

# FinRGAgents: A Multi-Agent Collaboration Framework For Multi-modal Chinese Financial Research Report Generation

Anonymous ACL submission

## Abstract

This paper introduces the Multi-Modal Chinese Financial Research Report Generation (MM-FinRG) task, which focuses on generating timely, visually enriched multimodal financial reports by synthesizing insights from diverse data sources. Distinct from conventional long-form content generation tasks (e.g., summarization, article generation), MM-FinRG prioritizes four critical dimensions: factual accuracy, forward-looking insights, logical consistency, and visual-textual coherence. To address these challenges, we propose FinRGAgents, a novel multi-agent collaboration framework that streamlines financial report generation through a structured three-stage workflow inspired by professional investment research practices: information summarization, plan generation, and report writing. Furthermore, we contribute the FinRG dataset, the first large-scale multi-modal benchmark for Chinese financial reports. Moreover, we design four metrics to rigorously assess report quality across the four key dimensions. Experimental results demonstrate that FinRGAgents achieves state-of-the-art performance, significantly outperforming existing methods on all evaluation criteria. The code, dataset, and evaluation toolkit will be publicly released at: <https://github.com/xxx>.

## 1 Introduction

Multimodal financial research report generation (MM-FinRG) is a complex task that requires close collaboration among team members, each possessing distinct professional skills (Bennett and Gadlin, 2012). However, these reports are typically produced manually by a few seasoned experts. This process not only consumes significant human resources but also incurs substantial intellectual costs. Such reports aim to provide investors with strategic guidance, risk assessment, and revenue forecasts, all crucial for making informed investment decisions (Baker and Nofsinger, 2010). Given the inherent complexity in the creation of financial research

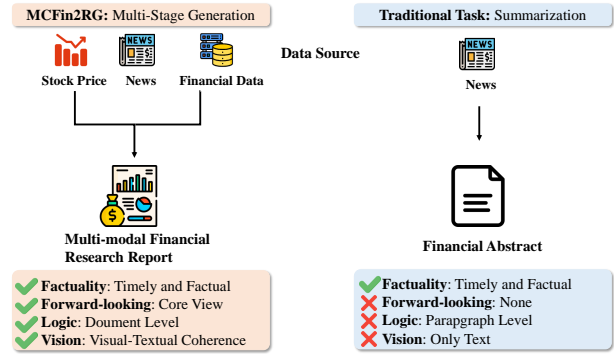


Figure 1: Comparison between our MM-FinRG task (left) and the traditional task (right). Our task can process multi-source data and generate multimodal document-level financial research reports.

reports, there is a growing interest in automating the process of generating these reports.

The existing approach regards financial report generation as a simple summarization task, which generates financial abstract from a single data source (e.g., news). For example, Wang et al. (2024) employs a CVAE with knowledge distillation to generate financial summarization from concise news. This methodology overlooks the rich variety of information sources available in financial markets. In the long-form text generation domain, STORM (Shao et al., 2024) constructs long-form articles through retrieval-enhanced multi-perspective questioning for outline generation, followed by parallel section writing. Simultaneously, LongWriter (Bai et al., 2024) develops a long-form text agent based on GPT-4o to generate the long-form dataset, with subsequent fine-tuning small LLMs to enhance coherent long-form text production capabilities, leaving visually enriched report generation unexplored.

However, these methods can not automatically create visually enriched and structurally clear financial research reports by simply applying them to MM-FinRG due to the following challenges (see Figure 1). First (C1), experts always collect and sum-

marize articles from various data sources to generate high-quality financial research reports. For instance, they will read the stock, news, and financial data and then capture the important information from these articles. Second (C2), the core views and their supporting evidence are important for financial research reports. These views are grounded in logical reasoning assumptions and reinforced by relevant visual charts to enhance the persuasiveness of the viewpoints. Third (C3), the MM-FinRG task generates the reports with multi-modal information text and visual charts while preserving semantic consistency across modalities. Fourth (C4), the MM-FinRG task faces a scarcity of relevant datasets, which hampers progress in this field. Most of the existing datasets focus on long-form text generation or summarization. These challenges highlight the urgent need for advanced methods in automated multimodal financial research report generation.

Motivated by the expert-like potential of autonomous agents, we propose a multi-agent collaborative system for MM-FinRG, named FinRGAgents. Unlike other textual plan-write paradigms (Shao et al., 2024) that generate an intermediate outline first to guide the final output, FinRGAgents can generate authentic and content-rich financial research reports based on real-time information. It operates through three core phases: information summarization (for C1) that reads multiple dataset sources, plan generation (for C2) that summarizes core views and outlines, and report writing (for C3) that generates the multi-modal reports. Additionally, during the plan generation phase, we incorporate the argument-counterargument method to enhance the forward-looking nature of the reports. This approach generates core viewpoints through debate based on external knowledge, which are then used to create the article outline. This ensures the complex reasoning structures and dialectical nature required for financial research reports.

For C4, we built and release a large-scale dataset over 25 domains for MM-FinRG, namely FinRG, to address data scarcity. Moreover, we propose a multidimensional evaluation framework leveraging the LLM-as-a-judge paradigm (Qian et al., 2024; He et al., 2024) to evaluate generated financial research reports. Current metrics such as PPL and ROUGE (Lin, 2004) fail to capture essential aspects of report quality. The framework assesses four dimensions: **Factuality**, **Forward-looking**, **Logic**, and **Vision**, providing both quantitative scores and qualitative feedback. Human evalua-

tion validated the framework’s reliability. Experiments show our method achieves a **3.44** average score across dimensions, demonstrating robust performance in diverse domains.

Our contributions are summarized as follows:

- We introduce the MM-FinRG task, a novel challenge in the field of finance that addresses multimodal report generation. We propose the FinRGAgents, which is the first capable of generating multi-modal financial research reports that include forward-looking views.
- We collect the high-quality FinRG datasets and design a novel metric framework; it is the first comprehensive evaluation framework that assesses multimodal financial reports.
- Using both automatic and human evaluations, we demonstrate that our framework can generate more visually enriched and viewpoints multimodal financial research reports and outperform several baselines.

## 2 Related Works

### 2.1 Report Generation

Extensive research has been conducted on report generation, but the methods vary across different fields. In the medical field (Chen et al., 2021; Wang et al., 2023), multi-modal models are often used to automatically generate a free-text description of a medical image (e.g., a chest X-ray), focusing more on the details within the medical images. In the fashion industry, some works (Ding et al., 2024) utilize GPT4-V to understand the types of clothing and attributes in walkcat images, and perform data analysis to ultimately generate reports for the fashion domain. In the financial sector, previous work (Yan, 2022; Wang et al., 2024) proposes a novel approach using a CVAE with KD to generate financial reports from concise news. While many studies are centered on generating narrative or descriptive texts, our research focuses on producing multimodal financial argumentative text. This approach places a greater emphasis on visually rich content and core viewpoints.

### 2.2 LLM As a Judge

Using large language models as an automatic evaluation metric is explored in some previous work, such as G-Eval (Liu et al., 2023) and LLM Evaluation (Chiang and Lee, 2023). The recent investiga-

| Dataset                          | Task            | InputDataType     | OuputDataType     | Avg.P       | Avg.R       | Domain    |
|----------------------------------|-----------------|-------------------|-------------------|-------------|-------------|-----------|
| MIMIC-CXR (Johnson et al., 2019) | MedicalRG       | Image             | Text              | —           | 102         | 1         |
| News-Reports (Wang et al., 2024) | LTG             | Text              | Text              | 30          | 200         | 1         |
| ArgEssay (Bao et al., 2022)      | AEG             | Text              | Text              | 39          | 342         | 1         |
| CHN-Editorial (He et al., 2024)  | AEG             | Text              | Text              | 171         | 1063        | 1         |
| <b>FinRG(ours)</b>               | <b>MM-FinRG</b> | <b>Multimodal</b> | <b>Multimodal</b> | <b>2054</b> | <b>5723</b> | <b>25</b> |

Table 1: Comparison of FinRG with existing report generation datasets. Avg.P/Avg.R indicates the average length of prompts/reports. AEG and LTG denote Argument Essay Generation and Long-form Text Generation, respectively.

| Industry                 | News  | Reports | Stock | Time Span         |
|--------------------------|-------|---------|-------|-------------------|
| Internet (互联网)           | 3,757 | 1,22    | 1,800 | 2023.01 - 2024.01 |
| Transportation (交通运输)    | 3,798 | 1,27    | 1,492 | 2023.01 - 2024.04 |
| Fiance (金融)              | 4,172 | 1,52    | 1,462 | 2023.01 - 2024.04 |
| Utilities (公用事业)         | 2,142 | 1,08    | 1,101 | 2023.01 - 2024.06 |
| Construction (建筑)        | 3,612 | 93      | 1,297 | 2023.02 - 2024.07 |
| Food and Beverage (食品饮料) | 3,591 | 1,15    | 1,326 | 2023.01 - 2024.07 |
| Culture and Media (文化传媒) | 3,415 | 1,38    | 1,467 | 2023.02 - 2024.08 |

Table 2: Partial basic information for the FinRG dataset is provided, with full descriptions of 25 domains contained in the Appendix A Table 6.

tion (Zheng et al., 2023) shows that such LLM-as-a-Judge methods perform differently on different tasks. Previous works focus on general language generation tasks, and to the best of our knowledge, none of these works targets multimodal financial research report evaluation, which requires specific designs to make the results practically meaningful.

## 2.3 Multi-agent Collaborate

The strong role-playing abilities of LLMs are largely attributed to foundational advances in the technology, as noted in the references (Zhou et al., 2023; Yu et al., 2024). (Hu et al., 2024) can generate multi roles based on topics, integrating various perspectives to produce argumentative text. ChatDev (Qian et al., 2024) proposes a multi-agent framework that introduces the chat chain method to divide each development phase into smaller sub-tasks, enhancing the robustness of software development. (Ni and Buehler, 2024) constructs a larger group of agents with an enhanced division of labor among planning, formulating, coding, executing, and critiquing the process and results, addressing elasticity problems in operations. Moreover, (Li et al., 2023b; Tsao, 2023) employ a multi-agent system to improve trading performance, while (Xing, 2024; Wan et al., 2024) gain valuable insights from its application to financial sentiment analysis and textual information processing. Motivated by recent advances in automated decision-making with language agent based societies, this paper intro-

duces FinRGAgents, a novel LLM-based multi-agent collaborative framework for end-to-end multimodal financial research reports generation.

## 3 FinRG Dataset

### 3.1 Dataset Processing

Our dataset is sourced from Eastmoney<sup>1</sup>, a platform where all financial disclosures are publicly accessible and compliant with open-data regulations. This online community, founded by a professional financial firm, is designed to help investors and investment institutions quickly understand the market trends and business conditions of listed companies. There are 25 industries and 5,128 listed companies in the A-share market; however, not every company possesses extensive financial data. Therefore, we selected the top 60 companies by market capitalization in each industry for analysis and performed several preprocessing steps, including:

- News Data Collection: Collect news data related to the company from the EastMoney information page<sup>2</sup> for the past month, including news titles and main content. Implement preprocessing steps to filter out articles containing fewer than 20 Chinese characters and subsequently perform deduplication based on textual similarity.
- Stock Data Retrieval: Retrieve detailed stock data for companies based on their names and stock codes, including opening price, highest price, lowest price, closing price, and trading volume, to analyze stock price fluctuations over the past months.
- Annual Report Processing: Obtain the latest annual report PDF document. To facilitate understanding by large models, use the Doc2x<sup>3</sup> parsing tool to convert the annual report into an md document, preserving table structures and organizing the content by sections.

<sup>1</sup><https://www.eastmoney.com/>

<sup>2</sup><https://so.eastmoney.com/news/s?keyword>

<sup>3</sup><https://doc2x.noedgeai.com/>

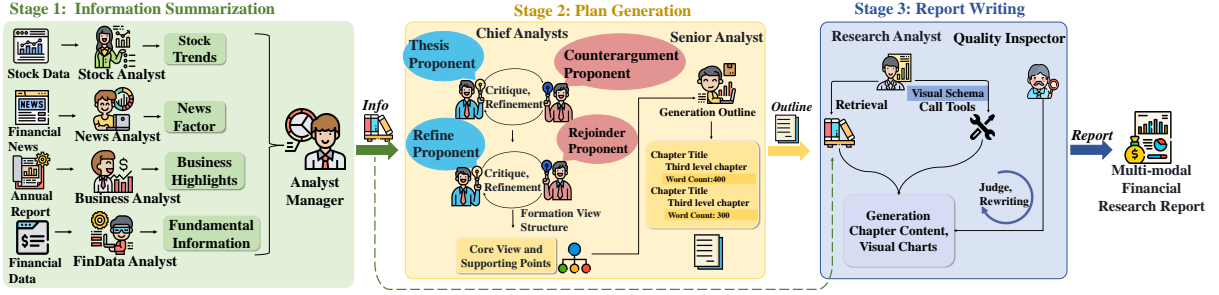


Figure 2: Overview of our framework. Upon receiving a company name, these agents conduct multi-source data organization, then form core insights through deliberative processes and develop a detailed outline plan. Following this outline, they execute a series of subtasks to produce a multimodal financial research report.

- **Financial Metrics Calculation:** Calculate key financial metrics such as the price-to-earnings ratio and price-to-book ratio, using financial formulas to reflect the company’s development indicators.

### 3.2 Data Statistics

The data presented in Table 2 from the Fin2RG dataset illustrates the diversity and volume of multimodal resources across various industries. The Internet and Finance sectors show particularly high activity with substantial news articles and stock data, reflecting the dynamic nature of these industries. Conversely, the Utilities and Construction sectors demonstrate lower figures in financial reports, suggesting less fluctuation and a more stable operational context. Utilizing this data, the model generates multimodal documents, which impose greater demands on the model’s capabilities.

### 3.3 Dataset Comparison

As shown in Table 1, our Fin2RG dataset substantially differs from existing benchmarks in three critical dimensions. First, while prior datasets (e.g., MIMIC-CXR) focus on unimodal inputs (images or text), Fin2RG introduces multimodal financial data integration, requiring joint analysis of news, stock, and financial metrics for report generation. Second, our dataset features significantly richer contextual interactions, with average prompt/report lengths (2,054/5,723 tokens) exceeding existing benchmarks by a factor of 5 to 54, better reflecting real-world financial analysis complexity. Third, FinRG covers 25 financial domains (vs. single-domain baselines), enabling cross-domain reasoning about market risks, and operational performance. This multimodal, long-context, and multi-domain design addresses critical gaps in current report generation research, particularly for financial decision-making scenarios requiring evidence fusion from heterogeneous data sources.

## 4 FinRGAgents

We introduce FinRGAgents (Figure 2), a collaborative multi-agent framework based on AutoGen (Wu et al., 2023) for generating multi-model financial research reports. Inspired by real-world organizational structures, FinRGAgents integrates multiple “report agents” with distinct societal roles (e.g., chief analysts, senior analysts, and research analysts) that collaborate during three core phases of report generation—information summarization, plan generation, and report writing. Each agent having a specific name, role, objective, constraints, along with predefined context, skills, and tools tailored to its functions. Further details can be found in the appendix B.

### 4.1 Problem Formulation

We propose MM-FinRG (Multi-Modal Chinese Financial Research Report Generation), a novel task that requires synthesizing heterogeneous financial data into coherent analytical reports. Formally, given a company profile  $\mathcal{I} = \{c, n, s, a, f\}$ , where  $c$ : company name,  $n$ : Market news (latest 30-day articles),  $s$ : stock indicators (latest 120-trading-day time series data),  $a$ : annual report (PDF document with structured tables and unstructured text),  $f$ : financial metrics (12-month structured data: P/E ratio, etc.) The system generate a multimodal report  $\mathcal{R}$  containing: narrative analysis interpreting market dynamics, visualizations, investment recommendations with risk assessments.

### 4.2 Stage 1: Information Summarization

Based on the type of data source, different agents handle the processing: Stock Analyst Agent utilize technical indicators such as Moving Averages (MA) and MACD to analyze short-term price patterns and forecast near-term market movements. News Analyst Agent summarize and analyze news



data, identifying emerging economic conditions or corporate developments that could trigger abrupt market shifts. Business Analyst Agent analyzes annual reports to summarize business highlights, revenue performance, cash flow status, and risk assessments. FinData Analyst Agent provides foundational insights through financial statement analysis, assessing operational metrics and intrinsic value to determine long-term investment viability.

$$Info = \langle M, F, B, I \rangle = \begin{cases} M = \mathcal{A}_s(s) \\ F = \mathcal{A}_n(n) \\ B = \mathcal{A}_b(a) \\ I = \mathcal{A}_f(f) \end{cases} \quad (1)$$

where  $\mathcal{A}_s$  denotes the Stock Analyst Agent (similarly for others), and  $Info$  represents the integrated multimodal market information.

Analyst Manager Agent cross-references heterogeneous information streams to mitigate inter-source conflicts and contradictions. The information  $Info$  is then passed to the second stage for plan generation.

$$Info = \mathcal{A}_s(Info) \quad (2)$$

### 4.3 Stage 2: Plan Generation

Based on the information from the first phase, Chief Analyst Agents, comprising agents who adopt both pro and con perspectives, engage in multi-round, multi-angle debates to develop the central thesis and sub-arguments for the research report, as detailed in Section 4.5. Subsequently, Senior Analyst Agent generates research report outlines based on the central thesis and sub-arguments, using structured language output to ensure the robustness of the outline. The outline includes second-level headings, third-level headings, and necessary visual schemas. The final report is then drafted in the third stage based on the outline.

$$\bar{V}^m = \mathcal{A}_c(Info), O = \mathcal{A}_{se}(\bar{V}^m) \quad (3)$$

where  $\bar{V}^m$  denotes the core view and  $O = \{o_1, o_2, \dots, o_n\}$  represents the report outline, with  $n$  indicating the total number of sections.

### 4.4 Stage 3: Multi-modal Report Writing

Inspired by the concept of residuals, the Research Analyst Agent retrieves knowledge from the first-phase dataset using predefined outlines. Leveraging Retrieval Augmented Generation (RAG) technology, it generates structured content to

prevent information loss during transfer. The Quality Inspector Agent then conducts compliance checks and quality assessments, ensuring adherence to privacy standards and sensitivity guidelines. If requirements are unmet, the Research Analyst Agent must revise the content.

$$\tilde{r}_i = \mathcal{A}_r(o_i, Info)_{\odot}, r_i = \mathcal{A}_q(\tilde{r}) \quad (4)$$

where  $\tilde{r}_i$  denotes the  $i$ -th unreviewed report content and  $r_i$  represents the corresponding reviewed version after quality control processing.

To maintain visual-textual consistency, the agent automatically identifies sections requiring graphical representation and constructs corresponding visual schemas. visual schemas define equation 5:

$$\text{Schema} \triangleq \begin{cases} \mathbf{x\_axis} \in \mathbb{R}^n & \text{s.t. } x_1 \leq \dots \leq x_n \\ \mathbf{y\_axis} \in \mathbb{R}^n \\ \text{chart\_type} \in \{\text{line, pie, bar}\} \end{cases} \quad (5)$$

where  $\mathbb{R}^n$  denotes the real numbers,  $\mathbf{y\_axis}$  represents the measured quantitative values.

The agent subsequently invokes integrated plotting tools through Python code execution to generate precise visualizations.

$$\mathcal{C}_i = \mathcal{A}_r(\text{Tool}, \text{Schema}_i) \quad (6)$$

where  $\mathcal{C}_i$  represents the  $i$ -th visual charts. All outputs are formatted in Markdown syntax to ensure cross-platform compatibility and facilitate flexible document format conversions.

$$\mathcal{R} = \text{Markdown}(r_1, \mathcal{C}_1, \dots, r_i, \dots, r_n, \mathcal{C}_n) \quad (7)$$

### 4.5 Core View Generation

Stakeholder theory in business (Freeman, 2010; Rohman, 1965) posits that where diverse stakeholders prioritize varying facets of a company, individuals with distinct perspectives may focus on different aspects when researching the same topic and uncover multifaceted information. Therefore, by emulating stakeholder theory, we utilize diverse viewpoints to distill core perspectives and determine research objectives from multiple angles.

We have developed several chief analyst agents, each offering unique perspectives by Freeman's theory (Freeman, 2011). These roles are defined as the thesis proponent agent, the counterargument proponent agent, and the rejoinder proponent agent. Initially, the thesis proponent will formulate a principal view, followed by several sup-

porting views in two sequential steps based on previously summarized information.

$$\begin{aligned} \mathcal{A}_{tp} : \{P, Info\} &\mapsto v^m, \\ &: \{P, Info, v^m\} \mapsto (v_1, v_2, \dots, v_n) \end{aligned} \quad (8)$$

where  $P$  and  $Info$  represent the writing prompt and information summarization knowledge, respectively; The symbol  $\mapsto$  indicates the action of prompting the Agent with a specific prompt to generate desired responses, while  $v^m$  and  $v_n$  denote the draft of the major view and  $n$  supporting views.

**Counterargument** Subsequently, the counterargument proponent agent generates an overriding rebuttal to challenge  $v_n$ . Then, conditioned on the rebuttal, agents are required to optimize  $v_n$ :

$$\begin{aligned} \mathcal{A}_{cp} : \{v_i\} &\mapsto r_i, \\ &: \{v_i, r_i\} \mapsto \bar{v}_i \end{aligned} \quad (9)$$

where  $r_i$  is the generated overriding rebuttal and  $\bar{v}_i$  is the refined view. The overriding rebuttal here functions as feedback, countering the input view and pointing out its inherent weaknesses.

**Rejoinder** In financial reporting, a refined view addresses this weakness, yet gaps remain, potentially involving the questioning of the validity of the view assumption. In addressing such issues, the Rejoinder Proponent Agent first generates an undercutting rebuttal to optimize  $\bar{v}_n$ :

$$\begin{aligned} \mathcal{A}_{rp} : \{\bar{v}_i\} &\mapsto r_i^u, \\ &: \{\bar{v}_i, r_i^u\} \mapsto \tilde{v}_i \end{aligned} \quad (10)$$

where  $r_i^u$  and  $\tilde{v}_i$  denote undercutting rebuttal and refine views respectively.

**Major View** Since each supporting views is revised, the draft major view should also be modified accordingly. Thus, we implement a bottom-to-top update of the major view with refined views  $\tilde{v}_i$ .

$$\mathcal{A}_{tp} : \{\tilde{v}_1, \tilde{v}_2, \dots, \tilde{v}_n\} \mapsto \bar{V}^m \quad (11)$$

## 5 Experiments

### 5.1 Baselines

We evaluate MM-FinRG in a zero-shot setting, where FinRGAgents generate a multi-model Chinese financial research report based multimodal market information, without any prior demonstrations. Our method is assessed using the FinRG dataset, which comprises research reports for 1,500 companies. Evaluations are conducted both automatically and through human assessment.

To verify the effectiveness of our framework, we compare it with the following zero-shot baselines that adopt different generation strategies by prompting LLMs and Multi-Agent Systems: (1) E2E LLM: An end-to-end generation that directly produces the target financial report. (2) CoT LLM: Chain-of-Thought generation (Wei et al., 2022) that initially generates a brief plan as an intermediate guideline and subsequently produces a financial report in the same response. (3) Dual-Agent System: This system involves two agents; one is responsible for generating an outline, while the other produces the financial report. (4) FinRobot (Yang et al., 2024): an open-source multi-agent platform for financial applications, with this paper focusing solely on its capability to generate financial research reports. All models mentioned above are based on GPT-4o, with the temperature set to 0.50.

### 5.2 Evaluation Metrics

**Automatic Evaluation** We adopt the following automatic metrics to evaluate the performance on the FinRG Datasets: (1) BLEU (B-n): We use  $n = 2$  to evaluate n-gram overlap between generated texts and human-written text. (2) Perplexity (PPL): Smaller perplexity scores generally indicate better fluency. We do not count the probability values at the positions where the sentence or discourse token is the golden truth. (3) Distinct-4 (D-4) (Li et al., 2015) : We adopt distinct-4, the ratio of distinct 4-grams to all the generated 4-grams, to measure the generation diversity.

**LLM Evaluation** Evaluating multimodal financial reports remains challenging, particularly in holistic assessment. Given the scarcity of benchmark resources, traditional text generation metrics (Sellam et al., 2020) are insufficient for comprehensive multimodal financial report evaluation. As an initial strategy, we apply four fundamental and objective dimensions that reflect different aspects of financial report quality to evaluate the agent-generated reports. We then integrate these dimensions to facilitate a more holistic evaluation. The detailed scoring criteria provided in Appendix E.

- **Factuality** measures the accuracy and reliability of information presented in a text. Factuality scores range from 1.00 to 5.00 and assess how well factual statements are supported by evidence and credible sources.
- **Forward-looking** evaluates the depth and quality of predictive analysis in a report, quantified by









| Method                   | Paradiagm   | Automatic Evaluation |               |              |             | LLM Evaluation  |             |             |             |
|--------------------------|---|----------------------|---------------|--------------|-------------|-----------------|-------------|-------------|-------------|
|                          |   | BLUE-2               | PPL           | Distinct-4   | Factuality  | Forward-looking | Logical     | Vision      | Avg         |
| E2E                      |  | 0.234                | 340.42        | 0.784        | 2.94        | 3.04            | 3.20        | —           | 3.06        |
| CoT                      |  | 0.235                | 318.87        | 0.827        | <u>3.13</u> | 2.96            | 3.35        | —           | 3.15        |
| Dual-Agent               |  | 0.312                | 283.23        | 0.842        | <b>3.14</b> | 3.02            | 3.41        | —           | <u>3.19</u> |
| FinRobot                 |  | <u>0.341</u>         | <u>273.12</u> | <u>0.861</u> | 2.71        | 2.63            | 2.89        | 2.79        | 2.75        |
| FinRGAgents              |  | <b>0.352</b>         | <b>216.46</b> | <b>0.872</b> | 2.88        | <b>3.60</b>     | <b>3.65</b> | <b>3.62</b> | <b>3.44</b> |
| <i>Ablation</i>          |   |                      |               |              |             |                 |             |             |             |
| <i>w/o Information</i>   |   | 0.285                | 296.21        | 0.734        | 2.43        | 3.55            | 3.46        | 3.45        | 3.22        |
| <i>w/o Core View</i>     |   | 0.278                | 305.47        | 0.741        | 2.64        | 3.46            | 3.38        | 3.52        | 3.25        |
| <i>w/o Visual Schema</i> |   | 0.282                | 294.24        | 0.715        | 2.52        | 3.32            | 3.46        | 3.26        | 3.14        |

Table 3: Overall performance of the LLM-powered financial report generation methods, encompassing both single-agent() , dual-agent() and multi-agent() paradigms. The top scores are in bold, with the second highest underlined. Avg is the average score of Factuality, Forward-looking, Logical and Vision.

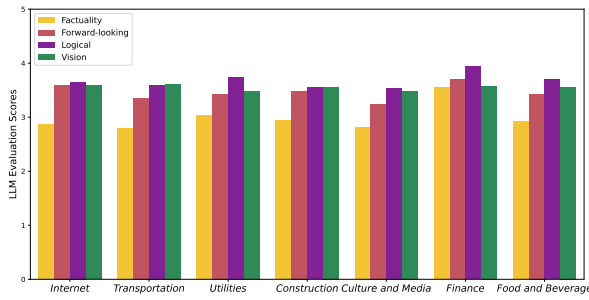


Figure 3: Evaluation results for FinRGAgents across 7 domains are presented using four dimensions: Factuality, Forward-looking, Logical and Vision. Details on the performance of the 25 domains are available in Appendix D Figure 7.

a score range from 1.00 to 5.00. A higher score indicates greater reliability and actionability of future projections.

- Logical evaluates the coherence and organization of a report’s arguments and conclusions, quantified on a scale of 1.00 to 5.00. A higher score reflects a more structured, reasoned, and persuasive analysis, with robust support for conclusions and a clear data-to-insights progression.
- Vision evaluates the alignment between the text and the chart presentation in a report, quantified on a scale from 1.00 to 5.00. A higher score indicates a better connection and consistency between the charts and the text.

**Human Evaluation** For a more comprehensive analysis, we conduct a human evaluation. We hire three well-educated financial master students to score the output quality following the three aspects in the automatic evaluation.

### 5.3 Main Results

We present the evaluation results of the FinRG dataset in Table 3. Our findings are as follows:

| Method     | Evaluator | Baseline Wins | FinRGAgents Wins | Draw  |
|------------|-----------|---------------|------------------|-------|
| E2E        | GPT-4     | 13.79%        | 85.28%           | 0.93% |
|            | Human     | 8.24%         | 91.32%           | 0.44% |
| COT        | GPT-4     | 17.86%        | 81.73%           | 0.41% |
|            | Human     | 10.56%        | 89.31%           | 0.13% |
| Dual-Agent | GPT-4     | 23.52%        | 70.55%           | 5.93% |
|            | Human     | 11.23%        | 87.23%           | 1.54% |
| FinRobot   | GPT-4     | 20.34%        | 73.21%           | 6.45% |
|            | Human     | 12.57%        | 84.21%           | 3.22% |

Table 4: Pairwise evaluation results.

FinRGAgents excel in both automated and LLM-based evaluations. First, it significant improvement over E2E, CoT, and Dual-Agent frameworks demonstrates that complex tasks require multi-step solutions rather than single-step approaches. By decomposing intricate problems into manageable subtasks, the system enhances task completion effectiveness by 9.82% compared to baseline models. Compared to FinRobot, FinRGAgents elevate the LLM evaluation score by 23.4% (from 2.75 to 3.38). This progress is largely attributed to the agents employing an argument-counterargument method, which continuously refines the core view through debate, rather than merely responding based on predefined instructions.

Our further analysis reveals that although FinRGAgents perform exceptionally in terms of forward-looking and logical aspects, they are slightly inferior to Dual-Agent on the factuality metric. This phenomenon may stem from the unique collaborative challenges of the multi-agent paradigm: First, the extended information transmission chain can lead to knowledge decay, where critical details may be lost through multiple processing stages by roles such as News Analyst Agent and Financial Data Analyst Agent. Second, the division of expertise can reduce the accuracy of factual

| Corelation | Factuality | Foward-looking | Logical | Vision | Avg  |
|------------|------------|----------------|---------|--------|------|
| Pearson    | 0.64       | 0.73           | 0.91    | 0.74   | 0.75 |
| Spearman   | 0.68       | 0.84           | 0.89    | 0.79   | 0.80 |

Table 5: The correlation scores between LLM ratings and human ratings under different dimensions.

alignment due to perspective differences, such as potential implicit biases between macroeconomic analysts and industry researchers in data interpretation standards. Furthermore, error propagation in sequential workflows where preprocessing inaccuracies compound downstream. Moreover, the detailed performance of FinRGAgents across 25 domains, as illustrated in Figure 3, underscores the robustness of our approach.

#### 5.4 Human Agreement Evaluation

Recent research highlights findings regarding the evaluation capabilities of large language models (LLMs) for complex tasks. (Tahmid Rahman Laskar et al., 2024) report that LLMs may not effectively gauge complex tasks, prompting investigation into user preferences in practical scenarios. Building on the experimental setup described by (Li et al., 2023a), we assess agent-generated solutions through paired comparisons both human participants and GPT-4 model. The results, in Table 4, reveal that FinRGAgents notably outperforms other baseline models in terms of average win rates across evaluations by both GPT-4 and humans. Furthermore, Table 5 illustrates a promising correlation in ratings between humans and LLMs, with an average Pearson correlation of 0.75 and Spearman correlation of 0.80. This human-LLM correlation not only surpasses other evaluation methods (Zheng et al., 2025) but also suggests that the metrics we developed closely align with human preferences. This reinforces their efficacy and relevance in practical applications.

#### 5.5 Ablation Study

To better understand the impact of each component in our proposed method, we conducted ablation studies using three different configurations. Specifically, we evaluated the method by: (1) providing raw information to the Chief Analyst (*w/o Information*), (2) omitting the argument-counterargument method (*w/o Core View*), and (3) removing the visual structure (*w/o Visual Schema*). Table 3 shows their average performance on the FinRG dataset. Removing *Information* reduces performance, confirming the importance of summarization. Omitting

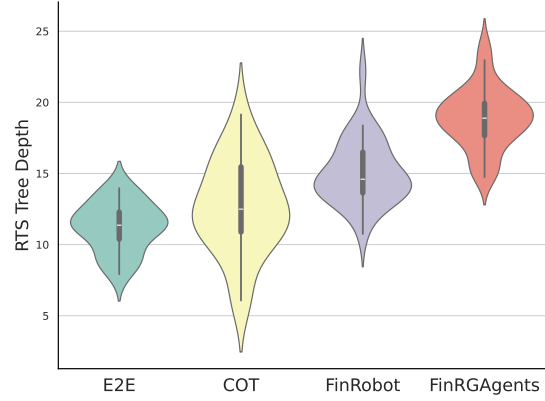


Figure 4: Distribution of RST tree depth of generation financial reports.

Core View decreases Forward-looking and Logical scores, indicating the argument-counterargument method improves report quality. Removing *Visual Schema* lowers the visual quality score from 3.628 to 3.26, suggesting pre-generated visual structures enhance text-chart alignment.

#### 5.6 Analysis on Report Structure

The structure of discourse provides insights into the advanced organization of texts, making the depth of Rhetorical Structure Theory (RST) trees a critical indicator of text quality. Therefore, we parse financial reports into RST trees and analyze the depth distribution of these trees to assess quality, as shown in Figure 4. FinRGAgents typically generate financial reports with deeper structural layers. Moreover, it exhibits a broader depth distribution, covering a wider range than methods like E2E, CoT, and FinRobot. This suggests that our approach effectively contributes to the creation of more diversified and complex structures.

### 6 Conclusion

We have introduced FinRGAgents, an innovative multi-agent collaboration framework for multimodal financial reports that utilizes multiple LLM-powered agents to integrate fragmented phases. It features an argument-counterargument mechanism to refine the core view. Our experiments across data from 25 domains have demonstrated the superiority of our method. Moreover, our proposed evaluation framework and Fin2RG dataset ensure the assessability of financial reports. This research provides a new paradigm for generating financial reports under unsupervised conditions and offers fresh insights for future work in this field.



## Limitations

While our approach demonstrates its capability to produce high-quality financial reports, it still faces inherent challenges that affect its general applicability. For example, achieving SOTA performance on our dataset is impressive but not absolute, which may limit its broader application. Additionally, the data sources are not fully comprehensive—for instance, ECC audio and company announcements are not fully included—which could result in sub-optimal report generation outcomes. Furthermore, although FinRGAgents shows significant improvements in chart generation compared to previous methods, it still lacks the ability to fully leverage visual information to enhance chart consistency. This is often reflected in issues such as incorrect elements in charts or errors in axis unit labeling. Addressing these limitations will require future enhancements, including better integration of visual information into the generation process.

## References

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. Longwriter: Unleashing 10,000+ word generation from long context llms. [arXiv preprint arXiv:2408.07055](#).

H Kent Baker and John R Nofsinger. 2010. [Behavioral finance: investors, corporations, and markets](#), volume 6. John Wiley & Sons.

Jianzhu Bao, Yasheng Wang, Yitong Li, Fei Mi, and Ruifeng Xu. 2022. [AEG: Argumentative essay generation via a dual-decoder model with content planning](#). In [Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing](#), pages 5134–5148, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

L Michelle Bennett and Howard Gadlin. 2012. Collaboration and team science: from theory to practice.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. [Cross-modal memory networks for radiology report generation](#). In [Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing \(Volume 1: Long Papers\)](#), pages 5904–5914, Online. Association for Computational Linguistics.

Cheng-Han Chiang and Hung-yi Lee. 2023. [Can large language models be an alternative to human evaluations?](#) In [Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics \(Volume 1: Long Papers\)](#), pages 15607–15631,

Toronto, Canada. Association for Computational Linguistics.

Yujuan Ding, Yunshan Ma, Wenqi Fan, Yige Yao, Tat-Seng Chua, and Qing Li. 2024. [Fashionregen: Llm-empowered fashion report generation](#). [Companion Proceedings of the ACM on Web Conference 2024](#).

James B Freeman. 2011. [Argument Structure:: Representation and Theory](#), volume 18. Springer Science & Business Media.

R Edward Freeman. 2010. [Stakeholder theory: The state of the art](#). Cambridge University Press.

Yuhang He, Jianzhu Bao, Yang Sun, Bin Liang, Min Yang, Bing Qin, and Ruifeng Xu. 2024. [Decomposing argumentative essay generation via dialectical planning of complex reasoning](#). In [Findings of the Association for Computational Linguistics ACL 2024](#), pages 12305–12322, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. 2024. Unlocking varied perspectives: A persona-based multi-agent framework with debate-driven text planning for argument generation. [arXiv preprint arXiv:2406.19643](#).

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. [arXiv preprint arXiv:1901.07042](#).

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023a. Camel: Communicative agents for "mind" exploration of large language model society. [Advances in Neural Information Processing Systems](#), 36:51991–52008.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. [arXiv preprint arXiv:1510.03055](#).

Yang Li, Yangyang Yu, Haohang Li, Z. Chen, and Khaldoun Khashanah. 2023b. [Tradinggpt: Multi-agent system with layered memory and distinct characters for enhanced financial trading performance](#).

Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In [Text Summarization Branches Out](#), pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuhang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In [Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing](#), pages 2511–2522, Singapore. Association for Computational Linguistics.

|     |   |     |
|-----|---|-----|
| 721 | Bo Ni and Markus J Buehler. 2024. Mechagents: Large language model multi-agent collaborations can solve mechanics problems, generate new data, and integrate knowledge. <i>Extreme Mechanics Letters</i> , 67:102131.   | 776 |
| 722 |   | 777 |
| 723 |   | 778 |
| 724 |   | 779 |
| 725 |   | 780 |
| 726 | Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, et al. 2024. Chatdev: Communicative agents for software development. In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 15174–15186.   | 781 |
| 727 |   | 782 |
| 728 |   | 783 |
| 729 |   | 784 |
| 730 |   | 785 |
| 731 |   | 786 |
| 732 |   | 787 |
| 733 | D Gordon Rohman. 1965. Pre-writing: The stage of discovery in the writing process. <i>College Composition &amp; Communication</i> , 16(2):106–112.  | 788 |
| 734 |   | 789 |
| 735 |   | 790 |
| 736 | Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In <i>Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics</i> , pages 7881–7892, Online. Association for Computational Linguistics.   | 791 |
| 737 |   | 792 |
| 738 |   | 793 |
| 739 |   | 794 |
| 740 |   | 795 |
| 741 |   | 796 |
| 742 | Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in writing Wikipedia-like articles from scratch with large language models. In <i>Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)</i> , pages 6252–6278, Mexico City, Mexico. Association for Computational Linguistics. | 797 |
| 743 |   | 798 |
| 744 |   | 799 |
| 745 |   | 800 |
| 746 |   | 801 |
| 747 |   | 802 |
| 748 |   | 803 |
| 749 |   | 804 |
| 750 |   | 805 |
| 751 | Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, et al. 2024. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. <i>arXiv e-prints</i> , pages arXiv–2407.  | 806 |
| 752 |   | 807 |
| 753 |   | 808 |
| 754 |   | 809 |
| 755 |   | 810 |
| 756 |   | 811 |
| 757 |   | 812 |
| 758 |   | 813 |
| 759 | Wen-Kwang Tsao. 2023. Multi-agent reasoning with large language models for effective corporate planning. 2023 International Conference on Computational Science and Computational Intelligence (CSCI), pages 365–370.   | 814 |
| 760 |   | 815 |
| 761 |   | 816 |
| 762 |   | 817 |
| 763 |   | 818 |
| 764 | Xiangpeng Wan, Haicheng Deng, Kai Zou, and Shiqi Xu. 2024. Enhancing the efficiency and accuracy of underlying asset reviews in structured finance: The application of multi-agent framework. <i>ArXiv</i> , abs/2405.04294.  | 819 |
| 765 |   | 820 |
| 766 |   | 821 |
| 767 |   | 822 |
| 768 |   | 823 |
| 769 |   | 824 |
| 770 | Siyuan Wang, Bo Peng, Yichao Liu, and Qi Peng. 2023. Fine-grained medical vision-language representation learning for radiology report generation. In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 15949–15956, Singapore. Association for Computational Linguistics.  | 825 |
| 771 |   | 826 |
| 772 |   | 827 |
| 773 |   | 828 |
| 774 |   | 829 |
| 775 |   |     |
|     | Ziao Wang, Yunpeng Ren, Xiaofeng Zhang, and Yiyuan Wang. 2024. Generating long financial report using conditional variational autoencoders with knowledge distillation. <i>IEEE Transactions on Artificial Intelligence</i> .   |     |
|     | Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. <i>Advances in neural information processing systems</i> , 35:24824–24837.  |     |
|     | Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Auto-gen: Enabling next-gen llm applications via multi-agent conversation framework. <i>arXiv preprint arXiv:2308.08155</i> .   |     |
|     | Frank Xing. 2024. Designing heterogeneous llm agents for financial sentiment analysis. <i>ArXiv</i> , abs/2401.05799.   |     |
|     | Sixing Yan. 2022. Disentangled variational topic inference for topic-accurate financial report generation. In <i>Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)</i> , pages 18–24, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.  |     |
|     | Hongyang Yang, Boyu Zhang, Neng Wang, Cheng Guo, Xiaoli Zhang, Likun Lin, Junlin Wang, Tianyu Zhou, Mao Guan, Runjia Zhang, et al. 2024. Finrobot: An open-source ai agent platform for financial applications using large language models. <i>arXiv preprint arXiv:2405.14767</i> .  |     |
|     | Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng, Yupeng Cao, Zhi Chen, Jordan W Suchow, Rong Liu, Zhenyu Cui, Denghui Zhang, et al. 2024. Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. <i>arXiv preprint arXiv:2407.06567</i> .   |     |
|     | Hao Zheng, Xinyan Guan, Hao Kong, Jia Zheng, Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and Le Sun. 2025. Pptagent: Generating and evaluating presentations beyond text-to-slides. <i>arXiv preprint arXiv:2501.03936</i> .  |     |
|     | Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. <i>Advances in Neural Information Processing Systems</i> , 36:46595–46623.   |     |
|     | Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, et al. 2023. Webarena: A realistic web environment for building autonomous agents. <i>arXiv preprint arXiv:2307.13854</i> .   |     |

| Industry             | News  | Reports | Stock | Time Span         | Industry         | News  | Reports | Stock | Time Span         |
|----------------------|-------|---------|-------|-------------------|------------------|-------|---------|-------|-------------------|
| Internet(互联网)        | 3,757 | 1,22    | 1,800 | 2023.01 - 2024.01 | TE(交运设备)         | 4,058 | 2,31    | 1,250 | 2023.01 - 2024.01 |
| Transportation(交通运输) | 3,798 | 1,27    | 1,492 | 2023.01 - 2024.04 | LLPS(生活及专业服务)    | 3,972 | 95      | 1,674 | 2023.05 - 2024.05 |
| IT(信息技术)             | 4,883 | 1,57    | 1,734 | 2023.01 - 2024.01 | Utilities(公用事业)  | 2,142 | 1,08    | 1,101 | 2023.01 - 2024.06 |
| AFAHF(农林牧渔)          | 3,358 | 1,24    | 1,342 | 2023.04 - 2024.06 | FE(石化能源)         | 4,521 | 83      | 1,573 | 2023.02 - 2024.03 |
| PB(医药生物)             | 4,312 | 1,43    | 1,384 | 2023.02 - 2024.03 | CR(商贸零售)         | 2,683 | 1,25    | 1,263 | 2023.01 - 2024.01 |
| NDE(国防与装备)           | 3,525 | 1,01    | 1,153 | 2023.01 - 2024.01 | BC(基础化工)         | 3,782 | 2,12    | 1,282 | 2023.03 - 2024.03 |
| HA(家电)               | 3,326 | 94      | 986   | 2023.03 - 2024.03 | BM(建材)           | 4,252 | 1,23    | 1,452 | 2023.02 - 2024.02 |
| Construction(建筑)     | 3,612 | 93      | 1,297 | 2023.02 - 2024.07 | RE(房地产)          | 3,233 | 1,34    | 1,182 | 2023.02 - 2024.02 |
| CM(文化传媒)             | 3,415 | 1,38    | 1,467 | 2023.02 - 2024.08 | Non-fM(有色金属)     | 3,292 | 98      | 1,213 | 2023.01 - 2024.01 |
| ME(机械设备)             | 2,973 | 1,23    | 2,346 | 2023.02 - 2024.02 | Electronic(电子设备) | 4,273 | 1,19    | 1,963 | 2023.02 - 2024.02 |
| Electrical(电气设备)     | 3,254 | 1,34    | 1,634 | 2023.03 - 2024.03 | TA(纺织服装)         | 4,215 | 1,23    | 1,324 | 2023.05 - 2024.06 |
| LIM(轻工制造)            | 3,123 | 87      | 972   | 2023.03 - 2024.04 | Fiance(金融)       | 4,172 | 1,52    | 1,462 | 2023.01 - 2024.04 |
| FB(食品饮料)             | 3,591 | 1,15    | 1,326 | 2023.01 - 2024.07 |                  |       |         |       |                   |

Table 6: Comprehensive overview of the full basic information for 25 domains in the FinRG dataset.

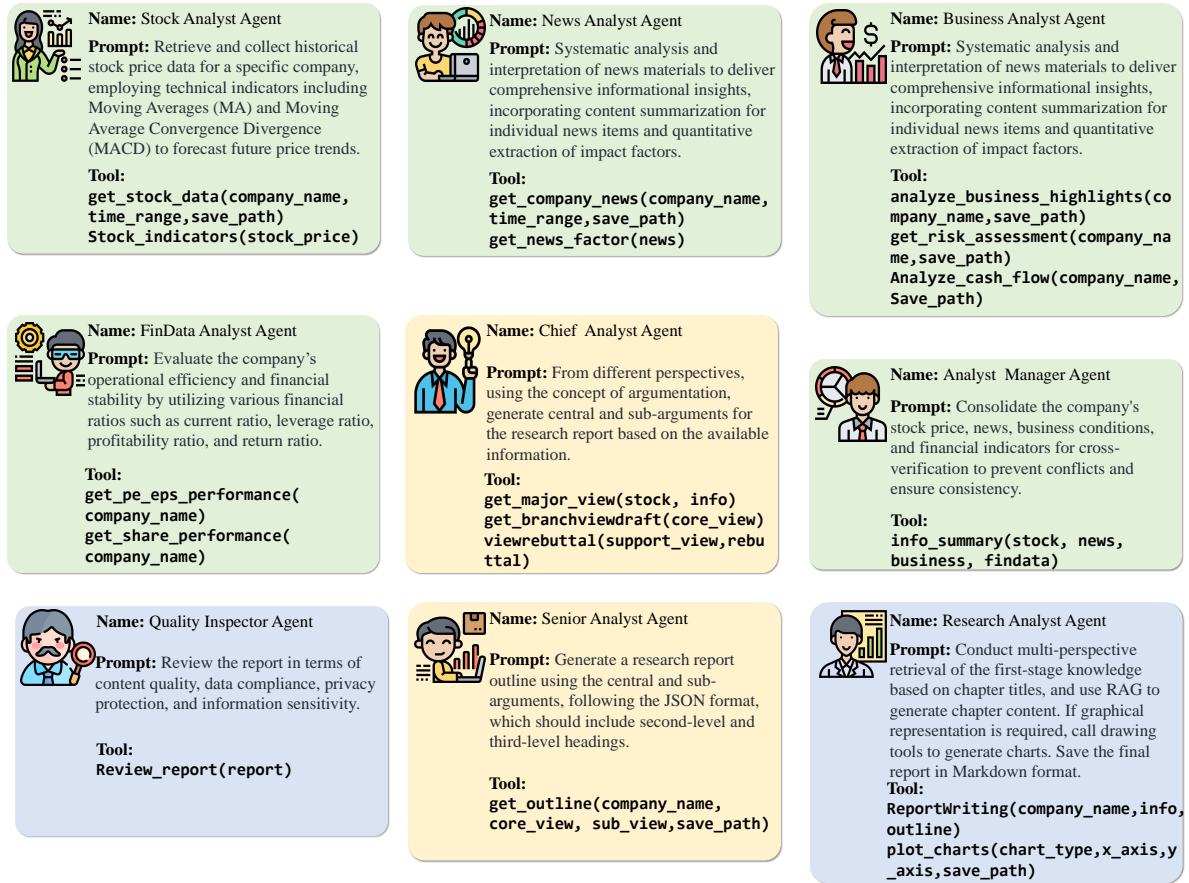


Figure 5: Detailed description of each agent in FinRGAgtns

## A Dataset Description

The basic information of the FinRG dataset is shown in Table 6.

## B Agent Role Specifications

The description of the agent in FinRGAgtns is depicted in the figures 5.

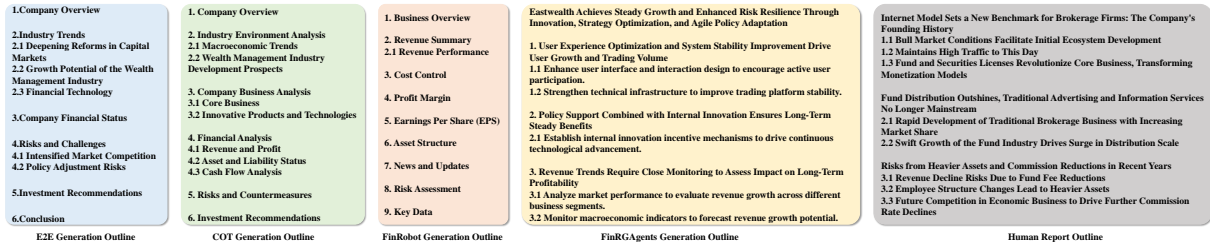


Figure 6: Case study of FinRGAgents and baselines on the generated financial research report outline for Eastwealth.

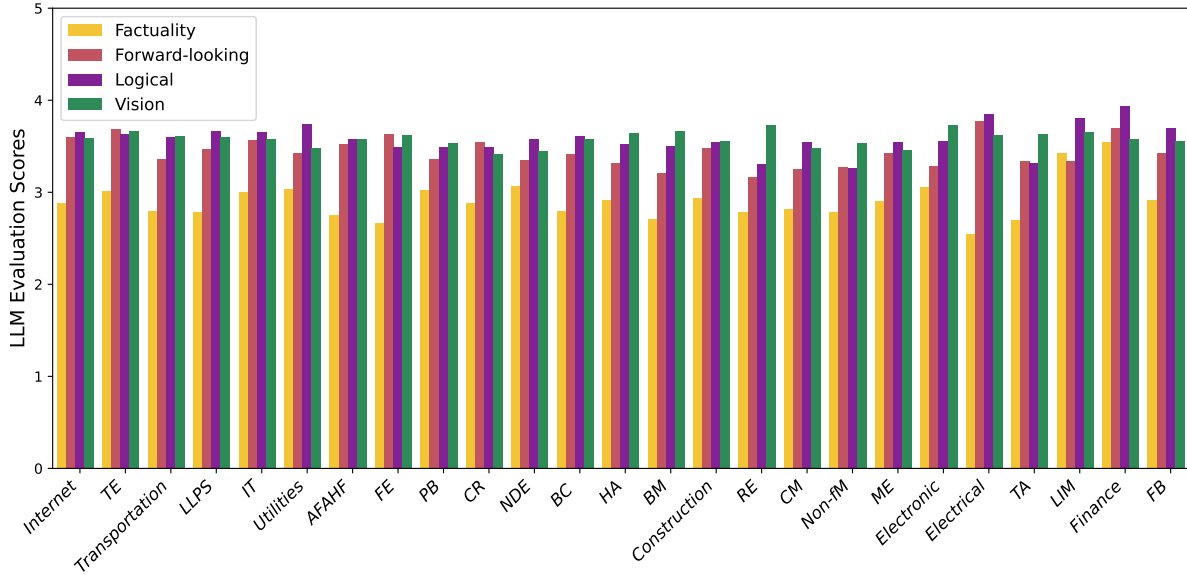


Figure 7: Evaluation results of 25 domains on the Fin2RG datasets.

## C In-depth Analysis

### C.1 Case Study

In Figure 6, we present the results of financial report outlines generated by FinRGAgents and baselines for Eastwealth Company. The outline on the far right represents the real report outline. We observe that the outlines generated by E2E, CoT, and FinRobot are relatively templated and fail to reflect content specific to Eastwealth. In contrast, the outline generated by FinRGAgents is more aligned with the real report outline. It first organizes a large volume of data through data analysts, after which the chief analyst extracts the central view from the processed data. This specialized division of labor enables the creation of more realistic and actionable financial report content.

## D Fine-Grained Evaluation Results

## E Prompts for LLM Evaluation

The prompts used in LLM Evaluation are depicted in Figures 8, 9, 10, and 11.



[1.00,1.49]-The text contains a significant amount of unverified information or major errors, with unclear or highly unreliable sources. Most factual statements lack evidence. It includes notably misleading content or false information, making it highly inaccurate.

[1.50,2.49]-The text has notable factual inaccuracies, with vague or unreliable sources. Some statements may be partially accurate, but the content often exhibits bias or significant gaps in background information, limiting its reliability.

[2.50,3.49]-The text demonstrates basic factual accuracy with some important inaccuracies or omissions. Some sources are reliable, while others are secondary or questionable. Overall, it provides reasonable coverage of major facts but lacks precision in details.

[3.50-4.49]-Most information in the text is factually accurate and supported by reliable sources. While a few details might lack verification or thoroughness, the content is consistent with known data and provides a solid background, enhancing its credibility.

[4.50-5.00]-All factual statements in the text have been rigorously verified and exhibit high accuracy. Data and information sources are clear and authoritative, fully supporting the content. The text provides a detailed and precise presentation of facts, making it fully trustworthy.

Output Format:

**Factuality Score:** [Insert Score Here] (Range: 1.00 to 5.00, rounded to two decimal places)

**Justification:** [Provide a brief explanation for the score based on the content's accuracy, source reliability, and evidence to support the statements. Highlight any significant issues that were noted which affected the score.]

Figure 8: Illustration of the prompt used to evaluate factuality in Fin2RGEval.

[1.00,1.49]-The report offers minimal or no forward-looking analysis. Predictions are speculative and lack credibility, with little to no consideration of macroeconomic, industry, or company-specific factors. Future projections are not actionable or reliable.

[1.50,2.49]-The report includes forward-looking analysis but is underdeveloped and lacks robustness. Predictions rely on limited data and weak or oversimplified assumptions. There is insufficient attention to macroeconomic trends, industry shifts, or strategic initiatives, and risk factors are minimally addressed.

[2.50,3.49]-The report provides a basic level of forward-looking analysis, considering some relevant macroeconomic and industry trends. Predictions may include key financial metrics and strategic factors but lack depth. Scenario and sensitivity analyses are minimal, and risk factors are identified but not thoroughly examined.

[3.50,4.49]-The forward-looking analysis is detailed and based on reasonable assumptions with sound data. It considers significant macroeconomic and industry trends, company strategies, and includes detailed financial forecasts. Risks are recognized and analyzed, with scenario analysis enhancing reliability and usability.

[4.50,5.00]-The report provides comprehensive and insightful forward-looking analysis, supported by thorough research and robust assumptions. It integrates extensive macroeconomic, industry, and company-specific factors, delivering in-depth financial forecasts and strategic insights. Risks are systematically addressed with advanced scenario and sensitivity analyses, resulting in highly reliable and actionable future projections.

Output Format:

**Forward-looking Score:** [Insert Score Here] (Range: 1.00 to 5.00, rounded to two decimal places)

**Justification:** [Provide a brief explanation for the score based on the level of detail, robustness, and credibility of future predictions, incorporating macroeconomic, industry, and company-specific factors. Point out any significant weaknesses or strengths.]

Figure 9: Illustration of the prompt used to evaluate forward-looking in Fin2RGEval.

[1.00,1.49]-The report lacks a coherent structure, with arguments that are disorganized or unclear. Logical fallacies are prevalent, and many conclusions are unsupported. Connections between data and conclusions are either weak or entirely absent, resulting in a fragmented and hard-to-follow analysis.

[1.50,2.49]-The report shows a basic attempt at organization but is still poorly structured. Arguments are weak, and some conclusions rely on inadequate evidence or overly simplistic assumptions. Logical connections are present but tenuous, and the analysis lacks depth and thorough reasoning.

[2.50,3.49]-The report demonstrates a generally clear structure with a reasonable flow of arguments. Most conclusions are supported by data, but the reasoning could lack depth or miss certain complexities. While logical connections exist, they are not consistently strong, and some sections might benefit from more clarity or rigor.

[3.50,4.49]-The report is well-organized with clear and well-founded arguments. Logical progression from data to conclusions is evident, with assumptions and reasoning clearly explained. Most conclusions are well-supported by evidence, though minor areas may require more depth or detail.

[4.50,5.00]-The report exhibits exceptional logical coherence and a compelling narrative. Arguments are sophisticated, well-structured, and deeply reasoned. Conclusions are robustly supported by comprehensive data and demonstrate a nuanced understanding of complex interrelationships. The analysis is both insightful and highly persuasive.

Output Format:

**Logical Score:** [Insert Score Here] (Range: 1.00 to 5.00, rounded to two decimal places)

**Justification:** [Provide a rationale for the score selected, detailing how the report's arguments and structure align with logical principles. Mention any significant logical fallacies or strengths in reasoning.]

Figure 10: Illustration of the prompt used to evaluate logical in Fin2RGEval.

[1.00, 1.49] - The text content and the chart presentation are significantly inconsistent. There is no clear connection between the charts and the data points or conclusions discussed, or the information in the charts directly contradicts the text description.

[1.50,2.49]-There is some connection between the text content and the charts, but it is weak. Some of the data in the charts do not fully align with the text description, or the explanations are insufficient, showing considerable discrepancies.

[2.50,3.49]-The text content and the charts are generally consistent. Most data and conclusions are supported by the charts, but the chart details may not fully showcase the complexity or depth discussed in the text.

[3.50,4.49]-The text content and the charts correspond well. Charts clearly display the key data and conclusions of the text section, with only minor details not fully corresponding or needing clearer presentation.

[4.50,5.00]-The text content and the chart presentation are perfectly consistent. Charts precisely and thoroughly reflect all the key points and complex data relationships described in the text, enhancing the persuasiveness of the report.

Output Format:

Consistency Score: [Insert Score Here] (Range: 1.00 to 5.00, rounded to two decimal places)

Justification: [Provide a rationale for the score selected, detailing how the text content aligns with the chart presentation. Mention any significant inconsistencies or strengths in consistency.]

Figure 11: Illustration of the prompt used to evaluate vision in Fin2RGEval.