FINAL PROJECT: QUANT-GPT

Yuzhe Yang 121090684 **Kangqi Yu 121090735** **Junquan Peng 120090556**

Chinese University of Hong Kong, Shenzhen
Shenzhen, China
{yuzheyang, kangqiyu, junquanpeng}@link.cuhk.edu.cn

ABSTRACT

This paper introduces Quant-GPT, a novel multi-agent optimized for A-share market investment decisions. Leveraging a fine-tuning combination of distilled sentiment analysis from ChatGPT and real-world market data, the prediction agent of Quant-GPT addresses the challenges of model collapse and weak causality between sentiment and expected returns. Our methodology integrates a Retrieval-Augmented Generation (RAG) agent and summary agent, enabling the model to access relevant news articles and corporate announcements summary with concise investment information to enhance investment decision-making. The inclusion of diverse datasets and RAG significantly improves the model's ability to forecast market trends and returns accurately. Experimental results demonstrate Quant-GPT's superior performance over existing open-source LLMs in terms of annualized return, maximum draw-down, and Sharpe ratio. These findings underscore the potential of advanced language models in financial applications, providing a robust framework for integrating natural language understanding with quantitative investment strategies. The code is available on GitHub: `https://github.com/TobyYang7/Quant-GPT`

1 INTRODUCTION

1.1 RESEARCH TOPIC

The objective of our research is the application of LLMs to quantitative finance, with a particular focus on enhancing investment strategies within China's A-share market. Recent advancements in LLMs, such as those seen in ChatGPT and its derivatives, have markedly improved natural language processing capabilities. These models, which are trained on extensive corpuses encompassing a broad spectrum of human knowledge, excel in tasks requiring deep contextual comprehension and nuanced language generation.

In financial analytics, the traditional reliance on quantitative data is being reevaluated with the potential integration of LLMs, which offer a systematic approach to assimilating unstructured textual data. Financial markets are influenced significantly by textual information sources, including news articles, financial statements, and regulatory filings. LLMs can process this voluminous textual data to extract sentiment and thematic trends that might elude conventional quantitative analysis.

However, the integration of LLMs into quantitative finance is fraught with challenges, notably the risk of model over-fitting and the phenomenon known as model collapse. To address these issues, our research introduces Quant-GPT, a specialized LLMs framework for mid-low frequency financial quantitative trading. This framework empowers the model with enhanced analytical capabilities by dynamically accessing and synthesizing relevant financial news and data, thereby enriching the model's predictive precision and robustness.

The integration of Quant-GPT into quantitative finance is pivotal due to the A-share market's complexity, where decisions are influenced by rapidly changing economic and social sentiments. Traditional quantitative models are often inadequate for processing unstructured textual data effectively,

such as news articles and financial reports, which contain critical sentiment and thematic trends. Quant-GPT's unique approach utilizes both distilled sentiment analysis and real-world market data, addressing common issues like model collapse and the tenuous link between sentiment and market returns. This not only enhances the model's predictive accuracy but also offers investors a robust tool for navigating the volatility of the A-share market, ultimately leading to improved investment outcomes.

1.2 ACHIEVEMENT

Our approach involved a meticulous process of fine-tuning a base LLM with a multi-source data framework, which integrates distilled sentiment information from existing LLM outputs and real-world financial announcements. This dual integration aimed to overcome the limitations of existing sentiment-only models.

Notably, most current related work primarily focuses on predicting stock trends (1) rather than explicitly defining five distinct return levels. Our method addresses this gap by specifying and predicting these five categories, providing a more comprehensive strategy framework.

The integration of a RAG system allowed for dynamic data retrieval, significantly reducing the lag between data acquisition and decision-making, thus enabling real-time analysis and response to market changes.

By incorporating real-world financial data into the training process, Quant-GPT demonstrated superior predictive capabilities compared to traditional models, reflected in key performance metrics such as annualized return, maximum drawdown, and the Sharpe ratio.

2 RELATED WORK

The intersection of LLMs and finance has garnered substantial attention, leading to significant developments in financial language models. Notable contributions in this arena include BloombergGPT (2), which leverages specialized datasets to enhance performance in financial forecasting and risk assessment. These models demonstrate the potential of LLMs when fine-tuned with domain-specific data, although their accessibility is often limited by proprietary data and high operational costs and restricted to the US stock market.

Our work is inspired by the DISC-FinLLM (3), a Chinese financial LLM that integrates multiple expert fine-tuning to address specific financial tasks, such as question answering and retrieval-augmented generation. This model exemplifies how diverse financial data sources, from regulatory filings to social media discussions, can be effectively utilized to train LLMs that are both comprehensive and capable of real-time financial analysis.

Furthermore, the introduction of open-source models FinGPT (4) marks a pivotal shift towards democratizing financial data and LLM technologies. These models emphasize a data-centric approach and leverage community-driven developments to foster innovation and practical applications in finance.

3 TASK

3.1 METHODOLOGY

3.1.1 FRAMEWORK

Our system is shown in Figure 1 (the screenshot of our demo system is shown in Appendix A), which employs a multi-agent framework to assist users in making informed investment decisions. Users can input multiple announcements regarding their concerns about a particular stock. The summary agent (ChatGLM3-6B-128k (5)) condenses the key points from these announcements. The retrieval agent then gathers relevant information from a knowledge base, providing additional context. This information is passed to the planning agent, which makes an initial assessment of the industry market or stock trend. Finally, the decision agent (DeepSeek-V2 (6)) uses this information to formulate detailed investment strategies.

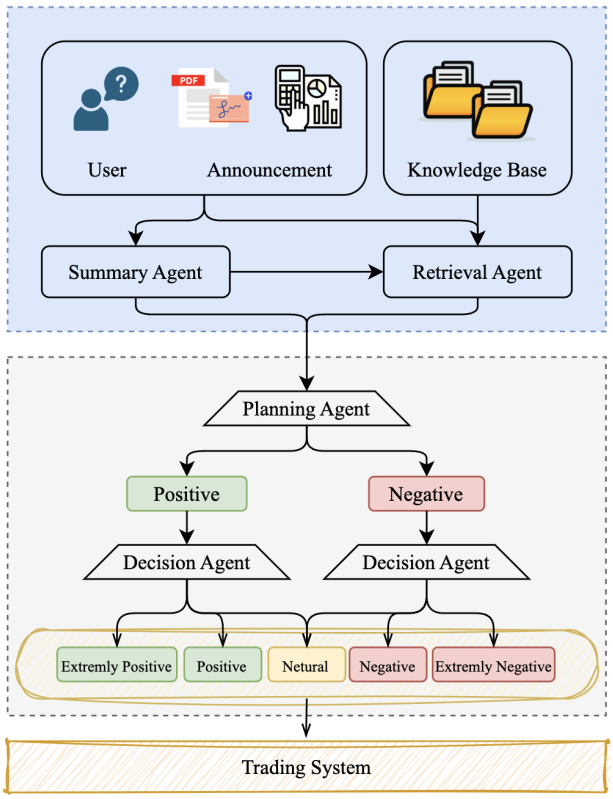

Figure 1: Framework of Quant-GPT

3.1.2 Financial News RAG

The RAG system is an agent that summarizes company announcements using a summary agent ChatGLM3-6B-128k. It is activated when a user inputs an announcement, focusing on key aspects such as the announcement time, company introduction, and relevant quantitative values by prompt engineering.

The system uses the BGE-M3 embedding model (7), which unifies dense retrieval and lexical retrieval functionalities. Dense embeddings capture the overall semantics of the text by converting it into vectors. The input announcement summary query q and news message p are transformed into the hidden states H_q and H_p based on a text encoder.

$$e_q = \text{norm}(H_q[0])$$

$$e_p = \text{norm}(H_p[0])$$

Thus, the relevance score between query and passage is measured by the inner product between the two embeddings e_q and e_p:

$$s_{\text{dense}} \leftarrow \langle e_p, e_q \rangle$$

The embeddings are also used to estimate the importance of each term to facilitate lexical retrieval. For each term t within the announcement summary (a term is corresponding to a token), the term weight is computed as

$$w_{qt} \leftarrow \text{ReLU}(W_{\text{lex}}^T H_q[i])$$

where $W_{\text{lex}} \in \mathbb{R}^{d \times 1}$ is the matrix mapping the hidden state to a float number. The relevance score between query and passage is computed by the joint importance of the co-existed terms (denoted as $q \cap p$):

$$s_{\text{lex}} \leftarrow \sum_{t \in q \cap p} (w_{qt} \cdot w_{pt})$$

Utilizing this hybrid similarity measure combining s_{dense} and s_{lex}, the system retrieves the top 10 news articles within a user-defined time window. Users have the flexibility to adjust the time window duration and specify the types of news in the user interface, allowing the model to conduct precise sentiment analysis of announcements.

Upon retrieval of the news articles, Quant-GPT is employed to conduct investment relevance scoring on each news by prompt, aiming to uncover deeper financial connections between company announcements and news. This approach facilitates the reranking of the top-10 news, with the final top-5 articles selected as the most pertinent texts required for the final prediction.

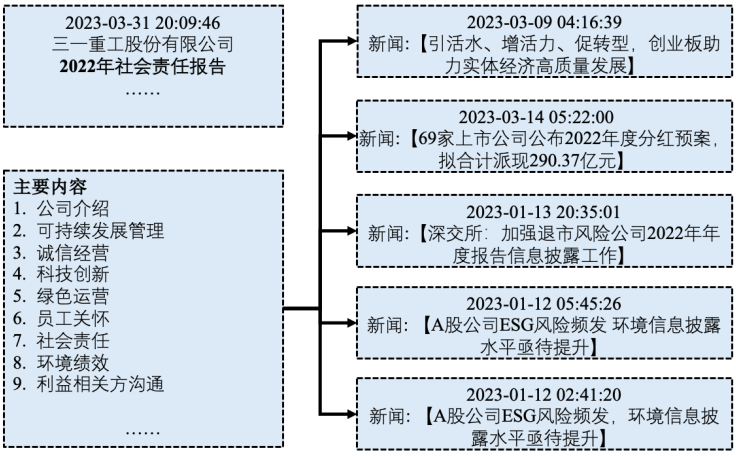

Figure 2: RAG Example

4 EXPERIMENT SETTING

4.1 DATASET

4.1.1 RAW DATASET

The time spread of all the datasets is from 2018-10-08 to 2024-04-14. The last year, 2023-04-14 to 2024-04-14, is treated as a test period, and the previous days are treated as a training period.

Listed Companies Announcement We select ten listed companies in Appendix B from different industries, and collect their announcement from Tushare [1]. This dataset covers changes in the shares capital, rules and regulations, shareholder changes, etc.

Sina News We collect all the Sina news through Tushare. Then, select the news related to the A-share markets containing approximately 10 thousand items.

History Trading Data To ensure the quality of history trading data, we select the data provided by MYQUANT [2]. The price adjustment method is post-adjust, which helps us to make the price sequence reflect real returns more accurately. The trading-based price is open, corresponding to our executive idea to change position at 09:30 a.m.

4.1.2 PROCESSED DATASET

Sentiment Analysis Sampled one thousand Sina News, we use GPT-4 to label their sentiment, which helps us to distill the sentiment analysis power of GPT-4.

Stock Return Level Prediction We will predict low-mid frequency return and uniformly cut continuous log return into 5 levels based on the training data distribution. For the announcements we collected, just like shown in the section framework 3.1.1, we use the summary agent to summarize

[1] https://tushare.pro/
[2] https://www.myquant.cn/

the upcoming announcement. Then, the retrieval agent will retrieve the news related to this company and the macroeconomic background. Finally, look back several days and concatenate a sequence of announcements to provide more background (8).

4.2 DESIGN

We build instruction tuning samples with the following template:

Instruction: [Task Prompts]	Text: [Financial News (RAG), Announcement (Summary)]
Response: [Corresponding Labels]	

We extract the keywords from the summarized announcement and then query for the relevant financial news from our knowledge base through RAG. This ensures that we can simultaneously learn about macroeconomic market factors and specific changes of the industry.

5 RESULT ANALYSIS

5.1 QUANTITATIVE EVALUATIONS

To evaluate the power of Quant-GPT, we use realistic trading data to simulate the investments from 2023/04/14 to 2024/04/14. On the 09:00 a.m. of each trading day, we receive the instructions of gpt-3.5-turbo, Qwen-7B, Quant-GPT without Decision Agent Instructions, and Quant-GPT with Decision Agent Instructions. We can get a target portfolio weight for each stock by synthesizing all the instructions with a time decay rate. Then, we change positions at 09:30 a.m. Their PnLs (Profit and Loss) and performance metrics are shown below (The PnLs of the first three models are in the Appendix C).

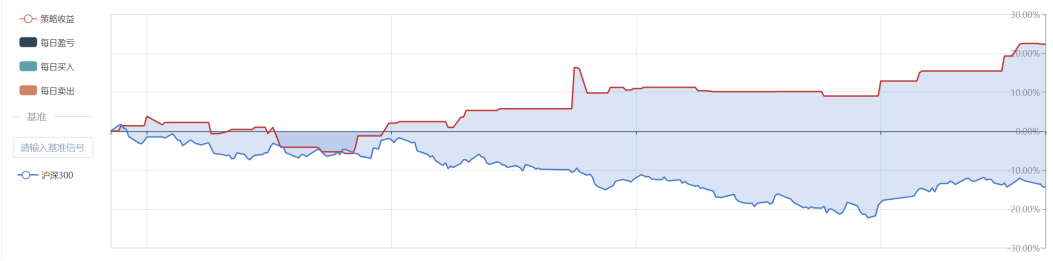

Figure 3: Trade with Quant-GPT with Decision Agent Instructions

Metrics	gpt-3.5-turbo	Qwen-7B	Quant-GPT without Decision Agent	Quant-GPT with Decision Agent
Annualized Return	-7.81%	-11.02%	4.83%	7.26%
Max Drawdown	19.96%	11.26%	12.39%	13.61%
Sharpe Ratio	-0.49	-2.02	0.31	0.40

5.2 CASE STUDY

In Appendix D, we illustrate a demo feedback of Quant-GPT. Our objective is to predict 5 labels for our log return level. However, if we directly use LLM to predict these 5 labels, we find that the accuracy is not particularly high and the results lack interpretability. Additionally, general LLM have limitations in understanding financial texts.

Therefore, our solution is first to use a multi-agent system to organize the key elements of the text. Then, we employ a fine-tuned model to predict the market trend roughly. Finally, we use a general-purpose large model to make specific decisions. This approach not only ensures that our decisions

align with conventional logic but also enhances the interpretability of our results. In real trading scenarios, this method can provide valuable insights for human traders.

6 CONCLUSION

Although we have proposed a brand new multi-agents framework and get relatively superior performance to GPT-3.5 and Qwen, which fills the blank of LLMs in Quantitative investment, this framework still has some limitations:

1. The data source is too signal (Sina News), which may make our investment decisions biased.
2. The financial market is dynamic, so the historical lessons learned by it have the risk of losing efficacy.
3. The financial market is noisy. It can not predict the return level exactly, so we need a trading strategy to synthesize all the instructions we get.
4. The computation resources limit us to train a larger model, which will perform better in a long context.

In the future, we may concern about:

1. Add more data from multiple resources by crawling from the Internet, buying processed data, and adding human text like the report from sell-side analysts.
2. Add special tokens to mark the temporal stock data and align it with the log return level prediction.
3. Replace the research target from stocks to industry indices will lower the noise in the corresponding return.

DIVISION OF TASKS

If it is a team task, please be sure to clearly write down the division of labor, which is **very important** for grading and assessment.

- Yuzhe Yang: Methodology, Fine-tuning
- Kangqi Yu: Methodology, Data Engineering, Trading Backtest
- Junquan Peng: Methodology, RAG

ACKNOWLEDGMENT

This is the final project for DDA 6307 / CSC 6052 / MDS6002, see details in `https://nlp-course-cuhksz.github.io/`.

REFERENCES

[1] Kelvin J.L. Koa, Yunshan Ma, Ritchie Ng, and Tat-Seng Chua. Learning to generate explainable stock predictions using self-reflective large language models. In *Proceedings of the ACM on Web Conference 2024*, WWW ' 24. ACM, May 2024.

[2] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. Bloomberggpt: A large language model for finance, 2023.

[3] Wei Chen, Qiushi Wang, Zefei Long, Xianyin Zhang, Zhongtian Lu, Bingxuan Li, Siyuan Wang, Jiarong Xu, Xiang Bai, Xuanjing Huang, and Zhongyu Wei. Disc-finllm: A chinese financial large language model based on multiple experts fine-tuning. *arXiv preprint arXiv:2310.15205*, 2023.

[4] Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. Fingpt: Open-source financial large language models. *FinLLM Symposium at IJCAI 2023*, 2023.

[5] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

[6] DeepSeek-AI. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model, 2024.

[7] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

[8] Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction, 2019.

A DEMO UI DESIGN

Figure 4: Demo UI Design with Gradio Framework

B SELECTED COMPANIES LIST

Here is the list of companies we selected:

- 600031.SH 三一重工
- 600036.SH 招商银行
- 600050.SH 中国联通
- 600104.SH 上汽集团
- 600346.SH 恒力石化
- 600570.SH 恒生电子

- 600887.SH 伊利股份
- 601390.SH 中国中铁
- 603160.SH 汇项科技
- 601668.SH 中国建筑

C Other PnLs

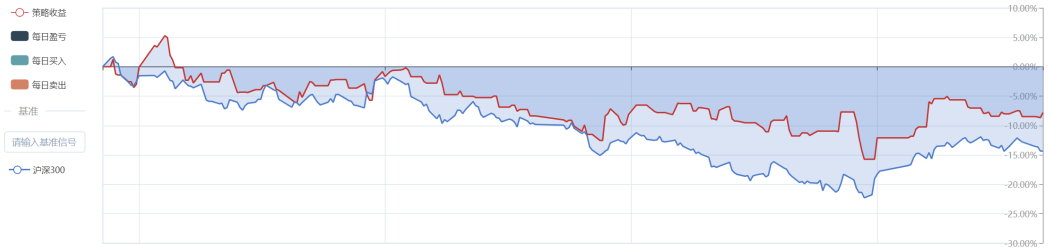

Figure 5: Trade with Zero-Shot Instructions (gpt3.5-turbo)

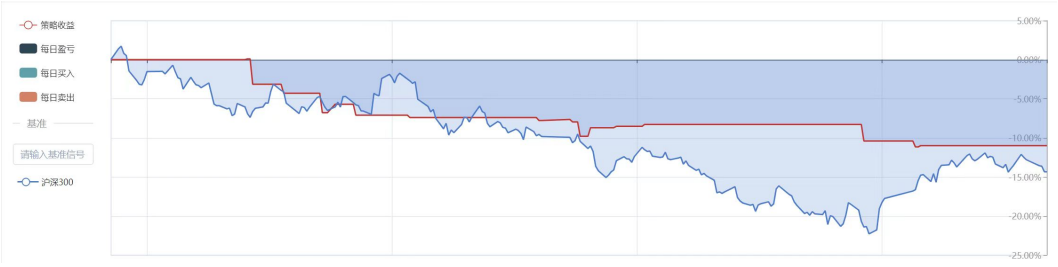

Figure 6: Trade with Zero-Shot Instructions (Qwen-7B)

Figure 7: Trade with Quant-GPT without Decision Agent Instructions

D Example Chat Demo

An example prompt for AI feedback

Prompt: 请根据以下新闻文本，预测三一重工股票的对数收益率属于以下哪一类别 (极度负面 / 负面 / 中性 / 正面 / 极度正面)

Text Input:

现在是2023-06-15 09:00:00+08:00前五个交易日上证指数的价格如下 3197.468,3194.835,3213.033,3225.305,3223.8962023-06-11 18:34:01 新闻:【今年来最大解禁潮来袭 下周解禁市值3874.83亿元】 Wind数据显示，除将上市的新股外，下周将有77只股票面临解禁，合计解禁量为611.19亿股，按最新收盘价计算，合计解禁市值为3874.83亿元，为2023年来单周解禁市值最高的一周，较本周解禁市值环比增加237.36%。从解禁股情况来看，邮储银行将解禁金额超3000亿元，占下周解禁市值总额的近8成，天合光能将解禁超310亿元，三安光电将解禁超90亿元。邮储银行、凯迪股份流通盘将增加超200%。（中国证券报）2023-05-17 07:33:21 新闻:【新能源汽车下乡催热光储充一体化产业链】 据上证报，在日前召开的中国汽车工业协会沟通会上，中汽协副总工程师许海东介绍，对于2023年新能源汽车下乡细则，相关部门正在准备过程中。记者近日采访获悉，随着新能源车的持续推广，打通光电利用通道的光储充产业链发展迅速，不仅国际巨头特斯拉积极探索布局，国内多家新能源产业链上市公司也于近期加码光储充一体化新业务。乘联会专家呼吁，应抓住新能源汽车下乡的窗口期，加快充电桩等农村地区光储充设施建设。2023-05-16 06:44:43 新闻:【业内人士：新能源板块整体估值处于近五年的相对低位 依然存在大量投资机会】新能源板块5月15日迎来“久违”的全线走强，光伏、储能、锂电池、新能源车等细分领域纷纷“反攻”。该板块经过近一年的调整，反复下探。近来，碳酸锂价格实现“十一连涨”，反弹幅度累计超40%，被市场视为新能源赛道触底回升的重要信号。业内人士表示，新能源板块整体估值处于近五年的相对低位，碳酸锂价格持续反弹主要受下游新能源企业需求超预期、锂矿供给收缩等因素影响，拉长周期来看，新能源板块依然存在大量投资机会。 (中证报)2023-05-20 21:52:09 新闻:【天风研究：“一带一路”新一轮行情或有更大空间和更好持续性】天风证券20日发布研报指出，“一带一路”新一轮行情或有更大空间和更好持续性，人民币国际化进程及疫后出海企业订单蕴含更大弹性或为此轮行情更突出的逻辑，同时应重视外交事件催化作用。建议沿三个方向寻找投资标的：1、在海外市场深耕的国际工程公司，推荐中材国际、中油工程、北方国际，建议关注中钢国际、上海港湾、中工国际；2、建筑央企亦为基建出海主力军，推荐中国化学、中国铁建、中国建筑、中国交建、中国中铁、中国中冶、中国电建、中国能建等；3、受益“一带一路”推进，新疆地区基建或景气向上，建议关注新疆交建、北新路桥、雪峰科技、青松建化等。2023-05-30 22:02:11 新闻:【瑞银：“中特估”叠加“一带一路”是今年比较重要的投资主线】“‘中特估’叠加‘一带一路’是今年比较重要的投资主线。”在5月20日举行的《中国“一带一路”倡议十周年解读及基建行业展望》电话会上，瑞银证券亚洲工业行业主管、中国研究部副总监徐宾指出，基建板块有望受益。（澎湃）2023-05-17 11:04:22 新闻:【军工股持续走高 长城军工涨停】长城军工涨停，中天火箭、中船防务此前涨停，新余国科、捷强装备、中船应急、天秦装备、中无人机、航发科技等涨超5%。银河证券发布研究报告表示，军工行业财报季靴子落地，板块预期将迎边际改善。2023-05-17 14:00:58 新闻:军工板块盘中走势活跃，新余国科涨近18%，盘中一度触及涨停；光力科技、中船汉光涨超10%，汉马科技、中天火箭、长城军工、中船防务等涨停，晨曦航空、天海防务、淳中科技等涨超7%。2023-05-19 10:19:39 新闻:A股军工板块震荡拉升，观典防务涨超9%，超卓航科涨逾6%，新兴装备、中无人机、航宇科技、迈信林等跟涨。2023-06-06 13:07:45 新闻:A股军工板块震荡走低，航天动力跌逾8%，华如科技跌超7%，航天软件、淳中科技、宝塔实业、国盾量子等跌超4%。时间：2023-06-14 21:58:04标题：三一重工股份有限公司关于召开2023年第三次临时股东大会的通知三一重工股份有限公司将于2023年6月30日召开2023年第三次临时股东大会，会议采用现场投票和网络投票相结合的方式。会议审议三项议案，包括《2023年员工持股计划（草案）及其摘要》的议案、《2023年员工持股计划管理办法》的议案和关于提请股东大会授权董事会办理公司2023年员工持股计划相关事宜的议案。会议登记时间为2023年6月23日上午9:00-12:00，下午14:30-17:30，地点为湖南省长沙市长沙县经开区三一产业园新研发楼4楼。

Expected Label: 正面
Log Return: 0.017732195963635

Predicted (By Planning Agent): 正面
Response (By Decision Agent):

首先，我们需要分析新闻文本对三一重工股份有限公司的影响。从提供的新闻中，我们可以看到以下几点：1. 新能源汽车下乡催热光储充一体化产业链，这可能对三一重工的业务产生积极影响，因为三一重工可能参与到相关的基础设施建设中。2. 新能源板块整体估值处于近五年的相对低位，存在投资机会。虽然这不一定直接关联到三一重工，但整个行业的复苏可能会带动相关企业的股价上涨。3. “一带一路”新一轮行情或有更大空间和更好持续性，这可能对三一重工等基建企业带来海外订单的增长。4. 军工股持续走高，虽然三一重工不是纯粹的军工股，但其业务范围广泛，可能受益于军工行业的增长。5. 三一重工即将召开股东大会审议员工持股计划等议案，这通常被视为公司管理层对未来发展的信心表现。结合这些信息，我们可以推断短期内（5天）市场对三一重工的情绪可能是正面的。因此，对于短期内的对数收益率属于正面这一判断是正确的。基于以上分析，我的投资建议是：B. 正面。这意味着我认为在当前的市场环境和新闻背景下，投资三一重工是有利的。然而，投资者应继续关注公司的具体动态和市场变化，以做出更为明智的投资决策

Result: 正面

Figure 8: Example