need for careful model selection based on task-specific requirements. This insight can inspire more nuanced approaches to deploying LLMs in various real-world applications, from finance to other industries where decision-making under uncertainty is crucial. Our findings could also encourage further exploration of LLM-based multi-agent frameworks and adaptive reasoning processes, which have broader relevance in fields such as economics, behavioral science, and automated systems.

References

Gourang Aggarwal, Vimal Patel, Gaurav Varshney, and Kimberly Oostman. 2019. Understanding the social factors affecting the cryptocurrency market. *arXiv preprint arXiv:1901.06245*.

Anamika Anamika and Sowmya Subramaniam. 2022. Do news headlines matter in the cryptocurrency market? *Applied Economics*, 54(54):6322–6338.

Alexandru Costin Baroiu, Vlad Diaconita, and Simona Vasilica Oprea. 2023. Bitcoin volatility in bull vs. bear market-insights from analyzing on-chain metrics and twitter posts. *PeerJ Computer Science*, 9:e1750.

Michele Cagan. 2024. *Stock Market 101: From Bull and Bear Markets to Dividends, Shares, and Margins—Your Essential Guide to the Stock Market*. Simon and Schuster.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Junyan Cheng and Peter Chin. 2024. Sociodojo: Building lifelong analytical agents with real-world text and time series. In *The Twelfth International Conference on Learning Representations*.

Min-Yuh Day, Yirung Cheng, Paoyu Huang, and Yensen Ni. 2023. The profitability of bollinger bands trading bitcoin futures. *Applied Economics Letters*, 30(11):1437–1443.

Barkha Dhingra, Shallu Batra, Vaibhav Aggarwal, Mahender Yadav, and Pankaj Kumar. 2024. Stock market volatility: a systematic review. *Journal of Modelling in Management*, 19(3):925–952.

Han Ding, Yinheng Li, Junhao Wang, and Hang Chen. 2024. Large language model agent in financial trading: A survey. *arXiv preprint arXiv:2408.06361*.

Fotini Economou, Konstantinos Gavriilidis, Bartosz Gebka, and Vasileios Kallinterakis. 2023. Feedback trading: a review of theory and empirical evidence. *Review of Behavioral Finance*, 15(4):429–476.

Ramazan Gencay. 1996. Non-linear prediction of security returns with moving average rules. *Journal of Forecasting*, 15(3):165–174.

Itay Goldstein. 2023. Information in financial markets and its real effects. *Review of Finance*, 27(1):1–32.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, Deyi Xiong, et al. 2023. Evaluating large language models: A comprehensive survey. *arXiv preprint arXiv:2310.19736*.

Hongye Jin, Xiaotian Han, Jingfeng Yang, Zhimeng Jiang, Zirui Liu, Chia-Yuan Chang, Huiyuan Chen, and Xia Hu. 2024. Llm maybe longlm: Self-extend llm context window without tuning. *arXiv preprint arXiv:2401.01325*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. In search of needles in a 10m haystack: Recurrent memory finds what llms miss. *arXiv preprint arXiv:2402.10790*.

Philippe Laban, Alexander R Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. Summary of a haystack: A challenge to long-context llms and rag systems. *arXiv preprint arXiv:2407.01370*.

Kangsan Lee and Daeyoung Jeong. 2023. Too much is too bad: The effect of media coverage on the price volatility of cryptocurrencies. *Journal of International Money and Finance*, 133:102823.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024a. LooGLE: Can long-context language models understand long contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16304–16333, Bangkok, Thailand. Association for Computational Linguistics.

Yuan Li, Bingqiao Luo, Qian Wang, Nuo Chen, Xu Liu, and Bingsheng He. 2024b. A reflective llm-based agent to guide zero-shot cryptocurrency trading. *arXiv preprint arXiv:2407.09546*.

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

Allan H Meltzer. 2002. Rational and irrational bubbles. *Central Banking Studies*.

OpenAI. 2024. https://openai.com/index/openai-o1-mini-advancing-cost-efficient-reasoning/. [Accessed 13-10-2024].

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2023. Refiner: Reasoning feedback on intermediate representations. *arXiv preprint arXiv:2304.01904*.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic prompt optimization with" gradient descent" and beam search. *arXiv preprint arXiv:2305.03495*.

Matthew Renze and Erhan Guven. 2024. Self-reflection in llm agents: Effects on problem-solving performance. *arXiv preprint arXiv:2405.06682*.

Mark Rubinstein. 2001. Rational markets: yes or no? the affirmative case. *Financial Analysts Journal*, 57(3):15–29.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

Jian Wang and Junseok Kim. 2018. Predicting stock price trend using macd optimized by historical volatility. *Mathematical Problems in Engineering*, 2018:1–12.

Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, et al. 2024. Needle in a multimodal haystack. *arXiv preprint arXiv:2406.07230*.

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.

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, 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*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

Seonghyeon Ye, Hyeonbin Hwang, Sohee Yang, Hyeongu Yun, Yireun Kim, and Minjoon Seo. 2024. Investigating the effectiveness of task-agnostic prefix prompt for instruction following. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19386–19394.

Xinli Yu, Zheng Chen, Yuan Ling, Shujing Dong, Zongyi Liu, and Yanbin Lu. 2023. Temporal data meets llm–explainable financial time series forecasting. *arXiv preprint arXiv:2306.11025*.

Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W Suchow, and Khaldoun Khashanah. 2024. 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.

Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, et al. 2024a. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4314–4325.

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*.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022.

Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Appendix

A Experimental Environment

Our experiments were conducted using four NVIDIA H100 PCIe GPUs, managed by the NVIDIA-SMI 555.42.06 driver and leveraging CUDA 12.5 for optimal performance. The models in these experiments were implemented using PyTorch 2.0.0 in Python 3.12.5, ensuring compatibility and efficient execution on this powerful hardware setup.

B Re-implementation of CryptoTrade

B.1 Dataset Splits

We base our re-implementation of CryptoTrade on the dataset CryptoTrade provides which covers several months, detailed in Table 7. This dataset captures the recent market performance of BTC, ETH, and SOL, highlighting challenges in identifying market trends and volatility. We divide the dataset into validation and test sets, using the validation set to fine-tune model hyperparameters and the test set to evaluate model performance.

B.2 Data and Code Source

We utilize the data and code available from CryptoTrade's public GitHub repository: https://github.com/Xtra-Computing/CryptoTrade.

B.3 Experiment Results

The experiment results shown in Table 8 and Table 9 indicate that stronger LLMs, such as o1-mini and GPT-4o, do not consistently outperform either traditional strategies or even simpler LLM models in terms of total returns and risk-adjusted performance. For instance, while GPT-4o performs reasonably well in Bull markets (28.47% total return on BTC and 115.18% on SOL), it fails to deliver the best results, trailing behind the simpler o1-mini model in BTC (36.50%) and behind the traditional SLMA strategy on SOL (169.98%). Furthermore, in Bear markets, o1-mini experiences significant reduction, with a -15.81% return on BTC and -25.68% on SOL, worse than the performance of weaker models like GPT-3.5-turbo. This pattern suggests that stronger LLMs, despite their advanced reasoning capabilities, do not necessarily make better trading decisions under all conditions, particularly in managing risk during downturns. Simpler models, such as GPT-3.5-turbo, and traditional strategies like SLMA, show better resilience and overall balanced performance across different market conditions, highlighting that more advanced LLMs may not always lead to superior results.

C License

The dataset used in this paper is released under the Creative Commons Attribution-NonCommercial-ShareAlike (CC BY-NC-SA) license. This permits use, distribution, and modification of the data for non-commercial purposes, provided that proper attribution is given and any derivative works are shared under the same licensing terms.

D Data Ethics

D.1 Statistical Data

We obtain cryptocurrency statistical data from CoinMarketCap[8] and Dune[9]. In line with CoinMarketCap's Terms of Service[10], we are provided with a limited, personal, non-exclusive, non-sublicensable, and non-transferable license to access and use the content and services solely for personal purposes. We strictly refrain from using the service or its content for any commercial activities, complying fully with these terms. As for Dune's Terms of Service[11], we are allowed to access Dune's APIs to perform SQL queries on blockchain data.

D.2 News

We utilize Gnews[12] to systematically collect cryptocurrency-related news articles. In accordance with Gnews' Terms of Service[13], we are allowed to download news for non-commercial, temporary viewing only. We are prohibited from modifying or copying the content, using it for commercial purposes or public displays, attempting to reverse engineer any software from Gnews, removing any copyright notices, transferring the content to others, or mirroring it on another server. We ensure that these conditions are strictly followed in our dataset.

E Baselines

1. **Buy and Hold:** A straightforward strategy where an asset is purchased at the beginning of the period and held until its end.

2. **SMA (Gencay, 1996):** The Simple Moving Average (SMA) strategy makes buy and sell decisions by comparing the asset's price to its average over a specified period. We experiment with different time windows $[5, 10, 15, 20, 25, 30]$, selecting the period that performs best on a validation dataset.

3. **SLMA (Wang and Kim, 2018):** The Staggered Moving Average (SLMA) method uses two moving averages with distinct durations. Trades are triggered when these averages cross. We evaluate various combinations of short and long moving averages, optimizing them based on validation set outcomes.

[8] https://coinmarketcap.com
[9] https://dune.com/home
[10] https://coinmarketcap.com/terms/
[11] https://dune.com/terms
[12] https://pypi.org/project/gnews/
[13] https://gnews.io/terms/

Type	Split	Start	End	Open	Close	Trend
BTC	Validation	2023-01-19	2023-03-13	20977.48	20628.03	-1.67%
	Test Bearish	2023-04-12	2023-06-16	30462.48	25575.28	-15.61%
	Test Bullish	2023-10-01	2023-12-01	26967.40	37718.01	39.66%
ETH	Validation	2023-01-13	2023-03-12	1417.13	1429.60	0.88%
	Test Bearish	2023-04-12	2023-06-16	1892.94	1664.98	-12.24%
	Test Bullish	2023-10-01	2023-12-01	1671.00	2051.76	22.59%
SOL	Validation	2023-01-14	2023-03-12	18.29	18.24	-0.27%
	Test Bearish	2023-04-12	2023-06-16	23.02	14.76	-36.08%
	Test Bullish	2023-10-01	2023-12-01	21.39	59.25	176.72%

Table 7: Dataset splits. Prices are in US dollars. In each split, the transaction days include the start date and exclude the end date. We evaluate the total profit on the end date.

Strategy	**Total Return**		**Daily Return**		**Sharpe Ratio**	
	Bull	**Bear**	**Bull**	**Bear**	**Bull**	**Bear**
Buy and Hold	39.66	-15.61	$0.56_{\pm 2.23}$	$-0.24_{\pm 2.07}$	0.25	-0.11
SMA	22.58	-21.74	$0.35_{\pm 1.89}$	$-0.36_{\pm 1.25}$	0.18	-0.29
SLMA	38.53	-7.68	$0.55_{\pm 2.21}$	$-0.11_{\pm 1.23}$	0.25	-0.09
MACD	13.57	-9.51	$0.22_{\pm 1.45}$	$-0.14_{\pm 1.56}$	0.15	-0.09
Bollinger Bands	2.97	-1.17	$0.05_{\pm 0.32}$	$-0.02_{\pm 0.51}$	0.15	-0.03
GPT-3.5-turbo	18.84	-9.12	$0.30_{\pm 1.69}$	$-0.14_{\pm 1.52}$	0.18	-0.09
GPT-4	26.35	-11.72	$0.40_{\pm 1.76}$	$-0.18_{\pm 1.67}$	0.23	-0.11
GPT-4o	28.47	-13.71	$0.43_{\pm 1.89}$	$-0.21_{\pm 1.71}$	0.23	-0.12
o1-mini	36.50	-15.81	$0.53_{\pm 2.17}$	$-0.25_{\pm 1.94}$	0.25	-0.13

Table 8: Performance comparison of CryptoTrade and baseline trading strategies on BTC during both Bull and Bear market conditions.

Strategy	**Total Return (%)**		**Daily Return (%)**		**Sharpe Ratio**	
	Bull	**Bear**	**Bull**	**Bear**	**Bull**	**Bear**
Buy and Hold	176.72	-36.08	$1.83_{\pm 6.00}$	$-0.61_{\pm 3.45}$	0.30	-0.18
SMA	119.37	1.04	$1.43_{\pm 5.67}$	$0.02_{\pm 0.10}$	0.25	0.16
SLMA	169.98	-8.11	$1.78_{\pm 5.93}$	$-0.11_{\pm 1.88}$	0.30	-0.06
MACD	23.25	-21.07	$0.35_{\pm 1.76}$	$-0.33_{\pm 2.44}$	0.20	-0.13
Bollinger Bands	2.92	-21.69	$0.05_{\pm 0.35}$	$-0.35_{\pm 1.75}$	0.13	-0.20
GPT-3.5-turbo	102.45	-24.08	$1.26_{\pm 4.54}$	$-0.39_{\pm 2.60}$	0.28	-0.10
GPT-4	99.84	-19.55	$1.24_{\pm 4.53}$	$-0.31_{\pm 2.35}$	0.27	-0.13
GPT-4o	115.18	-16.32	$1.38_{\pm 4.98}$	$-0.25_{\pm 2.35}$	0.28	-0.10
o1-mini	102.67	-25.68	$1.30_{\pm 5.27}$	$-0.41_{\pm 2.85}$	0.25	-0.15

Table 9: Performance comparison of CryptoTrade and baseline trading strategies on SOL during both Bull and Bear market conditions.

4. **MACD (Wang and Kim, 2018):** The Moving Average Convergence Divergence (MACD) strategy identifies buy and sell signals by analyzing momentum shifts. It calculates the difference between a 12-day and a 26-day Exponential Moving Average (EMA), with a 9-day EMA acting as a trigger line. EMAs assign greater significance to recent data points.

5. **Bollinger Bands (Day et al., 2023):** This approach generates signals by observing how the asset's price interacts with the Bollinger Bands, which consist of a 20-day SMA and bands placed at a set distance (typically two standard deviations) above and below. We adopt the standard settings for period length and band multiplier.

6. **CryptoTrade (Li et al., 2024b):** This strategy is an LLM-based trading agent designed specifically for cryptocurrency markets, expanding the typical application of LLMs beyond stock market trading. Experiments show that CryptoTrade outperforms time-series baselines in maximizing returns, though traditional trading signals still perform better under most of conditions.

F Author Statement

As authors of the this paper, we hereby declare that we assume full responsibility for any liability or infringement of third-party rights that may come up from the use of our data. We confirm that we have obtained all necessary permissions and/or licenses needed to share this data with others for their own use. In doing so, we agree to indemnify and hold harmless any person or entity that may suffer damages resulting from our actions.

Furthermore, we confirm that our dataset is released under the Creative Commons Attribution-NonCommercial-ShareAlike (CC BY-NC-SA) license. This license allows anyone to use, distribute, and modify our data for non-commercial purposes as long as they give proper attribution and share the derivative works under the same license terms. We believe that this licensing model aligns with our goal of promoting open access to high-quality data while respecting the intellectual property rights of all parties involved.

G Hosting Plan

After careful consideration, we have chosen to host our code and data on GitHub. Our decision is based on various factors, including the platform's ease of use, cost-effectiveness, and scalability. We understand that accessibility is key when it comes to data management, which is why we will ensure that our data is easily accessible through a curated interface. We also recognize the importance of maintaining the platform's stability and functionality, and as such, we will provide the necessary maintenance to ensure that it remains up-to-date, bug-free, and running smoothly.

At the heart of our project is the belief in open access to data, and we are committed to making our data available to those who need it. As part of this commitment,

we will be updating our GitHub repository regularly, so that users can rely on timely access to the most current information. We hope that by using GitHub as our hosting platform, we can provide a user-friendly and reliable solution for sharing our data with others.