FinRpt: Dataset, Evaluation System and LLM-based Multi-agent Framework for Equity Research Report Generation

Song Jin[1*], **Shuqi Li**[1,2*], **Shukun Zhang**[3], **Rui Yan**[1,4,5†]

[1]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
[2]King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
[3]School of Smart Governance, Renmin University of China, Suzhou, China
[4]School of Artificial Intelligence, Wuhan University, Wuhan, China
[5]School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan, China
jinsong8@ruc.edu.cn, shuqi.li@kaust.edu.sa, zhangshukun@ruc.edu.cn, rui.yan@whu.edu.cn

Abstract

While LLMs have shown great success in financial tasks like stock prediction and question answering, their application in fully automating Equity Research Report generation remains uncharted territory. In this paper, we formulate the Equity Research Report (ERR) Generation task for the first time. To address the data scarcity and the evaluation metrics absence, we present an open-source evaluation benchmark for ERR generation - FinRpt. We frame a Dataset Construction Pipeline that integrates 7 financial data types and produces a high-quality ERR dataset automatically, which could be used for model training and evaluation. We also introduce a comprehensive evaluation system including 11 metrics to assess the generated ERRs. Moreover, we propose a multi-agent framework specifically tailored to address this task, named FinRpt-Gen, and train several LLM-based agents on the proposed datasets using Supervised Fine-Tuning and Reinforcement Learning. Experimental results indicate the data quality and metrics effectiveness of the benchmark FinRpt and the strong performance of FinRpt-Gen, showcasing their potential to drive innovation in the ERR generation field. All code and datasets are publicly available.

Code — https://github.com/jinsong8/FinRpt
Dataset — https://huggingface.co/datasets/jinsong8/FinRpt
Extended version — http://arxiv.org/abs/2511.07322

Introduction

Recently, Large Language Models (LLMs) have reshaped the field of natural language processing and presented remarkable capabilities in many specialized domains across medicine (Tan et al. 2024), law (Izzidien, Sargeant, and Steffek 2024), physics (Polverini and Gregorcic 2024), and finance (Wu et al. 2023; Xie et al. 2023), etc. Within the financial domain, many recent studies have shown great progress in leveraging advanced LLMs into some traditional financial tasks, such as sentiment analysis (Zhang et al. 2023), information extraction (Sharma et al. 2023; Hamad et al. 2024), question answering (Yang, Liu, and Wang 2023), etc., which typically focus on short summaries or brief descriptions.

The media produces a large volume of information about individual companies every day, which includes both considerable noise and valuable insights. Equity research reports (ERRs) (Siantar and Saraswati 2024) play a crucial role in filtering and summarizing this information, providing investors with an in-depth assessment of the company's financial state, market position, and investment potential. An ERR usually consists of many segments, such as a company's financial status statement, risk evaluation, stock trend prediction, etc. Writing it requires specialized financial knowledge and experience, as well as a deep understanding of financial markets, industry trends, and company development, and is usually done by professional analysts.

Recent advancements in LLMs (Yang, Liu, and Wang 2023; Wu et al. 2023), have made automated ERR generation (Jejeniwa, Mhlongo, and Jejeniwa 2024; Adeyeri 2024) a feasible endeavor, which could shorten the time required for company-related information collection and analysis, providing timely insights and recommendations to organizations and researchers, allowing them to respond more quickly to market changes trends. Additionally, ERRs not only provide comprehensive company analyses but also offer good explanations for stock trends, paving the way for improved stock price prediction and advancing other Fintech applications. However, generating ERRs automatically remains an unexplored area due to the following reasons.

From the benchmark perspective, data scarcity (Yagamurthy, Azmeera, and Khanna 2023) is a major obstacle. The input information is typically unstructured and scattered across multiple sources, such as company announcements, industry reports, historical stock prices, and news articles, making it difficult to integrate these diverse data types into a cohesive and standardized format and further map it to the final ERR. Most existing evaluation metrics for generative financial tasks focus primarily on assessing the capabilities of the methods from an NLP view, such as ROUGE-L, and BERTScore (Xie et al. 2023), which are insufficient. Firstly, the evaluation framework should include metrics that assess the accuracy of key indicators in the generated report, such as the cash flow. Secondly, evaluating the accuracy of stock

*These authors contributed equally.
†Corresponding author: rui.yan@whu.edu.cn.

trend predictions is crucial, as it directly influences the potential gains and losses for investors. Finally, since ERRs rely heavily on finance knowledge, evaluating the professionalism of the generated reports is another essential component of the evaluation system.

On the method side, relying only on a single LLM (Wu et al. 2023; Yang, Liu, and Wang 2023) to generate such a complex financial report is hard to achieve. Recently, many LLM-based financial multi-agent frameworks have been developed that could deal with sophisticated scenarios, such as FinAgent (Zhang et al. 2024b) for trading and FinMem (Yu et al. 2024a) for decision-making. An ERR typically consists of multiple sections, each of which needs to reflect an aspect of a company and follow a coherent and logical structure. To tailor a multi-agent framework for generating ERRs is also an issue that requires attention.

To bridge the aforementioned gap, this work makes the first attempt to face the ERR generation task directly and address it. The main contributions of our work could be summarized as follows:

- We formally define the task of ERR generation for the first time, which could be generalized to report generation in other domains.
- We establish an open-source ERR generation evaluation benchmark, named FinRpt, consisting of a Chinese-English ERR dataset and a comprehensive evaluation system. The Dataset Construction Pipeline can automatically generate high-quality ERR data, which could be generalized to the dataset construction in other similar tasks.
- To tackle the newly defined ERR generation task, we tailor a multi-agent framework called FinRpt-Gen, which decomposes the complex task and assigns nine agents to address it collaboratively. Furthermore, we train these agents using Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL), enabling them to achieve optimal performance.

Related Work

Financial Agents Many LLM-based financial agents are developed to tackle complex tasks. FinMem (Yu et al. 2024a) is designed for automated trading. It is optimized by fine-tuning the agent's perception range and character profiles, which improves trading performance and results in higher cumulative investment returns. Zhang et al. (2024b) features a dual-level reflection module and a diversified memory retrieval system, which enhance its ability to make trading decisions. FinCon (Yu et al. 2024b) is structured with a hierarchical manager-analyst model, drawing inspiration from real-world investment firms. It also features a dual-level risk control system to optimize investment strategies and effectively mitigate risks. CryptoTrade (Li et al. 2024b) integrates both on-chain and off-chain data sources and employs a reflective mechanism to enhance trading strategies.

Evaluation of Financial LLMs There are many evaluation benchmarks for the financial domain, such as MME-Finacne (Gan et al. 2024), FinanceBench (Islam et al. 2023), BizBench (Koncel-Kedziorski et al. 2023), FinMME (Luo et al. 2025), PIXIU (Xie et al. 2023) and FinBen (Xie et al. 2024), which mainly focus on Financial NLP capabilities like information extraction and question answering. Among these, finance-related multiple-choice questions (Xie et al. 2023) are used to assess the model's understanding of financial knowledge. Common metrics like ROUGE (Lin 2004) and BERTScore (Zhang et al. 2019) are widely used to evaluate alignment, factual consistency, and information retention. However, this focus limits the ability to comprehensively assess LLMs across a broader range of complex financial applications. Equity Research Report generation encompasses many of these evaluation demands, making it crucial to design a specialized benchmark that can effectively assess ERR generation capabilities and foster its advancement.

Equity Research Reports Generation Some existing multi-agent frameworks are tailored for similar tasks with ERR generation. FinRobot (Yang et al. 2024) is an open-source AI agent platform for financial applications. By prompting LLMs directly, the system could generate well-structured financial analysis. A key limitation of FinRobot is its reliance on fixed annual reports, which restricts its ability to use real-time or diverse data. Another notable example is FinReport (Li et al. 2024a), an explainable stock earnings forecasting model. Unlike ERR, the report generated by the model emphasizes explanations of stock predictions and risk assessments obtained through specially trained modules. Fons et al. (2025) focuses on using LLMs to generate analytical reports for financial time series. However, due to the complexity of ERR generation task, there is still a need to design a customized framework specifically tailored to address it effectively.

FinRpt: Task, Benchmark and Method

ERR Generation Task Formulation

This work formally defines the ERR generation task for the first time. Given a company's stock ticker s and the research date t, the system automatically gathers and structures recently relevant information, and then utilizes it to generate an ERR R. This setup replicates the workflow of a real-world research analyst when drafting an ERR.

In this paper, the input information source $S = [O, F, A, N, P, M]$ includes Company Information O, Financial Indicators F, Company Announcements A, Company-related News N, Historical Stock Prices P, and Historical Market Indices M. To define the output ERR format, we summarize that an ideal ERR of a company should at least include 6 key segments, despite varying formats across securities firms: Financial Analysis R_{fin}, News Analysis R_{news}, Management and Development Analysis R_{manage}, Risks Analysis R_{risk}, Investment Potential Assessment R_{invest}, and Recommendation Rating R_{rec} (recommend a buy rating or a sell rating). We show a generated ERR case in the Appendix.

Dataset

To address the issue of data scarcity for this task, we construct a high-quality ERR generation dataset. In this section,

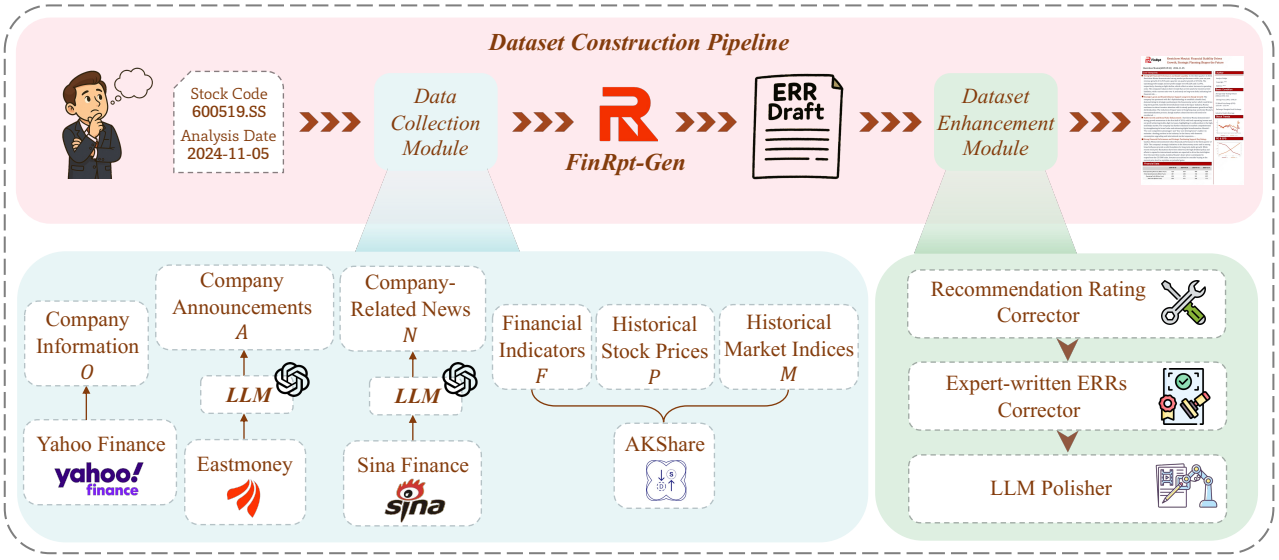

Figure 1: The Dataset Construction Pipeline, Data Collection Module, and the Dataset Enhancement Module.

we will thoroughly describe both the Data Collection Module and the Dataset Construction Pipeline. To enhance clarity, we have visualized the entire process in Figure 1, which illustrates the sequential steps involved in both data collection, dataset construction, and dataset enhancement.

Data Collection Module High-quality data ensures that the analysis, predictions, and insights derived from it are meaningful and trustworthy. Thus, a well-designed data collection module is necessary, which should be capable of gathering key information from various credible sources that provide timely and comprehensive information about a company.

Building on insights from previous research (Zhang et al. 2024b; Penman 2013; Greenwald et al. 2020; Yu et al. 2024a; Zhang et al. 2024a; Yu et al. 2024b; Fatemi and Hu 2024), we carefully selected six valuable and complementary types of company-related data and integrated them into our data collection module:

(1) Company Information O: providing foundational company information (Yu et al. 2024b). (2) Financial Indicators F: reflecting the company's operational status and released quarterly (Fatemi and Hu 2024). (3) Company Announcements A: about significant changes in investment decisions, personnel changes, or unexpected events, reflecting the company's management and development (Zhang et al. 2024a). Besides, we utilize GPT-4o-mini to summarize the announcements. (4)Company-related News N: reflecting events related to a certain company, influencing investor sentiment, and impacting stock trends (Yu et al. 2024a). Similar to announcements summarization, GPT-4o-mini is used to summarize news content and filter out news irrelevant to the company's stock. Besides, we remove brief articles and leverage BERTScore (Zhang et al. 2019) and MinHash (Broder 1997) to de-duplicate similar news. (5) Historical Stock Prices P: reflecting the value of a company to some extent and providing valuable insights into the potential investment assessment (Zhang et al. 2024b). (6) History Market Indices M: reflecting market conditions, as well as investor enthusiasm and confidence (Yoo et al. 2021).

Dataset Construction Pipeline To bridge the gap of data scarcity for ERR generation, we construct an ERR dataset, which consists of 800 stocks in the CSI800 Index of the Chinese market. The corresponding companies generally have a high market value, which results in a substantial amount of information being available in the media. The data range is from September 3, 2024, to November 5, 2024, with intervals of one week between analysis dates, resulting in 10 analysis dates for each company stock. The dataset has 6825 data samples (each sample including the input source information and the corresponding ERR). First, we use the Data Collection Module to gather the input information $S = [O, F, A, N, P, M]$ for each stock ticker s and analysis date t forming an input (s, t, S). We then apply a filtering process to enhance the quality of input data (s, t, S) that excludes data lacking financial indicators F, those with fewer than two news articles N, and those with summarized announcement A lengths under 300 Chinese characters.

Then we apply the multi-agent framework FinRpt-gen, introduced in the next Section, with GPT-4o as each LLM agent, to generate ERRs R automatically, which leads to a complete data input-output sample (s, t, S, R). To align the generated ERRs with the expert-written ERRs, we develop a Dataset Enhancement Module to enhance the quality of generated ERRs:

(1) Recommendation Rating Corrector: for each sample (s, t, S, R), the recommendation rate R_{rec} in R is compared to the ground truth trend label. If they are not consistent, the ERR is regarded as invalid, and this sample will be re-inferred until the correct predictions are generated. (2) Expert-written ERRs Corrector: for each sample (s, t, S, R), we retrieve reports related to the stock s during the week preceding the analysis date t from Eastmoney as the reliable

ERRs $R_{experts}$. Then the retrieved reports $R_{experts}$ along with the generated reports R, are used to prompt an LLM GPT-4o to review and refine the information accuracy, logical consistency, and writing style. The detailed prompt is shown in the Appendix. (3) LLM Polisher: last, we input each ERR R into an LLM GPT-4o for writing polish, enhancing its readability, coherence, and logical flow.

Based on the above processing steps, the high-quality ERR dataset FinRpt is constructed, including 6,825 ERRs. We also provide a corresponding English-translated version. It could be used for ERR generation evaluation, supervised fine-tuning, and reinforcement learning.

Dataset Statistics We analyze several statistics of the constructed dataset - FinRpt. It contains a total of 6,825 reports from 2024-09-03 to 2024-11-05. On average, there are about 9 reports per stock, with a total of 683 reports per analysis date. The detailed industry-wise statistics are shown in Figure 2.

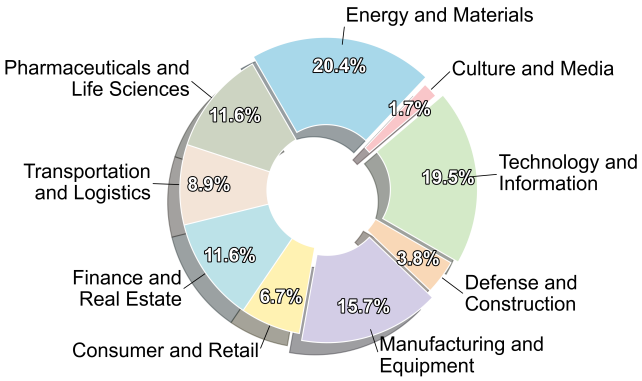

Figure 2: The proportion of reports from different industries of the FinRpt dataset.

We partitioned the dataset as follows: data before 2024-10-31 was randomly split into a training set and a validation set with a 9:1 ratio. Samples after 2024-10-31 were used as the test set. As a result, the training set contains 5,556 samples, the validation set contains 617 samples, and the test set consists of 652 samples.

Proposed Baseline: FinRpt-Gen

FinRpt-Gen The task of ERR generation requires the model to have extensive financial knowledge, a standardized report writing style, and exceptional logical analysis and forecasting abilities. In this work, we propose FinRpt-Gen, as shown in Figure 3, which is the first multi-agent framework specifically designed for the ERR generation task. Given the constructed dataset FinRpt, FinRpt-Gen consists of three modules: an Information Extraction Module, an Analysis Module, and a Prediction Module, involving nine agents playing different roles. We show the prompt examples for every agent in the Appendix.

Information Extraction Module The information extraction module extracts related information from the given input data (s, t, S). This module involves four different agents:

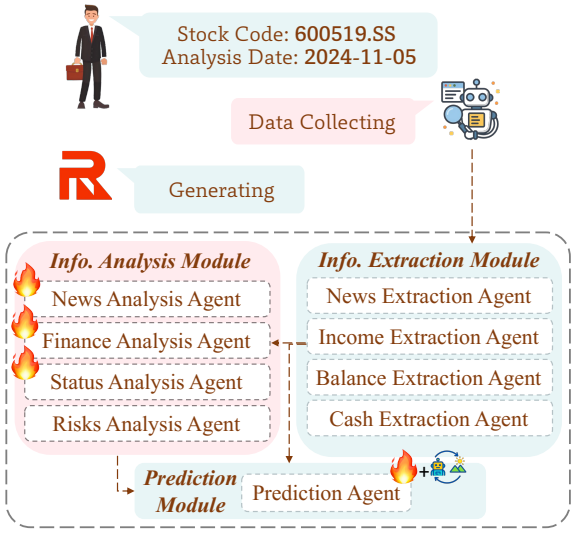

Figure 3: The framework of the proposed FinRpt-Gen.

(1) News Extraction Agent: ranking the provided news articles N by the impact of the news on the stock s and outputting the top 10 news articles most likely to influence stock prices. (2) Income Extraction Agent: given the income statement in financial indicators F, extracting key financial metrics such as revenue, net income, earnings per share, etc. (3) Balance Extraction Agent: given the balance sheet in financial indicators F, focusing on key financial indicators such as assets, liabilities, and equity. (4) Cash Extraction Agent: given the cash flow statement in financial indicators F, focusing on cash from operations, investing, and financing activities.

Based on the well-designed prompts and the table understanding ability of LLMs (Sui et al. 2024), these agents can effectively extract key financial indicators and news for further analysis. According to the ERR format that has been defined previously, we devise the following Analysis Module and Prediction Module to systematically complete the six specific sections of an ERR $R = [R_{fin}, R_{news}, R_{manage}, R_{risk}, R_{invest}, R_{rec}]$.

Information Analysis Module (1) Finance Analysis Agent: given the output of the Income, Balance, Cash Analyst Agent, this agent summarizes the company's financial health, profitability, and cash flow position, generating the financial analysis R_{fin}. (2) News Analysis Agent: given the output of the News Analyst Agent, this agent emphasizes how the selected news may affect the future stock performance, then produces news analysis R_{news}. (3) Status Analysis Agent: derives the management and development analysis R_{manage} from recent company announcements. (4) Risk Analysis Agent: integrates the financial, news, management, and development analysis content from the aforementioned analysis agents to analyze the key risks R_{risk} that should be paid attention to.

Prediction Module The Prediction Agent in the prediction module collects the analysis content of R_{fin}, R_{news}, R_{manage} and R_{risk} along with the historical stock prices P

and history market indices M, then forecasts the investment potential assessment R_{invest} and the recommendation rating R_{rec}. A recommendation rating refers to an evaluation or assessment given to an investment that reflects the analyst's opinion or suggestion regarding its performance potential, which is typically categorized as "buy" or "sell".

Supervised Fine-Tuning (SFT) Within the FinRpt-gen framework, we focus on fine-tuning the four most critical agents: Finance Analysis Agent, News Analysis Agent, Status Analysis Agent, and Prediction Agent. These agents handle the most complex tasks of generating deep, professional insights. We use demonstration samples from the corresponding sections of the FinRpt dataset. For example, given a data sample (s, t, S, R), for the Finance Analysis Agent, the input is the content generated from the Income Extraction Agent, Balance Extraction Agent, and Cash Extraction Agent. The output is the Financial Analysis R_{fin} section. The fine-tuning leverages SFT with LoRA (Hu et al. 2022), aiming to learn a set of low-rank adapter parameters $\Delta\theta$ to maximize the likelihood of generating the target text Y given an input X. The optimization objective is formulated as:

$$\max_{\Delta\theta} \sum_{(X,Y)\in D_{\text{demo}}} \log P(Y|X; \theta_0 + \Delta\theta),$$

where D_{demo} is the demonstration dataset for the respective agent, θ_0 represents the original parameters of the model, and $\Delta\theta$ are the low-rank parameters learned via LoRA.

Reinforcement Learning (RL) To further enhance the Prediction Agent, we introduce a reinforcement learning phase following SFT. This stage moves beyond pattern imitation to optimize the agent's output for real-world investment objectives. We employ DAPO (Yu et al. 2025), an advanced policy gradient algorithm derived from GRPO (Shao et al. 2024), to align the agent's generation with key metrics of accuracy and rationale quality.

First, We design a reward function, $\text{Reward}(Y, Y^*)$, to holistically evaluate a generated response $Y = [R_{\text{invest}}, R_{\text{rec}}]$ against its ground-truth $Y^* = [R^*_{\text{invest}}, R^*_{\text{rec}}]$. This reward is a weighted combination of the recommendation rating R_{rec} accuracy and the quality of the investment analysis R_{invest} measured by ROUGE (Lin 2004). It is defined as follows:

$$\begin{aligned}\text{Reward}(Y, Y^*) = \alpha \cdot \text{ACC}(R_{\text{rec}}, R^*_{\text{rec}}) \\ + \beta \cdot \text{ROUGE-1}(R_{\text{invest}}, R^*_{\text{invest}}) \\ + \gamma \cdot \text{ROUGE-L}(R_{\text{invest}}, R^*_{\text{invest}}),\end{aligned}$$

where α, β, and γ are hyperparameters that balance the importance of each component. In our configuration, we set these parameters to $\alpha = 0.6, \beta = 0.2, \gamma = 0.2$. This setup is designed to prioritize recommendation accuracy while also considering the quality of the analytical content.

The DAPO algorithm then optimizes the policy π_θ by maximizing a clipped surrogate objective, a principle inherited from PPO (Schulman et al. 2017) to ensure stable training. The objective can be conceptually expressed as:

$$\mathcal{J}(\theta) \approx \mathbb{E}\left[\min\left(r(\theta)\hat{A}, \text{clip}(r(\theta), 1-\epsilon_l, 1+\epsilon_h)\hat{A}\right)\right].$$

Here, $r(\theta)$ is the probability ratio between the new and old policies, and $\hat{A}$ represents the standardized advantage of a generated sequence. This objective encourages updates that improve rewards while penalizing large policy shifts. The full, detailed formulation is provided in the Appendix.

Evaluation System

To comprehensively evaluate the generated ERRs, we devise a comprehensive evaluation system, which along with the previously constructed FinRpt dataset, forms a complete benchmark.

Basic Metrics Basic metrics are used to evaluate the generated ERRs from the perspective of text similarity and prediction accuracy. (1) CompletionRate: reveals the proportion of cases where the method successfully generates the ERR in the required format. (2) Accuracy: evaluates the accuracy of the generated recommendation rating (buy or sell). (3) BERTScore (Zhang et al. 2019): evaluates semantic similarity using BERT embeddings. (4) ROUGE-L (Lin 2004): one of the ROUGE metric family, which is commonly used for evaluating the quality of summaries, text generation, and machine translation. (5) NumberRate: measures the size of the mathematical number in the generated report N_{gen} to that in the reference ERR N_{ref}, revealing the richness of numerical content in the generated report. It is calculated by $\text{NumberRate} = \min(N_{gen}/N_{ref}, 1)$.

LLM Evaluations To assess the generated ERRs from the aspects of semantic meaning and context, we developed a set of metrics by referring to relevant academic research (Penman 2013; Greenwald et al. 2020) and industry practices: (1) Financial Numeric (FN): evaluates the precision of the data presented and the depth of the financial analysis in the report. (2) News: assesses how relevant and comprehensive the news analysis is in relation to the company and its stock performance. (3) Company & Market & Industry (CMI): measures the model's insight into the company's management structure, development trajectory, market trends, and overall industry environment. (4) Invest: evaluates whether the investment recommendations are grounded in thorough, logical, and well-reasoned analysis. (5) Risk: assesses how thoroughly the report analyzes the potential risks associated with investing in the stock. (6) Writing: measures overall coherence, readability, and logical consistency.

We leverage a conservative approach to compare the performance of LLMs following previous works (Zheng et al. 2023; Liu et al. 2024; Fu et al. 2023). For each sample, the ERRs generated by different models are compared pairwise using a Judge Agent (GPT-4o). To eliminate position bias, the judge evaluates the ERRs twice, with their order swapped. A win is recorded only if one answer is preferred in both orders; otherwise, the result is marked as a tie. Once the judging process is finished, we could calculate the Win Counts, the Tie Counts, and the Loss Counts.

$$\text{Adjusted Win Rate} = \frac{\text{Win Counts} + 0.5 \cdot \text{Tie Counts}}{\text{Win Counts} + \text{Loss Counts} + \text{Tie Counts}}.$$

We further calculate the Adjusted Win Rate for the sake of comparison.

Category	Method	CompletionRate	Accuracy	ROUGE-L	BERTScore	NumberRate
Single LLMs	XuanYuan-13B-Chat	100%	33%	22.55	68.21	90.32%
	Gemma2-9B	100%	40%	24.97	70.83	38.92%
	Qwen2.5-7B-Instruct	100%	45%	28.08	72.06	82.52%
	Qwen2.5-14B-Instruct	100%	46%	29.74	74.24	83.40%
	Qwen2.5-72B-Instruct	100%	46%	30.17	73.77	90.00%
	Llama3.1-8B-Instruct	100%	45%	26.77	71.63	68.06%
	Llama3.1-70B	100%	46%	29.09	72.99	76.39%
	GLM4-9B-Chat	100%	45%	28.14	73.20	60.67%
	GPT-4o	100%	48%	40.72	79.57	95.72%
	GPT-4o-mini	100%	47%	39.45	78.89	95.75%
	Gemini-2.5-Pro	100%	50%	41.79	80.29	87.23%
FinRpt-Gen with open-source LLMs	FinRpt-Gen (Gemma2-9B)	95%	47%	32.83	72.63	80.99%
	FinRpt-Gen (Qwen2.5-7B-Instruct)	98%	48%	34.51	*72.65*	84.29%
	FinRpt-Gen (Qwen2.5-14B-Instruct)	100%	48%	36.27	77.05	93.16%
	FinRpt-Gen (Qwen2.5-72B-Instruct)	100%	49%	36.88	76.14	94.00%
	FinRpt-Gen (Llama3.1-8B-Instruct)	93%	45%	30.03	68.38	58.92%
	FinRpt-Gen (Llama3.1-70B-Instruct)	97%	48%	34.28	73.43	83.39%
	FinRpt-Gen (GLM4-9B-Chat)	100%	50%	38.35	76.66	76.09%
FinRpt-Gen with closed-source LLMs	FinRpt-Gen (GPT-4o)	100%	51%	48.44	82.09	**98.62%**
	FinRpt-Gen (GPT-4o-mini)	100%	50%	44.09	80.82	97.01%
	FinRpt-Gen (Gemini-2.5-Pro)	100%	51%	48.58	82.12	90.57%
FinRpt-Gen with fine-tuned LLMs	FinRpt-Gen (Llama3.1-8B-Instruct-SFT)	100%	50%	48.67	82.14	94.14%
	FinRpt-Gen (GLM4-9B-Chat-SFT)	100%	51%	48.64	82.16	93.86%
	FinRpt-Gen (Qwen2.5-7B-Instruct-SFT)	100%	54%	48.83	82.21	94.48%
Our	FinRpt-Gen (Qwen2.5-7B-Instruct-SFT-RL)	**100%**	**55%**	**49.06**	**82.43**	95.15%

Table 1: Performance comparison of FinRpt-Gen against baselines under the evaluation of basic metrics.

Experiments

Experiment Setting

Implement Detail The open-source models were accessed via the Ollama Python Library locally, while the closed-source models were accessed through their official APIs. The SFT phase was conducted on 8 NVIDIA 3090 GPUs, and the RL phase used 8 NVIDIA A100 GPUs. We randomly selected 100 samples from the FinRpt test set for evaluation. For detailed hyperparameters and further implementation specifics, please refer to the Appendix.

Baselines We evaluate our method against four types of baselines: (1) standalone state-of-the-art LLMs, (2) our FinRpt-Gen framework with closed-source LLMs, (3) the framework with open-source LLMs, and (4) the framework with fine-tuned open-source LLMs. Please see the Appendix for a more detailed description of all baselines.

Main Results

Main Results of Basic Metrics We compare the performance of FinRpt-Gen against strong baselines, and the results are shown in Table 1, from which we can draw the following conclusions: The performance of the multi-agent framework FinRpt-Gen is significantly better than that of single LLMs, which highlights the effectiveness of our multi-agent framework. In the case without SFT and RL, the performance of the closed-source models Gemini-2.5-Pro and GPT-4o is better than the selected open-source models with a clear margin. This is an expected outcome, as Gemini-2.5-Pro and GPT-4o are widely recognized as leading models in the field. After applying SFT on the constructed dataset FinRpt, there is an obvious improvement in performance compared to the results without SFT, and it even outperforms these two closed-source models in almost all evaluation aspects. After further enhancement through RL, optimal performance is achieved. This partially reflects the high quality of our dataset.

Main Results of LLM Evaluations Based on the LLM evaluation metrics previously detailed, we compare the performance of models from a financial professionalism perspective. The results are shown in Figure 4. And the detailed quantitative results are available in the Appendix. This radar chart illustrates our trained model achieves excellent performance comparable to GPT-4o and surpasses all other strong baselines. Notably, our trained model even excels FinRpt-Gen (GPT-4o) in CMI, News and FN metrics. By comparing the results, we can also conclude the effectiveness of the FinRpt-Gen framework and the training dataset FinRpt. To further validate the reliability of our LLM evaluation, we conducted a human evaluation study in Appendix.

Resource Requirements Analysis The framework's resource requirements are minimal. The entire process of generating an ERR, from data crawling to report creation, is completed in approximately 3 to 4 minutes. For a detailed breakdown of the resource requirements, including processing times and API costs, please refer to the Appendix.

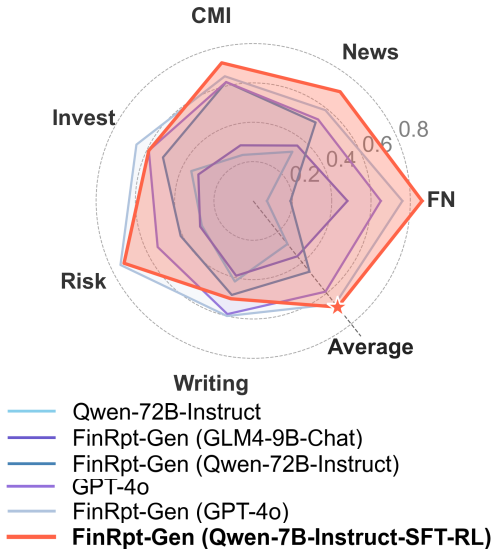

Figure 4: The performance comparison under the LLM evaluation metrics.

Ablation Study

To demonstrate the effectiveness of the components used in FinRpt-Gen (Qwen2.5-7B-Instruct-SFT-RL), abbreviated as FinRpt-Gen in this section, we compare it with four variants: (1) FinRpt-Gen w/o Finance Extraction: removing Income, Balance, and Cash Extraction Agents and inputting the corresponding information into the Information Analysis Module directly. (2) FinRpt-Gen w/o News Extraction: removing the News Extraction Agent and inputting news data into the Information Analysis Agent directly. (3) FinRpt-Gen w/o 3 Analysis Agents: replacing the Finance, Status, and Risks Analysis Agent with one single LLM (GPT-4o).

Method	Accuracy	ROUGE-L	BERTScore
w/o Finance Extraction	47	38,93	76.50
w/o News Extraction	49	46.02	81.20
w/o 3 Analysis Agents	51	45.92	81.38
FinRpt-Gen	**55**	**49.06**	**82.43**

Table 2: Ablation study results over 3 variants.

The ablation results are shown in Table 2, from which we can see that FinRpt-Gen outperforms other variants with clear advantages. The performance of FinRpt-Gen drops significantly without the Finance Extraction Agent, highlighting the necessity of this agent. Similarly, the absence of the News Extraction Agent leads to a noticeable decline in performance, demonstrating the value of pre-extracting and summarizing key news of the agent. Furthermore, utilizing three Analysis Agents outperforms relying on a single GPT-4 model, confirming the importance of the specialized design of each Analysis Agent.

ERRs	FN	News	Invest	Writing	Average
Expert-written	4.57	4.00	4.13	4.33	4.30
FinRpt	4.33	4.20	4.10	4.17	4.20
Kappa Score	0.85	0.86	0.89	0.84	0.86

Table 3: Comparing the quality of ERRs in dataset FinRpt and expert-written ERRs.

Dataset Quality Study

Human Evaluation In the Data Construction Pipeline, the Dataset Enhancement Module is devised to enhance the data quality. To investigate the dataset quality thoroughly, we also conduct a human evaluation. We randomly sample 30 ERRs from the FinRpt dataset and 30 ERRs written by experts. Then three senior financial analysts are invited to rate each ERR from four aspects as devised in LLM Evaluations Section with a scoring range of 0 to 5. These three senior analysts are carefully chosen based on their extensive experience and expertise in the field, ensuring they can provide reliable assessments. We also provide them with detailed evaluation guidelines and criteria to ensure consistency in their judgments. As the results presented in Table 3, the scores of FinRpt and expert-written are very close, suggesting a high level of data quality. Besides, we categorize the evaluation scores of each ERR into three classes and calculate the Fleiss' kappa score (Landis and Koch 1977) (ranging from -1 to 1) to evaluate agreement among the evaluators. The kappa scores are shown in the last line of Table 3, confirming the consistency and reliability of the human evaluation.

Case Study We randomly select an ERR case from FinRpt dataset. The case is presented in Appendix. This case shows the high quality of our dataset from the following four perspectives. (1) The detailed quantitative financial metrics underscore the report's accuracy and thoroughness in financial analysis. (2) The forward-looking strategic analysis offers investors valuable insights into long-term growth drivers and potential risks. (3) The clear investment thesis, supported by quantitative data and strategic context, reflects a balanced approach to stock valuation. (4) The well-structured format ensures clarity and easy navigation, enhancing the report's usability for investors. Collectively, these perspectives highlight the high quality of our dataset, underscoring its utility as a valuable resource for the ERR generation task and other Fintech fields.

Conclusion

This paper formulated the Equity Research Report generation task, and proposed an open-source benchmark FinRpt consisting of a high-quality ERR dataset and a comprehensive evaluation system designed to assess various aspects of ERR generation. Additionally, we tailored a multi-agent framework FinRpt-Gen for this task by applying SFT and RL to our proposed datasets. Experimental results demonstrate the data quality, metric effectiveness of benchmark FinRpt, and strong performance of FinRpt-Gen, highlighting its potential to advance ERR generation.

Acknowledgments

This work was supported by the Public Computing Cloud, Renmin University of China and by fund for building worldclass universities (disciplines) of Renmin University of China.

References

Adeyeri, T. B. 2024. Automating Accounting Processes: How AI is Streamlining Financial Reporting. *Journal of Artificial Intelligence Research*, 4(1): 72–90.

Broder, A. Z. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, 21–29. IEEE.

Fatemi, S.; and Hu, Y. 2024. FinVision: A Multi-Agent Framework for Stock Market Prediction. In *Proceedings of the 5th ACM International Conference on AI in Finance*, 582–590.

Fons, E.; Kochkina, E.; Kaur, R.; Zeng, Z.; Hlavaty, B.; Smiley, C.; Vyetrenko, S.; and Veloso, M. 2025. AI Analyst: Framework and Comprehensive Evaluation of Large Language Models for Financial Time Series Report Generation. *arXiv preprint arXiv:2507.00718*.

Fu, J.; Ng, S.-K.; Jiang, Z.; and Liu, P. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Gan, Z.; Lu, Y.; Zhang, D.; Li, H.; Liu, C.; Liu, J.; Liu, J.; Wu, H.; Fu, C.; Xu, Z.; et al. 2024. MME-Finance: A Multimodal Finance Benchmark for Expert-level Understanding and Reasoning. *arXiv preprint arXiv:2411.03314*.

Greenwald, B. C.; Kahn, J.; Bellissimo, E.; Cooper, M. A.; and Santos, T. 2020. *Value investing: from Graham to Buffett and beyond*. John Wiley & Sons.

Hamad, H.; Thakur, A. K.; Kolleri, N.; Pulikodan, S.; and Chugg, K. 2024. FIRE: A Dataset for Financial Relation Extraction. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 3628–3642.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Islam, P.; Kannappan, A.; Kiela, D.; Qian, R.; Scherrer, N.; and Vidgen, B. 2023. Financebench: A new benchmark for financial question answering. *arXiv preprint arXiv:2311.11944*.

Izzidien, A.; Sargeant, H.; and Steffek, F. 2024. LLM vs. Lawyers: Identifying a Subset of Summary Judgments in a Large UK Case Law Dataset. *arXiv preprint arXiv:2403.04791*.

Jejeniwa, T. O.; Mhlongo, N. Z.; and Jejeniwa, T. O. 2024. A comprehensive review of the impact of artificial intelligence on modern accounting practices and financial reporting. *Computer Science & IT Research Journal*, 5(4): 1031–1047.

Koncel-Kedziorski, R.; Krumdick, M.; Lai, V.; Reddy, V.; Lovering, C.; and Tanner, C. 2023. Bizbench: A quantitative reasoning benchmark for business and finance. *arXiv preprint arXiv:2311.06602*.

Landis, J. R.; and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *biometrics*, 159–174.

Li, X.; Shen, X.; Zeng, Y.; Xing, X.; and Xu, J. 2024a. FinReport: Explainable Stock Earnings Forecasting via News Factor Analyzing Model. In *Companion Proceedings of the ACM on Web Conference 2024*, 319–327.

Li, Y.; Luo, B.; Wang, Q.; Chen, N.; Liu, X.; and He, B. 2024b. A Reflective LLM-based Agent to Guide Zero-shot Cryptocurrency Trading. *arXiv preprint arXiv:2407.09546*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, Y.; Zhou, H.; Guo, Z.; Shareghi, E.; Vulić, I.; Korhonen, A.; and Collier, N. 2024. Aligning with human judgement: The role of pairwise preference in large language model evaluators. *arXiv preprint arXiv:2403.16950*.

Luo, J.; Kou, Z.; Yang, L.; Luo, X.; Huang, J.; Xiao, Z.; Peng, J.; Liu, C.; Ji, J.; Liu, X.; et al. 2025. FinMME: Benchmark Dataset for Financial Multi-Modal Reasoning Evaluation. *arXiv preprint arXiv:2505.24714*.

Penman, S. H. 2013. *Financial statement analysis and security valuation*. McGraw-hill.

Polverini, G.; and Gregorcic, B. 2024. How understanding large language models can inform the use of ChatGPT in physics education. *European Journal of Physics*, 45(2): 025701.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Sharma, S.; Khatuya, S.; Hegde, M.; Shaikh, A.; Dasgupta, K.; Goyal, P.; and Ganguly, N. 2023. Financial numeric extreme labelling: A dataset and benchmarking. In *Findings of the Association for Computational Linguistics: ACL 2023*, 3550–3561.

Siantar, A. L.; and Saraswati, D. 2024. Implementation of Preparation of Financial Reports of MSMEs Based on Financial Accounting Standards for Micro, Small and Medium Entities (SAK-EMKM) Case Study at CV Hubol's. *International Journal of Integrative Sciences*, 3(9): 937–948.

Sui, Y.; Zhou, M.; Zhou, M.; Han, S.; and Zhang, D. 2024. Table meets llm: Can large language models understand structured table data? a benchmark and empirical study. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 645–654.

Tan, Y.; Zhang, Z.; Li, M.; Pan, F.; Duan, H.; Huang, Z.; Deng, H.; Yu, Z.; Yang, C.; Shen, G.; et al. 2024. MedChatZH: A tuning LLM for traditional Chinese medicine consultations. *Computers in Biology and Medicine*, 172: 108290.

Wu, S.; Irsoy, O.; Lu, S.; Dabravolski, V.; Dredze, M.; Gehrmann, S.; Kambadur, P.; Rosenberg, D.; and Mann, G. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Xie, Q.; Han, W.; Chen, Z.; Xiang, R.; Zhang, X.; He, Y.; Xiao, M.; Li, D.; Dai, Y.; Feng, D.; et al. 2024. The finben: An holistic financial benchmark for large language models. *arXiv preprint arXiv:2402.12659*.

Xie, Q.; Han, W.; Zhang, X.; Lai, Y.; Peng, M.; Lopez-Lira, A.; and Huang, J. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*.

Yagamurthy, D. N.; Azmeera, R.; and Khanna, R. 2023. Natural language generation (NLG) for automated report generation. *Journal of Technology and Systems*, 5(1): 48–59.

Yang, H.; Liu, X.-Y.; and Wang, C. D. 2023. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*.

Yang, H.; Zhang, B.; Wang, N.; Guo, C.; Zhang, X.; Lin, L.; Wang, J.; Zhou, T.; Guan, M.; Zhang, R.; et al. 2024. FinRobot: An Open-Source AI Agent Platform for Financial Applications using Large Language Models. *arXiv preprint arXiv:2405.14767*.

Yoo, J.; Soun, Y.; Park, Y.-c.; and Kang, U. 2021. Accurate multivariate stock movement prediction via data-axis transformer with multi-level contexts. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2037–2045.

Yu, Q.; Zhang, Z.; Zhu, R.; Yuan, Y.; Zuo, X.; Yue, Y.; Dai, W.; Fan, T.; Liu, G.; Liu, L.; et al. 2025. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*.

Yu, Y.; Li, H.; Chen, Z.; Jiang, Y.; Li, Y.; Zhang, D.; Liu, R.; Suchow, J. W.; and Khashanah, K. 2024a. Finmem: A performance-enhanced llm trading agent with layered memory and character design. In *Proceedings of the AAAI Symposium Series*, volume 3, 595–597.

Yu, Y.; Yao, Z.; Li, H.; Deng, Z.; Cao, Y.; Chen, Z.; Suchow, J. W.; Liu, R.; Cui, Z.; Xu, Z.; et al. 2024b. FinCon: A Synthesized LLM Multi-Agent System with Conceptual Verbal Reinforcement for Enhanced Financial Decision Making. *arXiv preprint arXiv:2407.06567*.

Zhang, C.; Liu, X.; Jin, M.; Zhang, Z.; Li, L.; Wang, Z.; Hua, W.; Shu, D.; Zhu, S.; Jin, X.; et al. 2024a. When ai meets finance (stockagent): Large language model-based stock trading in simulated real-world environments. *arXiv preprint arXiv:2407.18957*.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhang, W.; Deng, Y.; Liu, B.; Pan, S. J.; and Bing, L. 2023. Sentiment analysis in the era of large language models: A reality check. *arXiv preprint arXiv:2305.15005*.

Zhang, W.; Zhao, L.; Xia, H.; Sun, S.; Sun, J.; Qin, M.; Li, X.; Zhao, Y.; Zhao, Y.; Cai, X.; et al. 2024b. FinAgent: A Multimodal Foundation Agent for Financial Trading: Tool-Augmented, Diversified, and Generalist. *arXiv preprint arXiv:2402.18485*.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623.