FinAgent: Benchmarking Financial Analysis with Multimodal Stock-Agent

Anonymous ACL submission

Abstract

Constructing stock agents that facilitate investment analysis is an important research direction in finance. The key technology is the agent's capability to automatically identify user queries and integrate multimodal data for analysis by large language models (LLMs). Currently, LLMs have made some progress, primarily in retrieving text and time-series data from knowledge bases based on user queries and providing these data in a basic combination to LLMs. However, they have not efficiently integrated these data to enhance the performance of the LLMs. Also, they do not fully exploit image information and depend on extensive knowledge bases that require real-time updates. To overcome these limitations, we propose the FinAgent dataset, which encompasses research datasets, financial Q&A, stock charts, and handwritten chain-of-thought (CoT) data. Moreover, we innovate a Stock-Agent that efficiently discerns user intent and retrieves necessary information via APIs and knowledge bases to tackle financial tasks. Additionally, we propose an efficient multimodal information fusion method that enhances data sorting and organization, thereby improving the analytical quality of LLMs. We conduct extensive experiments to demonstrate the effectiveness of our framework in financial analysis.

1 Introduction

With the advancement of generative agents (Park et al., 2023), building an intelligent agent that can tackle financial analysis tasks is an important direction. Financial analysis tasks involve stock trend prediction and financial Q&A (Saad et al., 1998; Shah et al., 2022). Due to the outstanding performance of LLMs in content generation (Dredze et al., 2016; Araci, 2019; Bao et al., 2021), they have attracted widespread attention within the financial sector. Thus, there's a keen interest in using LLMs as agents to improve financial analysis.

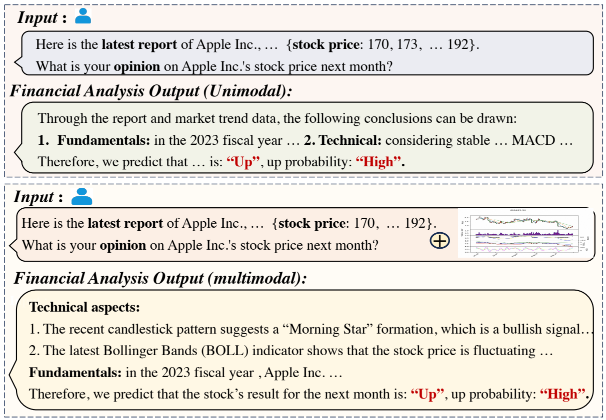

Figure 1: Traditional LLMs rely solely on textual information for stock price prediction, lacking the analysis of image data. An ideal LLMs should combine multimodal information for a more comprehensive analysis.

Leveraging the prowess of LLMs, FinGPT(Yang et al., 2023), FinMA(Xie et al., 2023) and BloombergGPT(Wu et al., 2023) are tailored as specialized FinLLMs. They can be applied to a variety of financial tasks, aligning with industry demands. As shown in Figure 1, they can handle diverse textual data, including news and reports. Thus, FinLLMs assist investors in making investment and trading decisions.

However, due to the limitations of FinLLMs and lack of stock image's training data, they have shortcomings in processing visual data. They support only a single modality, relying on textual data and weakening their ability in financial analysis.

Moreover, the current mechanism of FinLLMs involves matching user queries with retrieval-augmented generation (RAG) (Lewis et al., 2021), retrieving relevant text and data to feed into FinLLMs. This process necessitates daily updates to the knowledge base, which requires a vast capacity. Also, FinLLMs merely merge the knowledge together without delving deeper into knowledge and effectively integrating them. Thus, this approach fails to enhance the capabilities of the FinLLMs.

To effectively tackle these challenges, we first formalize the tasks of financial analysis. Additionally, we propose the FinAgent dataset for fine-tuning LLMs. This dataset encompasses not only financial report and stock Q&A, but also an extensive collection of paired stock candlestick charts and their corresponding textual annotations, along with manually crafted image analysis corpora. Building on this, we introduce an integrated stock investment analysis agent, termed Stock-Agent. It not only recognizes investors' queries and employs precise retrieval tools (APIs and knowledge bases) based on the queries, but also deeply mines and applies effective integration for the retrieved information. The approach delivers more accurate stock trend predictions and nuanced financial analysis for investors.[1]

The main contributions of the model are shown:

- We propose an FinAgent dataset, which contains research data, financial reports, handwritten CoT data, and stock charts, driving the development of multimodal LLMs.
- We propose the Stock-Agent framework, which integrates a fine-tuned StockGPT based on FinAgent with automated agents. It adeptly identifies investor queries and utilizes precise retrieval tools suited to the queries, enhancing retrieval precision and efficiency.
- We propose a multimodal information fusion method, which deeply mines and effectively integrates the retrieved information, offering stock predictions and financial analysis.
- Our extensive experiments reveal that Stock-Agent outperforms the existing methods on FinAgent datasets. The ablation studies show the efficiency of each module in Stock-Agent.

2 FinAgent Datasets

We present the FinAgent dataset as shown in Figure 2, comprising research datasets, StockQA, financial news, reports, and 50,000 stock charts, sourced from various databases. Table 1 highlights the issue of inadequate label length in existing research datasets, which impedes FinLLMs training. FinAgent remedies these issues by providing enhanced quality and length. We provide the details of FinAgent's data sources and construction process in this section.

[1]The proposed datasets, source codes, and model checkpoints will be released upon paper acceptance.

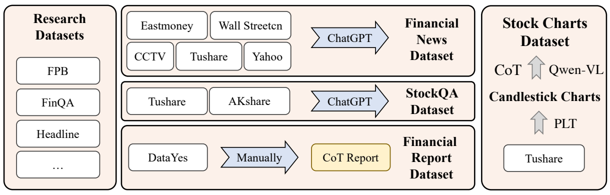

Figure 2: The data source and preprocessing of the proposed FinAgent datasets.

2.1 Data Sources

- **Research datasets**: This part encompasses a range of traditional financial datasets drawn from scholarly sources, such as FPB (Malo et al., 2014), FinQA (Maia et al., 2018), convFinQA (Chen et al., 2022), and Headline (Sinha and Khandait, 2020). These datasets are instrumental in augmenting the capabilities of LLMs in terms of extracting information and generating summaries.
- **StockQA dataset**: This section contains financial data including stock price obtained from Tushare (Tushare, 2021) and AKshare (AKshare, 2021). It adopts a sequential data format to capture stock price movements (e.g. $\{..., 170, 173, 171, ...\}$).
- **Financial news dataset**: To furnish FinLLMs with financial insights, we integrate financial news, including the economic segments of CCTV and Wall Street CN.
- **Financial reports dataset**: We construct datasets of financial report through DataYes (DataYes, 2021), encompassing expert evaluations and corporate insights performed by institutional analysts.
- **Stock charts**: We create stock candlestick charts using time-series data for all listed companies from the last 100 trading weeks, sourced from TuShare (Tushare, 2021).

Dataset	Size	Input	Label	Type
Research	42,373	712.8	5.6	en
StockQA	21,000	1313.6	40.8	zh
Fin. News	79,000	497.8	64.2	zh
Fin. Reports	120,000	2203.0	17.2	zh
Fin. Reports CoT	200	2184.8	407.8	zh
Stock Charts	80000	180.3	411.08	zh

Table 1: The details of the FinAgent datasets. "Input" and "Label" denote their text length.

2.2 Data Preprocessing

As shown in Figure 2, we explore the details of FinAgent preprocessing:

- **Research dataset**: Conventional research datasets predominantly consist of extensive volumes of English language material. To improve FinLLM's proficiency in Chinese, we selectively extract a subset from them.
- **StockQA dataset**: Given that the source data is presented sequentially, we deploy ChatGPT with a tailored prompt to create financial questions. This method aids in training and expands the pool of questions. We then use ChatGPT to formulate answers, creating QA pairs for LLMs training.
- **Financial news dataset**: Using ChatGPT, we summarize news to create a financial news dataset. This method enhances FinLLM's skills in financial news summaries.
- **Financial eports dataset**: We align the financial reports of companies with their respective stock prices on the day the reports are made public, using a predefined template to construct financial report datasets. Furthermore, to enhance the capabilities of LLMs, we manually create 200 CoT financial report data using expert financial insight and more descriptive labels.
- **Stock charts dataset**: We utilize Tushare to collect historical stock data and convert it into candlestick charts, integrating financial indicators such as moving averages to reflect realistic trading scenarios. Then, we construct image-text pairs for stock chart analysis using Qwen-VL. Moreover, we manually create 100 CoT pairs to enhance the stock chart analysis capabilities of FinLLMs.

3 Stock-Agent Framework

We define the tasks of financial analysis as comprising two components: stock trend prediction and the financial (Q&A). Accordingly, we develop the Stock-Agent framework, which tackles two financial analysis tasks, as depicted in Figure 3. Within this section, we begin by formally defining the scope of the tasks and provide an in-depth exposition of each component within our framework.

3.1 Task Definition

For the task 1, given a set of companies $C = \{c_i\}_{i=1}^{N}$ along with the associated knowledge documents $D = \{d_j\}_{j=1}^{M}$, we can predict the stock trends. These documents encompass three kinds of information, including textual reports, stock price time series data, and stock candlestick charts, etc.

$$Pred_i = \phi(c_i, d_j),\ \ Pred_i \in \{up, down\} \quad (1)$$

where ϕ denotes a stock predicting system, and d_j is identified as the documentation pertinent to company c_i. The objective is to select a subset of companies C_{chosen} that are forecast to rise.

$$C_{chosen} = \{c_i | c_i \in C \wedge Pred_i = up\} \quad (2)$$

For the task 2, automated agents adeptly identify investor queries and utilize API and retrieval tools to be suited for the queries Q. We denote Q_t and R_t as the user query and agent response at the current time step t, and $H_t = [Q_0, R_0, ..., Q_{t-1}, R_{t-1}]$ as the dialogue history.

$$r_k = Agent(Q) \quad (3)$$

$$d_k = \begin{cases} API(r_k), & r_k = 0 \\ DB(r_k), & r_k = 1 \end{cases} \quad (4)$$

where r_k represents the intent of user query returned by the agent. Meanwhile, $r_k = 0$ stands for using api, and $r_k = 1$ is for using database retrieval.

Then we formalize the financial Q&A task as obtaining the response based on a current query, dialogue history, and corresponding documents:

$$R_t = \pi(d_k, H_t, Q_t) \quad (5)$$

where π represents the framework, and d_k is the retrieved document related to Q_t.

3.2 Task 1: Stock Trend Prediction

Our first task is to focus on predicting stock trends for each specific company c_i based on its correlated documents d_j, which comprise both textual and image information.

3.2.1 StockGPT Fine-Tuning

Figure 3 presents a three-step LoRA-based fine-tuning of GPT: training on FinAgent's financial reports, enhancing reasoning with CoT reports and adding analysis skills using annotated stock charts. Through fine-tuning, we obtain

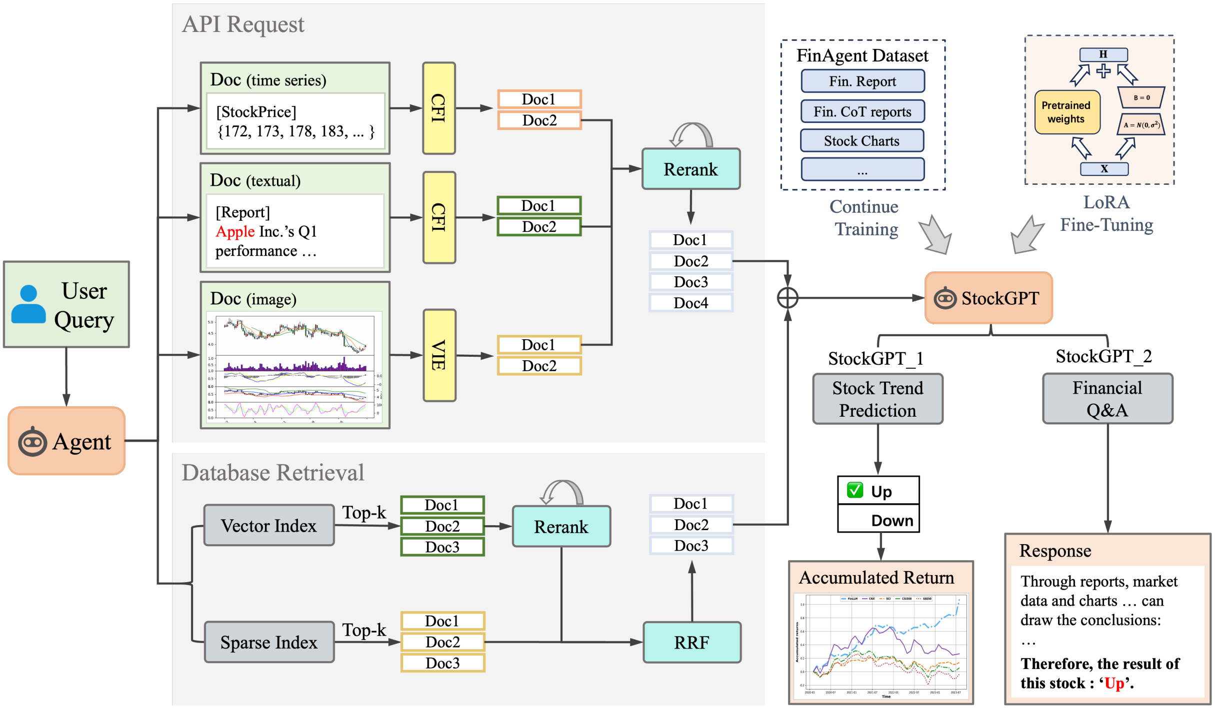

Figure 3: An illustration of the Stock-Agent framework of the two tasks in financial analysis.

$StockGPT_1$, which can predict the trend of c_i based on d_j more accurately, as well as providing detailed analysis and explanations.

3.2.2 CFI and VIE

We propose Comprehensive Financial Insight (CFI) and Visual Investment Explorer (VIE) technologies to pose in-depth questions about the retrieved documents d_j, as shown in Appendix C. In Appendix A, we show performance evaluation of different open-source multimodal LLMs from professional financial analysts. Thus, for textual documents, we employ the ChatGLM3-6B (Du et al., 2021), while for image documents, we utilize the Qwen-VL (Bai et al., 2023). These questions, along with the documents d_j, are fed into a LLM to generate new d_j^1, to enhance document comprehension and analysis.

$$d_j^1 = CFI\&VIE(d_j) \tag{6}$$

To enhance the FinLLM's utilization efficiency of multimodal information, including research reports, stock data, and chart's textual analysis, we employ an adaptive method $Rerank$ (Xiao et al., 2023) to integrate and optimize the sequence of information input. This method mitigates the oversight of middle information by LLMs, improving the accuracy of stock trend predictions. Then, we design a template $Prompt_1$.

$$d_{api} = Rerank(d_j^1) \tag{7}$$

$$I_i = concat(Prompt_1, d_{api}) \tag{8}$$

where `<report>`, `<market data>` and `<image>` compose d_{api}. We concatenate the prompt with the documents as the input I_i into the StockGPT to analyze the stock trend prediction $Pred_i$ for c_i.

$$Res_i = StockGPT_1(I_i) \tag{9}$$

3.2.3 Prediction and Post-Process

We manually extract the prediction result $Pred_i$ from Res_i. Finally, we choose all stocks predicted as "up" as C_{chosen}.

$$Pred_i = \begin{cases} up, & if \text{ “}up\text{” } \in Res_i \\ down, & else \end{cases} \tag{10}$$

$$C_{chosen} = \{c_i | Pred_i = \text{“}up\text{”}\} \tag{11}$$

Additionally, we implement a monthly rolling strategy. We hold all the c_i in C_{chosen} for one month. The proportion of each stock in the investment portfolio is calculated using the capitalization weighting method.

$$AR_m = AR_{m-1} + \sum_{c_i \in C_{chosen}} w * R_{c_i} \tag{12}$$

where AR_m is the accumulated return of month m, and R_{c_i} is the return of stock c_i. w represents the proportion of stock c_i in the portfolio. v_i is the market value of company c_i.

$$w = \frac{v_i}{\sum_{n=1}^{N} v_n} \quad (13)$$

3.3 Task 2: Financial Q&A

Stock-Agent has the ability to perform financial Q&A, which includes four parts: agent decision, vector DB construction, knowledge rerank, and response generation.

Given a dialogue history H_t, user query Q_t, and the document d_j retrieved by $agent$, conversation system π can give a response R_t.

3.3.1 Agent Decision

When investors pose a query, our agent automatically selects the appropriate retrieval method based on the intent of the user query. The agent searchs for knowledge within a local database and interacts with API by generating a valid JSON object to obtain information. This adaptive retrieval approach enhances precision, minimizes the storage needs of the local database, increases efficiency, and lower costs:

$$r_j = Agent(Q) \quad (14)$$

$$d_j = \begin{cases} API(r_j), & r_j = 0 \\ DB(r_j), & r_j = 1 \end{cases} \quad (15)$$

where d_j is the document related to Q.

3.3.2 Vector DB Construction

Vector DB is an important part of RAG, which is used for efficient storage and retrieval of knowledge documents. To improve the accuracy and efficiency of document retrieval, we leverage ChatGPT to extract summary s_k from each document.

$$s_j = ChatGPT(d_j) \quad (16)$$

Given s_j, we obtain the embedding vector e_{s_j} via an embedding model. This vector will be stored in the database for subsequent retrieval.

$$e_{s_j} = SentEmbed(s_j) \quad (17)$$

where $SentEmbed$ is a sentence embedding model, such as BGE (Xiao et al., 2023) and SGPT (Muennighoff, 2022). We adopt BGE as the embedding model in our framework.

3.3.3 Knowledge Rerank

To retrieve knowledge in the vector DB, user query Q would also be fed into the same sentence embedding model to obtain the embedding vector e_Q.

$$e_Q = SentEmbed(Q) \quad (18)$$

In the retrieval approach, our framework integrates traditional sparse retrieval (BM25) with vector search methods (Cosine similarity). We take the top k documents retrieved by both methods and apply Reciprocal Rank Fusion (RRF) for re-ranking to acquire the top n documents for the query.

$$d_{emb} = \arg\max_{d_j} \frac{e_Q^\top \cdot e_{s_j}}{|e_Q||e_{s_j}|} \quad (19)$$

$$d_{bm25} = BM25(D, Q) \quad (20)$$

$$d_{db} = \text{RRF}(d_{emb}, d_{bm25}) \quad (21)$$

We combine the locally retrieved d_{db} with d_{api} within API request to obtain the final document d^*.

$$d^* = d_{api} \cup d_{db} \quad (22)$$

3.3.4 LLM Fine-Tuning

We inherit $StockGPT_1$ as the base LLM in this part, then continue training $StockGPT_1$ on the research dataset, financial news, and StockQA datasets of FinAgent to obtain $StockGPT_2$.

3.3.5 Response Generation

Given a dialogue history H_t, user query Q_t, and retrieved document d^* related to Q_t, the goal is to give the response R_t in a conversation of turn t.

We provide a template $Prompt_2$. Then, we concatenate the $Prompt_2$, retrieved knowledge, conversation history, and query to get the input I_t for StockGPT to get the response R_t.

$$I_t = concat(Prompt_2, d^*, H_t, Q_t) \quad (23)$$

$$R_t = StockGPT_2(I_t) \quad (24)$$

4 Experiments

In the following section, we conduct a series of experiments designed to validate the efficacy of Stock-Agent. Owing to the architecture or our framework, the experiments are divided into two parts. The initial part examines the model's ability to predict stock trends. In the second part, we evaluate the performance of our Stock-Agent via ablation study and preference evaluation with human & GPT-4.

4.1 FinAgent -Test Datasets

We select a subset from the FinAgent dataset as the test set, which is excluded from the training set. For task 1, we choose a certain number of samples from the financial reports dataset.

In task 2, the test set comprised some samples from StockQA dataset.

Model	ARR ↑	AERR ↑	ANVOL ↓	SR ↑	MD ↓	CR ↑	MDD ↓
SSE50	-1.0%	-2.7%	19.3%	-0.054	45.9%	-0.023	29
CSI 300	1.7%	0	18.2%	0.092	39.5%	0.043	30
SCI	3.9%	2.2%	14.8%	0.266	21.5%	0.183	19
CNX	7.6%	5.9%	26.5%	0.287	41.3%	0.185	20
Randomforest	9.8%	8.1%	19.5%	0.501	16%	0.608	22
RNN	8.1%	6.4%	10.9%	0.742	15.7%	0.515	12
BERT	10.7%	9.0%	16.1%	0.664	13.5%	0.852	14
GRU	11.2%	9.5%	13.7%	0.814	14.6%	0.765	21
LSTM	11.8%	10.1%	15.4%	0.767	15.3%	0.768	19
Logistic	12.5%	10.8%	27.1%	0.463	32.5%	0.385	18
XGBoost	13.1%	11.4%	20.5%	0.633	20.9%	0.619	17
Decision Tree	13.4%	11.7%	19.6%	0.683	11.9%	1.126	20
ChatGLM2	8.1%	6.4%	24.9%	0.324	62.6%	0.126	26
ChatGPT(Turbo3.5)	14.3%	12.6%	27.7%	0.516	53.6%	0.267	23
FinMa	15.7%	14.0%	37.1%	0.422	66.3%	0.236	25
FinGPT	17.5%	15.8%	28.9%	0.605	55.5%	0.312	24
Tongyi-Fin	24.1%	22.4%	18.9%	0.827	36.9%	0.653	28
Stock-Agent	**30.7%**	**29.0%**	**18.9%**	**1.6243**	**11.3%**	**2.73**	**11**

Table 2: The main experimental results on FinAgent-Test. ARR (Annualized rate of return) is a core indicator, while the middle indicators (like AERR, ANVOL, etc.) could assist investors in evaluating the model's effectiveness. Since the rate of return usually fluctuates wildly, to ensure the stability of the performance, we run each model 10 times and obtain the average result.

4.2 Baselines

To validate the effectiveness of our Stock-Agent on the test datasets, we select baselines as follows:

- **Major Indices**: We select indices in the Chinese capital market, including the SCI, CSI 300, SSE50, and CNX.
- **ML&DL Algorithms**: We use ML algorithms such as Logistic (Sperandei, 2014) and XGBoost (Chen et al., 2015), and DL models like LSTM (Yu et al., 2019) and GRU, which are employed for time-series prediction.
- **General LLMs**: We choose the general-purpose LLMs like ChatGPT and ChatGLM2. These LLMs have been chosen due to their ability and wide range of applications in NLP.
- **FinLLMs**: In the financial domain, we focus on open-source FinLLMs, such as FinGPT and FinMA, which have been trained for financial tasks.

4.3 Metrics

In task 1, we employ core metrics (ARR) along with supplementary (MD&SR) for a comprehensive assessment of model ability. The ARR gauges profitability, while MD and SR measure risk assessment. In task 2, the Ragas (Ragas, 2023) metric evaluates the output quality of LLMs, complemented by scoring from GPT-4 and experts, establishing a multidimensional framework for performance assessment.

4.4 Comparison Results

As shown in Appendix D, Figure 8 depicts various models' Accumulated Returns (AR) with curves. Notably, starting in 2023, the Stock-Agent achieve the highest AR and continued to rise, demonstrating its effectiveness in stock investment.

By referring to Table 2, Stock-Agent leads with a 30.7% Annualized Return Rate (ARR), further confirming its validity. We derive the following observations: Firstly, ML&DL show potential in stock trend prediction with noteworthy ARR results. Secondly, LLMs often outperform ML&DL when combining financial reports with market data, enhancing stock trend prediction. ChatGPT achieved a 14.3% ARR. Despite LLMs being trained on extensive text data, they lack financial domain expertise. Thus, FinLLMs, fine-tuned for finance, are expected to improve stock prediction efficacy, with Tongyi reaching a 24.1% ARR.

Finally, fine-tuning Stock-Agent on stock charts and report data yielded a 30.7% ARR, enhancing prediction accuracy and returns. Through comprehensive financial data for fine-tuning, we improve prediction accuracy and return, thereby validating the efficacy of Stock-Agent.

4.5 Ablation Study

Model	ARR ↑	SR ↑	Out_len ↑	N/A ↓
ChatGLM3-6B	8.1%	0.324	228.1	52.3%
+ Fin. Reports	15.8%	0.636	17.2	-
+ Charts & CoT	30.7%	1.6243	**260.5**	23.1%

Table 3: The ablation results for task 1 under different training data. Out_len: average length of LLMs outputs. N/A: invalid answer ratio.

Model	ROUGE-1 ↑	ROUGE-2 ↑	ROUGE-L ↑
ChatGLM3-6B	0.2794	0.1944	0.2642
+ Fin. News	0.3838	0.2553	0.3745
+ StockQA	0.4115	0.3028	0.4082

Table 4: The ablation results for task 2 under different training data.

Method	Precision ↑	Recall ↑	Faithfulness ↑
Embed. Retrieval	0.6028	0.8195	0.7412
+ BM25 & RRF	0.6189	0.8324	0.7691
+ CFI & VIE	0.6389	0.8517	0.7784
+ Rerank	0.6717	0.8430	0.8005

Table 5: In the Ragas evaluation framework (three key indicators), we present the ablation study in the task 2 using different retrieval methods.

We conduct three ablation experiments. Firstly, we observe the stock trend prediction ability of our Stock-Agent under different training datasets.

As shown in Table 3, all parts of our training data, including Fin. Reports, Fin. Reports CoT, and Stock Charts, contribute to the final performance. In particular, Stock Charts and CoT data greatly improves ARR and output length, which implies that it helps to enhance the financial analysis ability of the model.

It is worth mentioning that after fine-tuning the financial report data, the LLM's output only includes rise and fall, thus resolving the issue of invalid answers. After fine-tuning with CoT and Stock Charts, our Stock-Agent achieves optimal performance with 30.7% ARR, and the proportion of invalid answers also decreased, reaching 23.1%.

As for the second ablation experiment, we investigate whether the quality of the output improved after fine-tuning the LLM at different data. As shown in Table 4, we observe that the scores of Stock-Agent in terms of rouge1 and rouge2 reach 0.3838 and 0.2553 respectively after fine-tuning with News. Furthermore, it is noteworthy that Stock-Agent achieves optimal performance after fine-tuning with both news and StockQA.

Table 5 illustrates the impact of various components on the performance of Stock-Agent in financial Q&A. We select key metrics from Ragas to assess the quality of the Stock-Agent output. The results indicate that Stock-Agent performs best when all components are utilized.

4.6 Preference Evaluation

We utilize GPT-4 and humans to assess the output quality of each LLM on test datasets. As shown in Figure 4 and Figure 5, we observe that any basic LLMs with our Stock-Agent has an improvement in performance in financial analysis. Moreover, in the assessment by humans and GPT-4, Stock-Agent surpasses ChatGPT in delivering effective content shown in Appendix B.

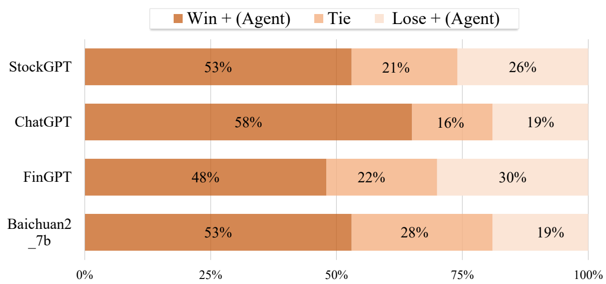

Figure 4: Preference evaluations via human.

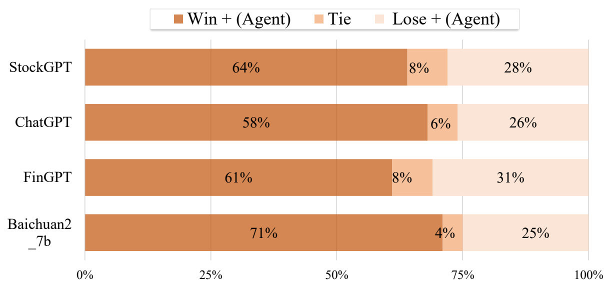

Figure 5: Preference evaluations via GPT-4.

4.7 Case Study

We offer Stock-Agent's results for qualitative analysis. Figure 6 demonstrates that when queries about Apple Inc.'s latest reports and investment guidance, Stock-Agent can retrieve the latest data and reports to feed the LLM. This process refines the LLM's responses, ensuring the information provided is

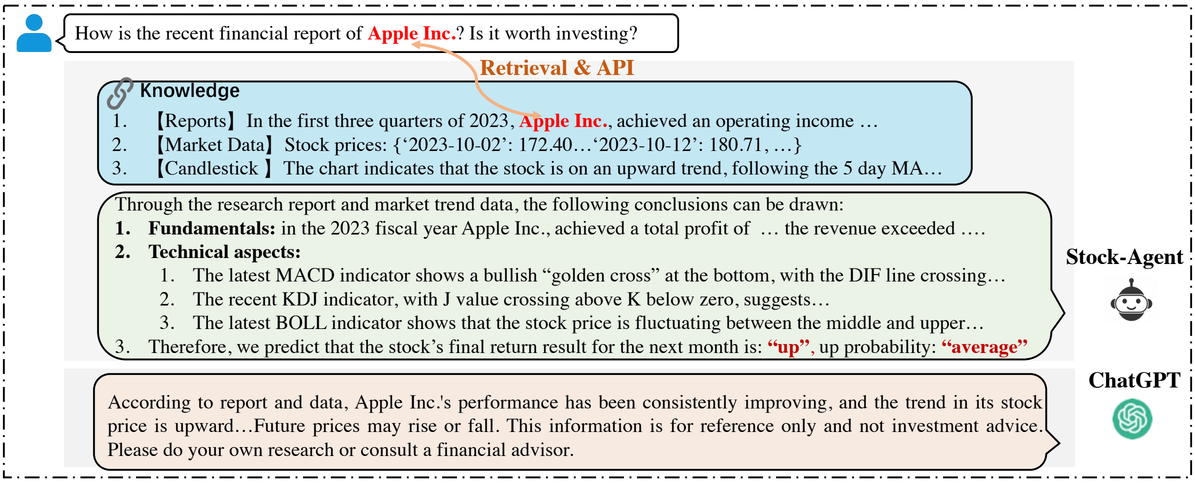

Figure 6: Case of outputs of Stock-Agent and ChatGPT

current, which aids investors in making informed decisions and obtaining relevant advice.

Conversely, with ChatGPT, there are shortcomings in the timeliness and quality of its responses. Therefore, by incorporating Agent, Stock-Agent improves the LLM's utility and performance.

5 Related Work

5.1 Financial Datasets

With the continuous development of financial datasets, DISC-FinLLM (Chen et al., 2023) enhanced the financial-domain capabilities, including various financial-task instructional samples. FinGPT (Zhang et al., 2023) is an open-sourced dataset promoting the development of financial LLMs for transparent and scalable finance-related research. PIXIU (Xie et al., 2023) introduces the instruction tuning dataset specifically tailored for financial LLMs. The scarcity of financial image datasets impedes multimodal LLM development, To make up for this gap, we propose FinAgent to support the training of multimodal LLMs.

5.2 Algorithms in Financial Domain

Traditional ML&DL algorithms, such as LSTM, Logistic, and BERT (Devlin et al., 2018), have been applied in stock trend prediction. However, ML&DL focuses on the final result, without analyzing the underlying factors driving market trend. As for FinLLM, although BloombergGPT, FinMA, and FinGPT play important roles in the community, they are mainly based on English-language datasets. In contrast, Stock-Agent relies on Chinese language and is specifically designed for stock trend prediction.

5.3 Large Multimodal Models

With the continuous development of image and text datasets, multimodal LLMs have made rapid progress. VISCPM (Hu et al., 2023)is a multimodal LLM series from Tsinghua University, specialized in image-text dialogue and optimized for Chinese. Ziya-Visual (Lu et al., 2023), trained on the Ziya general model, excels in visual QA and dialogue, particularly for Chinese users. VisualGLM-6B (Du et al., 2022; Ding et al., 2021), an open-source LLM with bilingual support, blends vision and language, fine-tuned for improved semantic alignment. In Appendix A, the Qwen-VL (Bai et al., 2023) shows strong performance across all metrics, leading us to select it as our primary model for analyzing stock charts.

6 Conclusion

In our study, we define the tasks of financial analysis and propose the FinAgent datasets for fine-tuning StockGPT. Then, we propose a Stock-Agent that efficiently discerns user queries and retrieves information via APIs or knowledge bases. Moreover, we propose an efficient multimodal information fusion method that deeply mines and effectively integrates the retrieved information, thereby improving the analytical quality of LLMs. We conduct extensive experiments on the FinAgent datasets, as well as some supplementary experiments such as ablation studies and GPT4&human preference evaluation, to reveal that Stock-Agent outperforms all the baseline methods, and shows effectiveness for the tasks of financial analysis.

Limitations

Despite the positive contributions of this study, we recognize that there is still great room for development in our work. In our future work, we will further contribute to the open-source FinLLMs, improve their generalization, enhance their ability in other financial tasks, create more powerful open-source FinLLMsand more intelligent investment agents.

Ethics Statement

We assert that there are no ethical dilemmas surrounding the submission of this article and have no known competing financial interests or personal relationships that could have had an impact on the research work presented.

References

AKshare. 2021. Akshare - a financial online documentation.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhihua Wu, Zhen Guo, Hua Lu, Xinxian Huang, et al. 2021. Plato-xl: Exploring the large-scale pre-training of dialogue generation. *arXiv preprint arXiv:2109.09519*.

Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, Rory Mitchell, Ignacio Cano, Tianyi Zhou, et al. 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

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

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. *arXiv preprint arXiv:2210.03849*.

DataYes. 2021. Datayes - financial data and analytics platform.

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*.

Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, et al. 2021. Cogview: Mastering text-to-image generation via transformers. *Advances in Neural Information Processing Systems*, 34:19822–19835.

Mark Dredze, Prabhanjan Kambadur, Gary Kazantsev, Gideon Mann, and Miles Osborne. 2016. How twitter is changing the nature of financial news discovery. In *proceedings of the second international workshop on data science for macro-modeling*, pages 1–5.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335.

Jinyi Hu, Yuan Yao, Chongyi Wang, Shan Wang, Yinxu Pan, Qianyu Chen, Tianyu Yu, Hanghao Wu, Yue Zhao, Haoye Zhang, Xu Han, Yankai Lin, Jiao Xue, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Large multilingual models pivot zero-shot multimodal learning across languages.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Junyu Lu, Dixiang Zhang, Xiaojun Wu, Xinyu Gao, Ruyi Gan, Jiaxing Zhang, Yan Song, and Pingjian Zhang. 2023. Ziya-visual: Bilingual large vision-language model via multi-task instruction tuning.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: financial opinion mining and question answering. In *Companion proceedings of the the web conference 2018*, pages 1941–1942.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Niklas Muennighoff. 2022. Sgpt: Gpt sentence embeddings for semantic search.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior.

Ragas. 2023. Evaluation framework for your retrieval augmented generation (rag) pipelines.

Emad W Saad, Danil V Prokhorov, and Donald C Wunsch. 1998. Comparative study of stock trend prediction using time delay, recurrent and probabilistic neural networks. *IEEE Transactions on neural networks*, 9(6):1456–1470.

Jaimin Shah, Darsh Vaidya, and Manan Shah. 2022. A comprehensive review on multiple hybrid deep learning approaches for stock prediction. *Intelligent Systems with Applications*, page 200111.

Ankur Sinha and Tanmay Khandait. 2020. Impact of news on the commodity market: Dataset and results.

Sandro Sperandei. 2014. Understanding logistic regression analysis. *Biochemia medica*, 24(1):12–18.

Tushare. 2021. Tushare - a financial data interface.

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*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

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.

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

Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. 2019. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270.

Boyu Zhang, Hongyang Yang, and Xiao-Yang Liu. 2023. Instruct-fingpt: Financial sentiment analysis by instruction tuning of general-purpose large language models. *arXiv preprint arXiv:2306.12659*.

This is a section in the appendix.

A Multimodal LLM Evaluation

Acc(Accuracy) reflects whether the LLMs accurately describe the content of the image. Int(Integrity) measure whether the LLMs fully describe the information in the image content. Opera(Operation) evaluates whether the LLMs provide detailed and specific operational recommendations. Log(Logicality) measures whether the description of the LLMs is consistent and rigorous. Expre(Expression) measures whether the expression of a LLMs is professional. Overall is the average of all items, reflecting the comprehensive ability of the LLMs.

We assess advanced multimodal LLMs like VisCPM and Ziya-Visual, which can handle text-to-image generation and dialogue. These models demonstrate outstanding image comprehension and robust performance in stock prediction tasks.

After undergoing an evaluation by experts in the finance sector, we recognize that both Qwen-VL and gpt4 demonstrate exceptional capabilities. However, due to necessary considerations for risk control, gpt4 appears relatively weaker when it comes to providing specific operational recommendations. Therefore, we decided to select Qwen-VL as our multimodal LLMs, dedicated to interpreting stock charts from the stock market.

LLM	Acc	Int	Opera	Log	Expre	Overall
Qwen-VL	0.61	0.55	0.705	0.775	0.83	0.694
GPT-4	0.67	0.57	0.85	0.85	0.515	0.691
VisualGLM	0.55	0.54	0.735	0.82	0.53	0.636
VisCPM	0.46	0.395	0.445	0.475	0.42	0.439
Ziya-Visual	0.3	0.225	0.37	0.385	0.325	0.327

Table 6: The details of the Multimodal LLM Evaluation.

B GPT-4 VS Stock-Agent

Figure 7 shows the preference evaluations between ChatGPT and Stock-Agent, with experts and GPT-4 serving as referees respectively.

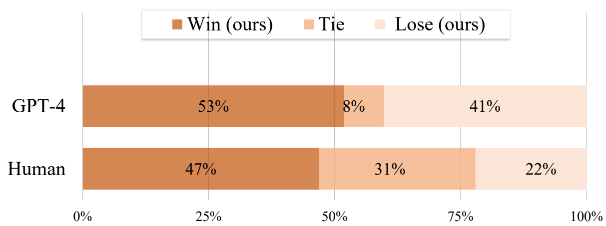

Figure 7: The preference evaluations between ChatGPT and Stock-Agent, with experts and GPT-4 serving as referees respectively.

C Prompt Details

Table C shows the prompts we used in the experiments.

Type	Prompt
CFI-time-series	**# Stock Data** 1. Stock name: (stock_name). 2. Stock price (yuan) : (stock_price). 3. Rise/Fall (%) : (stock_change). 4. Volume (billion) : (stock_volume). **# Restrictions** Your analysis should be limited to: 1. What investment strategy should we adopt at the current stock price? 2. What risk factors exist in the trend of the stock, and how to manage the risk in the current market environment? 3. Based on the current trend, what is the likely future change in the stock price? 4. What kind of technical structure and K-line combination are included in the stock price trend, and what is the impact of these k-line combinations on the future trend? **# Tasks** Your task is to analyze and mine the stock data I provided, and generate a more comprehensive and professional analysis report. Your analysis can ONLY be limited to the above limits. Answers are not allowed to contain fabrications!
CFI-reports	**# Stock Reports** 1. Title: (report_title). 2. Date: (report_date). 3. Target price: (target_price). 4. Report content: (report_content). **# Restrictions** Your analysis should be limited to: 1. Business and market overview: Analysis of the company's main products or services, market positioning, revenue sources and market share; 2. Financial position (core) : in-depth analysis of the company's key financial indicators such as revenue growth, profit margin, cash flow, and debt position; 3. Market trend and macro environment (core) : Consider the influence of industry trend, industry development, policy environment and macro economic factors, including interest rates and policies on the company; 4. Market expectation and enterprise valuation (core) : analyze the company's future revenue and profit expectations in the report, the level of enterprise valuation (pe, pb and other indicators), whether there is a bubble, and evaluate the rationality of these expectations and the possibility of realization; 5. R&d and innovation capabilities: evaluate the company's R&D strength and new product development; 6. Competitive advantage: understand the company's competitive position and advantages in the same industry; 7. Risk factors: identify risks that may affect the company's performance, including operational risks, industry risks, market risks, legal and regulatory risks, etc. **# Tasks** Your task is to sort out and analyze the financial research report I provided, and mine the deeper information of the financial research report from the above several analysis angles, so as to generate a more professional, detailed and comprehensive financial research report. Answers are not allowed to contain fabrications!
VIE	Please describe or analyze in detail based on the diagram provided to you. The description includes seven directions: 1. What is the overall trend of the moving averages (down, up, or sideways); 2. What is the technical structure of the recent K-line portfolio, whether there are buy or sell signals; 3. The latest BOLL indicator, whether there are buy or sell signals; 4. The latest MACD indicator, whether there are buy or sell signals; 5. The recent KDJ indicator, whether there is a buy or sell signal in the volume; 6. Whether the overall fluctuation range of the k line is large; 7. And what the future might hold. The description should be no less than 600 words!

Table 7: Instruction prompts for CFI and VIE technology.

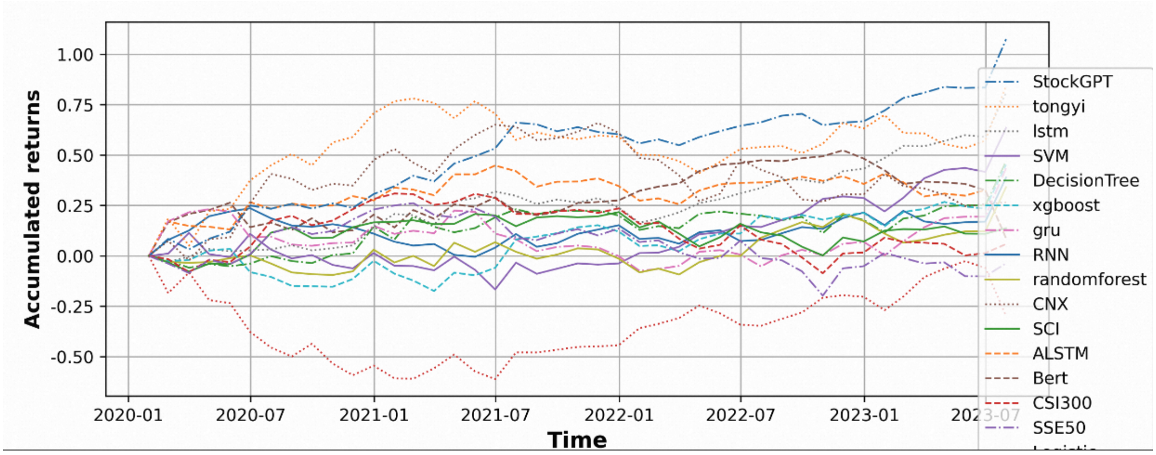

Figure 8: Accumulated returns (AR) of each baseline under the test set of the financial report dataset from January 2020 to July 2023. The figure shows the curves of some baselines.

D Accumulated returns

Figure 8 shows the accumulated returns (AR) of each baseline.