

text

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

text

Anonymous ACL submission

text

Abstract

text

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

text

1 Introduction

text

030 Multimodal financial research report generation
031 (MM-FinRG) is a complex task that requires close
032 collaboration among team members, each possessing
033 distinct professional skills (Bennett and Gadlin,
034 2012). However, these reports are typically
035 produced manually by a few seasoned experts. This
036 process not only consumes significant human
037 resources but also incurs substantial intellectual
038 costs. Such reports aim to provide investors with
039 strategic guidance, risk assessment, and revenue
040 forecasts, all crucial for making informed
041 investment decisions (Baker and Nofsinger, 2010).
042 Given the inherent complexity in the creation of
financial research

... #0 (57..)

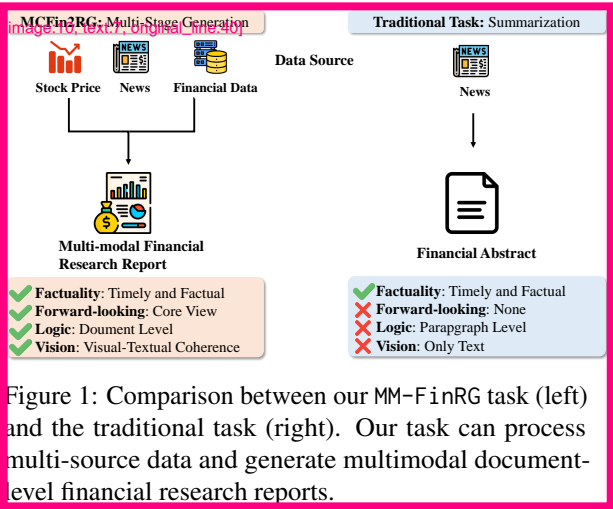


Figure 1: Comparison between our MM-FinRG task (left) and the traditional task (right). Our task can process multi-source data and generate multimodal document-level financial research reports.

text

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-

text

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-

Qwen: #1.00).

line #1	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
line #3	FinRG(ours)	MM-FinRG	Multimodal	Multimodal	2054	5723	25

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.

line #9	Industry	News	Reports	Stock	Time Span
line #14	Internet (互联网)	3,757	1,512	1,800	2023.01-2024.01
line #19	Transportation (交通运输)	3,708	2,012	1,492	2023.01-2024.04
line #24	Finance (金融)	4,172	2,515	1,462	2023.01-2024.04
line #29	Utilities (公用事业)	2,142	3,019	1,101	2023.01-2024.06
line #34	Construction (建筑)	3,612	3,573	1,297	2023.02-2024.07
line #39	Food and Beverage (食品饮料)	3,591	4,013	1,326	2023.04-2024.07
line #44	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-

text

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 Eastmoney¹, 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:

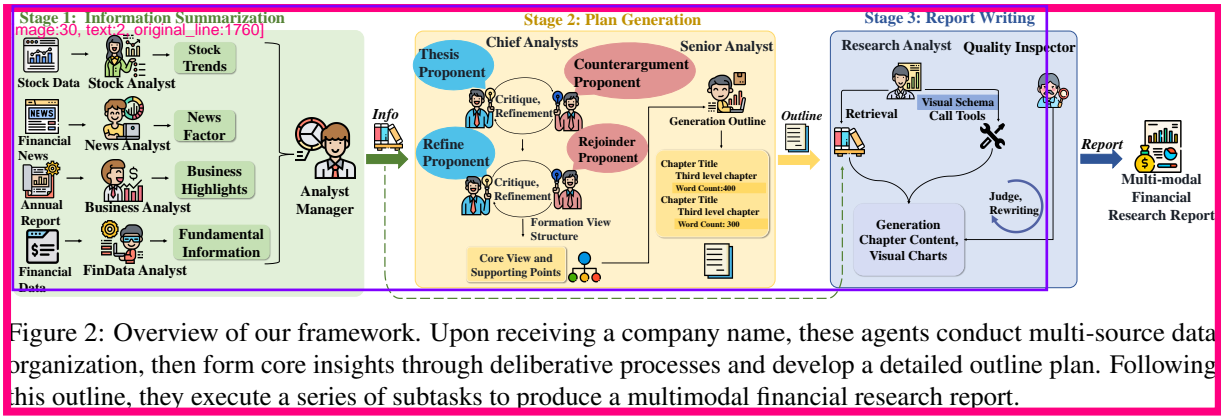
- News Data Collection: Collect news data related to the company from the EastMoney information page² for the past month, including news titles and main content. Implement preprocessing steps to filter out articles containing fewer than 20 Chinese characters and subsequently perform deduplication based 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 Doc2x³ parsing tool to convert the annual report into an md document, preserving table structures and organizing the content by sections.

¹<https://www.eastmoney.com/>

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

³<https://doc2x.noedgeai.com/>

...Given7920.)



- 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:

$$\text{Schema} \triangleq \begin{cases} \text{x_axis} \in \mathbb{R}^n & \text{s.t. } x_1 \leq \dots \leq x_n \\ \text{y_axis} \in \mathbb{R}^n \\ \text{chart_type} \in \{\text{line, pie, bar}\} \end{cases} \quad (5)$$

where \mathbb{R}^n denotes the real numbers, 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-

text
 396 porting views in two sequential steps based on pre-
 397 viously summarized information.

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

text
 399 where P and $Info$ represent the writing prompt
 400 and information summarization knowledge, respec-
 401 tively; The symbol \mapsto indicates the action of
 402 prompting the Agent with a specific prompt to gen-
 403 erate desired responses, while v^m and v_n denote
 404 the draft of the major view and n supporting views.

405 **Counterargument** Subsequently, the counterar-
 406 gument proponent agent generates an overriding
 407 rebuttal to challenge v_n . Then, conditioned on the
 408 rebuttal, agents are required to optimize v_n :

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

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

414 **Rejoinder** In financial reporting, a refined view
 415 addresses this weakness, yet gaps remain, poten-
 416 tially involving the questioning of the validity of
 417 the view assumption. In addressing such issues,
 418 the Rejoinder Proponent Agent first generates an
 419 undercutting rebuttal to optimize \bar{v}_n :

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

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

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

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

428 5 Experiments

429 5.1 Baselines

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

text
 To verify the effectiveness of our framework,
 we compare it with the following zero-shot base-
 lines 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 Sys-
 tem: This system involves two agents; one is re-
 sponsible 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 focus-
 ing 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

text
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.

469 **LLM Evaluation** Evaluating multimodal finan-
 470 cial reports remains challenging, particularly in
 471 holistic assessment. Given the scarcity of bench-
 472 mark resources, traditional text generation met-
 473 rics (Sellam et al., 2020) are insufficient for compre-
 474 hensive multimodal financial report evaluation. As
 475 an initial strategy, we apply four fundamental and
 476 objective dimensions that reflect different aspects
 477 of financial report quality to evaluate the agent-
 478 generated reports. We then integrate these dimen-
 479 sions to facilitate a more holistic evaluation. The
 480 detailed scoring criteria provided in Appendix E.

- text
- **Factuality** measures the accuracy and reliabil-
 ity 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 qual-
 ity of predictive analysis in a report, quantified by

Overall #0

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	3.14	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		0.352	216.46	0.872	2.88	3.60	3.65	3.62	3.44
Ablation									
w/o Information		0.285	296.21	0.734	2.43	3.55	3.46	3.45	3.22
w/o Core View		0.278	305.47	0.741	2.64	3.46	3.38	3.52	3.25
w/o Visual Schema		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.

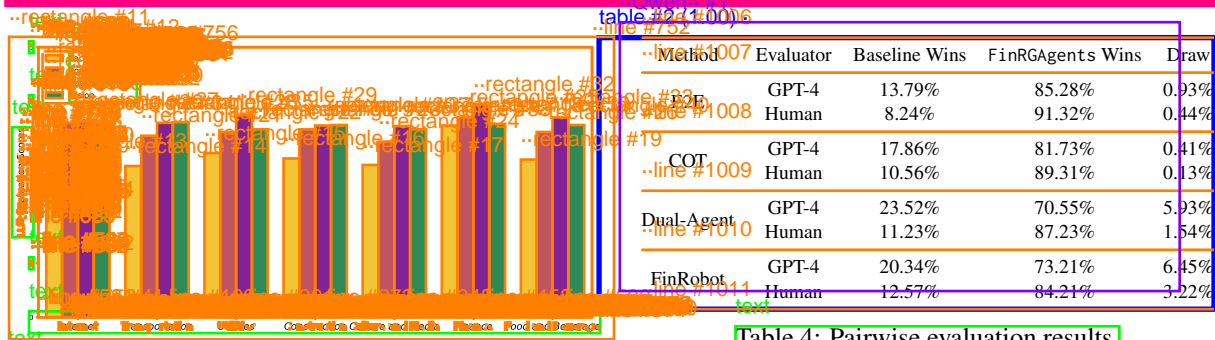


Table 4: Pairwise evaluation results.

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:

text

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

text

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

Correlation	Factuality	Forward-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

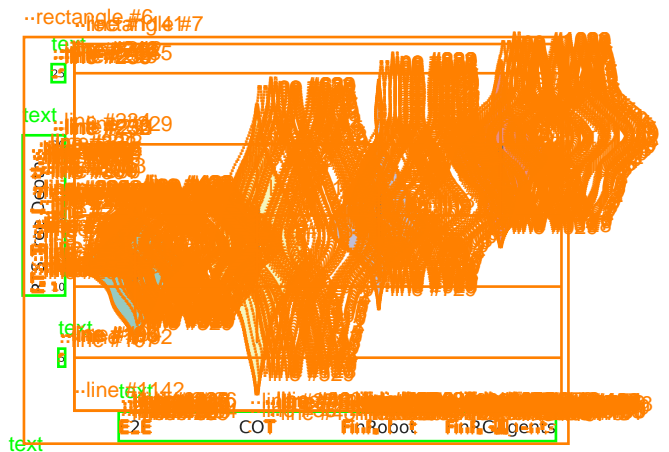


Figure 4: Distribution of RST tree depth of generation financial reports.

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.

text

515 Limitations

text

516 While our approach demonstrates its capability to
517 produce high-quality financial reports, it still faces
518 inherent challenges that affect its general applica-
519 bility. For example, achieving SOTA performance
520 on our dataset is impressive but not absolute, which
521 may limit its broader application. Additionally, the
522 data sources are not fully comprehensive—for in-
523 stance, ECC audio and company announcements
524 are not fully included—which could result in sub-
525 optimal report generation outcomes. Furthermore,
526 although FinRGAgents shows significant improve-
527 ments in chart generation compared to previous
528 methods, it still lacks the ability to fully leverage
529 visual information to enhance chart consistency.
530 This is often reflected in issues such as incorrect
531 elements in charts or errors in axis unit labeling.
532 Addressing these limitations will require future en-
533 hancements, including better integration of visual
534 information into the generation process.

text

535 References

text

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

541 H Kent Baker and John R Nofsinger. 2010. [Behavioral
542 finance: investors, corporations, and markets](#), vol-
543 ume 6. John Wiley & Sons.

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

551 L Michelle Bennett and Howard Gadlin. 2012. Collabo-
552 ration and team science: from theory to practice.

553 Zhihong Chen, Yaling Shen, Yan Song, and Xi-
554 ang Wan. 2021. Cross-modal memory net-
555 works for radiology report generation. In [Proceedings of the 59th Annual Meeting of the
556 Association for Computational Linguistics and the
557 11th International Joint Conference on Natural
558 Language Processing \(Volume 1: Long Papers\)](#),
559 pages 5904–5914, Online. Association for Computa-
560 tional Linguistics.

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

text

Toronto, Canada. Association for Computational Lin- 667
guistics. 668

text

Yujuan Ding, Yunshan Ma, Wenqi Fan, Yige Yao, Tat- 669
Seng Chua, and Qing Li. 2024. [Fashionregen: Llm-
513 empowered fashion report generation](#). Companion
514 [Proceedings of the ACM on Web Conference 2024](#). 670
671 672

text

James B Freeman. 2011. [Argument Structure: 129
Representation and Theory](#), volume 18. Springer
513 Science & Business Media. 673
674 675

text

R Edward Freeman. 2010. [Stakeholder theory: The 127
state of the art](#). Cambridge University Press. 676
677

text

Yuhang He, Jianzhu Bao, Yang Sun, Bin Liang, Min 678
Yang, Bing Qin, and Ruifeng Xu. 2024. [Decom-
679 posing argumentative essay generation via dialectical
680 planning of complex reasoning](#). In [Findings of
681 the Association for Computational Linguistics ACL
682 2024](#), pages 12305–12322, Bangkok, Thailand and
683 virtual meeting. Association for Computational Lin-
684 guistics. 685

text

Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. 686
2024. Unlocking varied perspectives: A persona-
687 based multi-agent framework with debate-driven text
688 planning for argument generation. [arXiv preprint
689 arXiv:2406.19643](#). 690

text

Alistair EW Johnson, Tom J Pollard, Nathaniel R Green- 691
baum, Matthew P Lungren, Chih-ying Deng, Yifan
692 Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz,
693 and Steven Horng. 2019. Mimic-cxr-jpg, a large pub-
694 licly available database of labeled chest radiographs.
695 [arXiv preprint arXiv:1901.07042](#). 696

text

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii 697
Khizbullin, and Bernard Ghanem. 2023a. Camel:
698 Communicative agents for "mind" exploration of
699 large language model society. [Advances in Neural
700 Information Processing Systems](#), 36:51991–52008.
701

text

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, 702
and Bill Dolan. 2015. A diversity-promoting objec-
703 tive function for neural conversation models. [arXiv
704 preprint arXiv:1510.03055](#). 705

text

Yang Li, Yangyang Yu, Haohang Li, Z. Chen, and Khal- 706
doun Khashanah. 2023b. [Tradinggpt: Multi-agent
707 system with layered memory and distinct characters
708 for enhanced financial trading performance](#). 709

text

Chin-Yew Lin. 2004. [ROUGE: A package for automatic
710 evaluation of summaries](#). In [Text Summarization
711 Branches Out](#), pages 74–81, Barcelona, Spain. Asso-
712 ciation for Computational Linguistics. 713

text

Yang Liu, Dan Iter, Yichong Xu, Shuhang Wang, 714
Ruochen Xu, and Chenguang Zhu. 2023. [G-eval:
715 NLG evaluation using gpt-4 with better human align-
716 ment](#). In [Proceedings of the 2023 Conference on
717 Empirical Methods in Natural Language Processing](#),
718 pages 2511–2522, Singapore. Association for Com-
719 putational Linguistics. 720

text

text

721 Bo Ni and Markus J Buehler. 2024. Mechagents:
722 Large language model multi-agent collaborations can
723 solve mechanics problems, generate new data, and
724 integrate knowledge. [Extreme Mechanics Letters](#),
725 67:102131.

text

726 Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yu-
727 fan Dang, Jiahao Li, Cheng Yang, Weize Chen,
728 Yusheng Su, Xin Cong, et al. 2024. Chatdev: Com-
729 municative agents for software development. In
730 [Proceedings of the 62nd Annual Meeting of the
731 Association for Computational Linguistics \(Volume
732 1: Long Papers\)](#), pages 15174–15186.

text

733 D Gordon Rohman. 1965. Pre-writing: The stage of dis-
734 covery in the writing process. [College Composition
735 & Communication](#), 16(2):106–112.

text

736 Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020.
737 BLEURT: Learning robust metrics for text genera-
738 tion. In [Proceedings of the 58th Annual Meeting
739 of the Association for Computational Linguistics](#),
740 pages 7881–7892, Online. Association for Computa-
741 tional Linguistics.

text

742 Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu,
743 Omar Khattab, and Monica Lam. 2024. Assisting
744 in writing Wikipedia-like articles from scratch with
745 large language models. In [Proceedings of the 2024
746 Conference of the North American Chapter of the
747 Association for Computational Linguistics: Human
748 Language Technologies \(Volume 1: Long Papers\)](#),
749 pages 6252–6278, Mexico City, Mexico. Association
750 for Computational Linguistics.

text

751 Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Sai-
752 ful Bari, Mizanur Rahman, Mohammad Abdullah
753 Matin Khan, Haidar Khan, Israt Jahan, Amran
754 Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, et al.
755 2024. A systematic survey and critical review on
756 evaluating large language models: Challenges, limi-
757 tations, and recommendations. [arXiv e-prints](#), pages
758 arXiv:2407.

text

759 Wen-Kwang Tsao. 2023. Multi-agent reasoning
760 with large language models for effective corpo-
761 rate planning. [2023 International Conference
762 on Computational Science and Computational
763 Intelligence \(CSCI\)](#), pages 365–370.

text

764 Xiangpeng Wan, Haicheng Deng, Kai Zou, and Shiqi
765 Xu. 2024. Enhancing the efficiency and accuracy
766 of underlying asset reviews in structured finance:
767 The application of multi-agent framework. [arXiv](#),
768 abs/2405.04294.

text

769 Siyuan Wang, Bo Peng, Yichao Liu, and Qi Peng.
770 2023. Fine-grained medical vision-language repre-
771 sentation learning for radiology report generation. In
772 [Proceedings of the 2023 Conference on Empirical
773 Methods in Natural Language Processing](#), pages
774 15949–15956, Singapore. Association for Compu-
775 tational Linguistics.

text

776 Ziao Wang, Yunpeng Ren, Xiaofeng Zhang, and Yiyuan
777 Wang. 2024. Generating long financial report us-
778 ing conditional variational autoencoders with knowl-
779 edge distillation. [IEEE Transactions on Artificial
780 Intelligence](#).

text

781 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten
782 Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou,
783 et al. 2022. Chain-of-thought prompting elicits rea-
784 soning in large language models. [Advances in neural
785 information processing systems](#), 35:24824–24837.

text

786 Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu,
787 Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang,
788 Xiaoyun Zhang, and Chi Wang. 2023. Auto-
789 gen: Enabling next-gen llm applications via multi-
790 agent conversation framework. [arXiv preprint
791 arXiv:2308.08155](#).

text

792 Frank Xing. 2024. Designing heterogeneous llm
793 agents for financial sentiment analysis. [arXiv](#),
794 abs/2401.05799.

text

795 Sixing Yan. 2022. Disentangled variational topic
796 inference for topic-accurate financial report gen-
797 eration. In [Proceedings of the Fourth Workshop
798 on Financial Technology and Natural Language
799 Processing \(FinNLP\)](#), pages 18–24, Abu Dhabi,
800 United Arab Emirates (Hybrid). Association for Com-
801 putational Linguistics.

text

802 Hongyang Yang, Boyu Zhang, Neng Wang, Cheng Guo,
803 Xiaoli Zhang, Likun Lin, Junlin Wang, Tianyu Zhou,
804 Mao Guan, Runjia Zhang, et al. 2024. Finrobot: An
805 open-source ai agent platform for financial applica-
806 tions using large language models. [arXiv preprint
807 arXiv:2405.14767](#).

text

808 Yangyang Yu, Zhiyuan Yao, Haohang Li, Zhiyang Deng,
809 Yupeng Cao, Zhi Chen, Jordan W Suchow, Rong Liu,
810 Zhenyu Cui, Denghui Zhang, et al. 2024. Fincon: A
811 synthesized llm multi-agent system with conceptual
812 verbal reinforcement for enhanced financial decision
813 making. [arXiv preprint arXiv:2407.06567](#).

text

814 Hao Zheng, Xinyan Guan, Hao Kong, Jia Zheng,
815 Hongyu Lin, Yaojie Lu, Ben He, Xianpei Han, and
816 Le Sun. 2025. Pptagent: Generating and evaluating
817 presentations beyond text-to-slides. [arXiv preprint
818 arXiv:2501.03936](#).

text

819 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan
820 Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,
821 Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023.
822 Judging llm-as-a-judge with mt-bench and chatbot
823 arena. [Advances in Neural Information Processing
824 Systems](#), 36:46595–46623.

text










825 Shuyan Zhou, Frank F Xu, Hao Zhu, Xuhui Zhou,
826 Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue
827 Ou, Yonatan Bisk, Daniel Fried, et al. 2023. We-
828 barena: A realistic web environment for building au-
829 tonomous agents. [arXiv preprint arXiv:2307.13854](#).

text

... #0 (93..)

Industry	News #	Reports #	Stock #	Time Span	Industry	News #	Reports #	Stock #	Time Span
Internet(互联网)	1,757	1,224	1,890	2023.01-2024.01	TE(交运设备)	1,058	1,233	1,250	2023.01-2024.01
Transportation(交通运输)	1,798	1,734	1,492	2023.01-2024.04	LLPS(生活及专业服务)	1,072	1,085	1,674	2023.01-2024.05
IT(信息技术)	1,488	1,744	1,734	2023.01-2024.01	Utilities(公用事业)	1,142	1,084	1,190	2023.01-2024.06
AFAHF(农林牧渔)	1,358	1,454	1,342	2023.01-2024.06	FE(石化能源)	1,452	1,833	1,573	2023.01-2024.03
PB(医药生物)	1,434	1,641	1,384	2023.01-2024.03	CR(商贸零售)	1,683	1,669	1,263	2023.01-2024.01
NDE(国防与装备)	1,525	1,741	1,153	2023.01-2024.01	BC(基础化工)	1,782	2,113	1,283	2023.01-2024.03
HA(家电)	1,326	1,841	1,986	2023.01-2024.03	BM(建材)	1,252	1,239	1,452	2023.01-2024.02
Construction(建筑)	1,612	1,838	1,297	2023.01-2024.07	RE(房地产)	1,223	1,499	1,183	2023.01-2024.02
CM(文化传媒)	1,415	1,088	1,146	2023.01-2024.08	Non-fM(有色金属)	1,292	1,088	1,101	2023.01-2024.01
ME(机械设备)	1,973	1,186	1,236	2023.01-2024.02	Electronic(电子设备)	1,423	1,186	1,196	2023.01-2024.02
Electrical(电气设备)	1,254	1,284	1,634	2023.01-2024.03	TA(纺织服装)	1,245	1,286	1,126	2023.01-2024.06
LIM(轻工制造)	1,123	1,388	1,972	2023.01-2024.04	Fiance(金融)	1,172	1,386	1,136	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.

 <p>Name: Stock Analyst Agent Prompt: Retrieve and collect historical stock price data for a specific company, employing technical indicators including Moving Averages (MA) and Moving Average Convergence Divergence (MACD) to forecast future price trends. Tool: get_stock_data(company_name, time_range, save_path) Stock_indicators(stock_price)</p>	 <p>Name: News Analyst Agent Prompt: Systematic analysis and interpretation of news materials to deliver comprehensive informational insights, incorporating content summarization for individual news items and quantitative extraction of impact factors. Tool: get_company_news(company_name, time_range, save_path) get_news_factor(news)</p>	 <p>Name: Business Analyst Agent Prompt: Systematic analysis and interpretation of news materials to deliver comprehensive informational insights, incorporating content summarization for individual news items and quantitative extraction of impact factors. Tool: analyze_business_highlights(company_name, save_path) get_risk_assessment(company_name, save_path) Analyze_cash_flow(company_name, Save_path)</p>
 <p>Name: FinData Analyst Agent Prompt: Evaluate the company's operational efficiency and financial stability by utilizing various financial ratios such as current ratio, leverage ratio, profitability ratio, and return ratio. Tool: get_pe_eps_performance(company_name) get_share_performance(company_name)</p>	 <p>Name: Chief Analyst Agent Prompt: From different perspectives, using the concept of argumentation, generate central and sub-arguments for the research report based on the available information. Tool: get_major_view(stock, info) get_branchviewdraft(core_view) viewrebuttal(support_view, rebuttal)</p>	 <p>Name: Analyst Manager Agent Prompt: Consolidate the company's stock price, news, business conditions, and financial indicators for cross-verification to prevent conflicts and ensure consistency. Tool: info_summary(stock, news, business, findata)</p>
 <p>Name: Quality Inspector Agent Prompt: Review the report in terms of content quality, data compliance, privacy protection, and information sensitivity. Tool: Review_report(report)</p>	 <p>Name: Senior Analyst Agent Prompt: Generate a research report outline using the central and sub-arguments, following the JSON format, which should include second-level and third-level headings. Tool: get_outline(company_name, core_view, sub_view, save_path)</p>	 <p>Name: Research Analyst Agent Prompt: Conduct multi-perspective retrieval of the first-stage knowledge based on chapter titles, and use RAG to generate chapter content. If graphical representation is required, call drawing tools to generate charts. Save the final report in Markdown format. Tool: ReportWriting(company_name, info, outline) plot_charts(chart_type, x_axis, y_axis, save_path)</p>

text

Figure 5: Detailed description of each agent in FinRGAgtns

text

A Dataset Description

830

text

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

831

text

B Agent Role Specifications

832

text

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

833

text

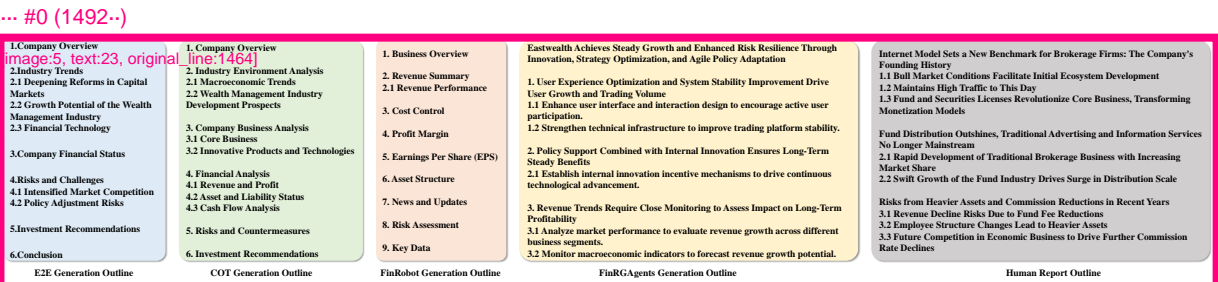


Figure 6: Case study of FinRGAgents and baselines on the generated financial research report outline for Eastwealth.

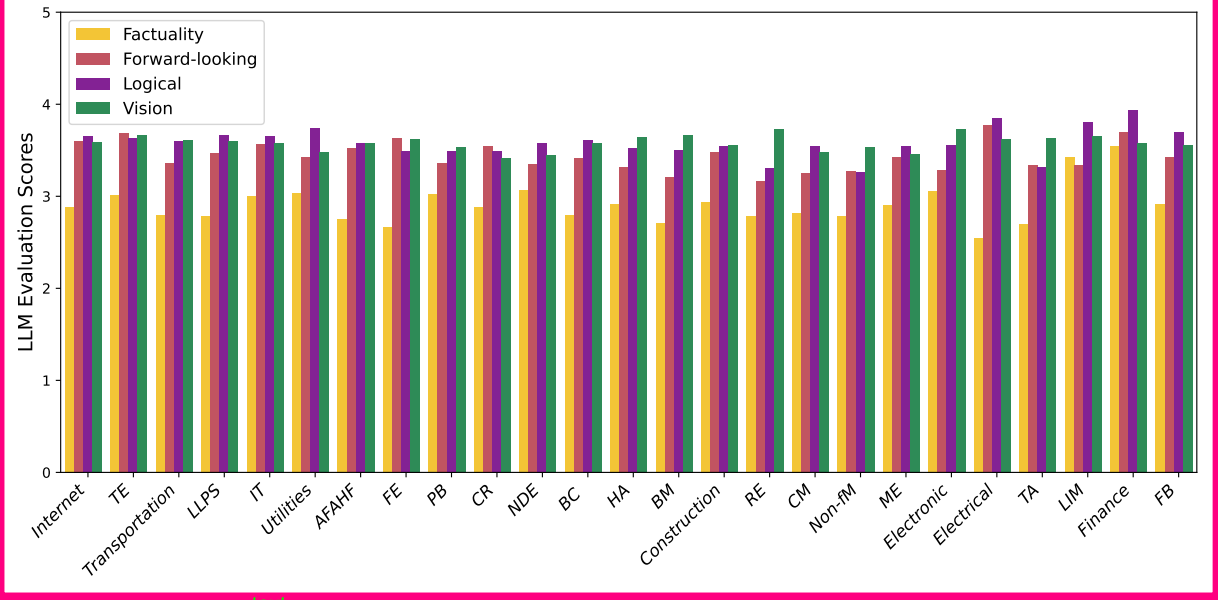


Figure 7: Evaluation results of 25 domains on the Fin2RG datasets.

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.

rectangle #0
line #1
line #2
line #3
line #4

[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.]

text

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

rectangle #5
line #5
line #6
line #7
line #8
line #9
line #10
line #11

[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.]

text

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

rectangle #10
line #10
line #11
line #12
line #13
line #14
line #15
line #16
line #17
line #18
line #19
line #20
line #21
line #22
line #23
line #24
line #25
line #26
line #27
line #28
line #29
line #30
line #31
line #32
line #33
line #34
line #35
line #36
line #37
line #38
line #39
line #40
line #41
line #42
line #43
line #44
line #45
line #46
line #47
line #48
line #49
line #50
line #51
line #52
line #53
line #54
line #55
line #56
line #57
line #58
line #59
line #60
line #61
line #62
line #63
line #64
line #65
line #66
line #67
line #68
line #69
line #70
line #71
line #72
line #73
line #74
line #75
line #76
line #77
line #78
line #79
line #80
line #81
line #82
line #83
line #84
line #85
line #86
line #87
line #88
line #89
line #90
line #91
line #92
line #93
line #94
line #95
line #96
line #97
line #98
line #99
line #100
line #101
line #102
line #103
line #104
line #105
line #106
line #107
line #108
line #109
line #110
line #111
line #112
line #113
line #114
line #115
line #116
line #117
line #118
line #119
line #120
line #121
line #122
line #123
line #124
line #125
line #126
line #127
line #128
line #129
line #130
line #131
line #132
line #133
line #134
line #135
line #136
line #137
line #138
line #139
line #140
line #141
line #142
line #143
line #144
line #145
line #146
line #147
line #148
line #149
line #150
line #151
line #152
line #153
line #154
line #155
line #156
line #157
line #158
line #159
line #160
line #161
line #162
line #163
line #164
line #165
line #166
line #167
line #168
line #169
line #170
line #171
line #172
line #173
line #174
line #175
line #176
line #177
line #178
line #179
line #180
line #181
line #182
line #183
line #184
line #185
line #186
line #187
line #188
line #189
line #190
line #191
line #192
line #193
line #194
line #195
line #196
line #197
line #198
line #199
line #200
line #201
line #202
line #203
line #204
line #205
line #206
line #207
line #208
line #209
line #210
line #211
line #212
line #213
line #214
line #215
line #216
line #217
line #218
line #219
line #220
line #221
line #222
line #223
line #224
line #225
line #226
line #227
line #228
line #229
line #230
line #231
line #232
line #233
line #234
line #235
line #236
line #237
line #238
line #239
line #240
line #241
line #242
line #243
line #244
line #245
line #246
line #247
line #248
line #249
line #250
line #251
line #252
line #253
line #254
line #255
line #256
line #257
line #258
line #259
line #260
line #261
line #262
line #263
line #264
line #265
line #266
line #267
line #268
line #269
line #270
line #271
line #272
line #273
line #274
line #275
line #276
line #277
line #278
line #279
line #280
line #281
line #282
line #283
line #284
line #285
line #286
line #287
line #288
line #289
line #290
line #291
line #292
line #293
line #294
line #295
line #296
line #297
line #298
line #299
line #300
line #301
line #302
line #303
line #304
line #305
line #306
line #307
line #308
line #309
line #310
line #311
line #312
line #313
line #314
line #315
line #316
line #317
line #318
line #319
line #320
line #321
line #322
line #323
line #324
line #325
line #326
line #327
line #328
line #329
line #330
line #331
line #332
line #333
line #334
line #335
line #336
line #337
line #338
line #339
line #340
line #341
line #342
line #343
line #344
line #345
line #346
line #347
line #348
line #349
line #350
line #351
line #352
line #353
line #354
line #355
line #356
line #357
line #358
line #359
line #360
line #361
line #362
line #363
line #364
line #365
line #366
line #367
line #368
line #369
line #370
line #371
line #372
line #373
line #374
line #375
line #376
line #377
line #378
line #379
line #380
line #381
line #382
line #383
line #384
line #385
line #386
line #387
line #388
line #389
line #390
line #391
line #392
line #393
line #394
line #395
line #396
line #397
line #398
line #399
line #400
line #401
line #402
line #403
line #404
line #405
line #406
line #407
line #408
line #409
line #410
line #411
line #412
line #413
line #414
line #415
line #416
line #417
line #418
line #419
line #420
line #421
line #422
line #423
line #424
line #425
line #426
line #427
line #428
line #429
line #430
line #431
line #432
line #433
line #434
line #435
line #436
line #437
line #438
line #439
line #440
line #441
line #442
line #443
line #444
line #445
line #446
line #447
line #448
line #449
line #450
line #451
line #452
line #453
line #454
line #455
line #456
line #457
line #458
line #459
line #460
line #461
line #462
line #463
line #464
line #465
line #466
line #467
line #468
line #469
line #470
line #471
line #472
line #473
line #474
line #475
line #476
line #477
line #478
line #479
line #480
line #481
line #482
line #483
line #484
line #485
line #486
line #487
line #488
line #489
line #490
line #491
line #492
line #493
line #494
line #495
line #496
line #497
line #498
line #499
line #500
line #501
line #502
line #503
line #504
line #505
line #506
line #507
line #508
line #509
line #510
line #511
line #512
line #513
line #514
line #515
line #516
line #517
line #518
line #519
line #520
line #521
line #522
line #523
line #524
line #525
line #526
line #527
line #528
line #529
line #530
line #531
line #532
line #533
line #534
line #535
line #536
line #537
line #538
line #539
line #540
line #541
line #542
line #543
line #544
line #545
line #546
line #547
line #548
line #549
line #550
line #551
line #552
line #553
line #554
line #555
line #556
line #557
line #558
line #559
line #560
line #561
line #562
line #563
line #564
line #565
line #566
line #567
line #568
line #569
line #570
line #571
line #572
line #573
line #574
line #575
line #576
line #577
line #578
line #579
line #580
line #581
line #582
line #583
line #584
line #585
line #586
line #587
line #588
line #589
line #590
line #591
line #592
line #593
line #594
line #595
line #596
line #597
line #598
line #599
line #600
line #601
line #602
line #603
line #604
line #605
line #606
line #607
line #608
line #609
line #610
line #611
line #612
line #613
line #614
line #615
line #616
line #617
line #618
line #619
line #620
line #621
line #622
line #623
line #624
line #625
line #626
line #627
line #628
line #629
line #630
line #631
line #632
line #633
line #634
line #635
line #636
line #637
line #638
line #639
line #640
line #641
line #642
line #643
line #644
line #645
line #646
line #647
line #648
line #649
line #650
line #651
line #652
line #653
line #654
line #655
line #656
line #657
line #658
line #659
line #660
line #661
line #662
line #663
line #664
line #665
line #666
line #667
line #668
line #669
line #670
line #671
line #672
line #673
line #674
line #675
line #676
line #677
line #678
line #679
line #680
line #681
line #682
line #683
line #684
line #685
line #686
line #687
line #688
line #689
line #690
line #691
line #692
line #693
line #694
line #695
line #696
line #697
line #698
line #699
line #700
line #701
line #702
line #703
line #704
line #705
line #706
line #707
line #708
line #709
line #710
line #711
line #712
line #713
line #714
line #715
line #716
line #717
line #718
line #719
line #720
line #721
line #722
line #723
line #724
line #725
line #726
line #727
line #728
line #729
line #730
line #731
line #732
line #733
line #734
line #735
line #736
line #737
line #738
line #739
line #740
line #741
line #742
line #743
line #744
line #745
line #746
line #747
line #748
line #749
line #750
line #751
line #752
line #753
line #754
line #755
line #756
line #757
line #758
line #759
line #760
line #761
line #762
line #763
line #764
line #765
line #766
line #767
line #768
line #769
line #770
line #771
line #772
line #773
line #774
line #775
line #776
line #777
line #778
line #779
line #780
line #781
line #782
line #783
line #784
line #785
line #786
line #787
line #788
line #789
line #790
line #791
line #792
line #793
line #794
line #795
line #796
line #797
line #798
line #799
line #800
line #801
line #802
line #803
line #804
line #805
line #806
line #807
line #808
line #809
line #810
line #811
line #812
line #813
line #814
line #815
line #816
line #817
line #818
line #819
line #820
line #821
line #822
line #823
line #824
line #825
line #826
line #827
line #828
line #829
line #830
line #831
line #832
line #833
line #834
line #835
line #836
line #837
line #838
line #839
line #840
line #841
line #842
line #843
line #844
line #845
line #846
line #847
line #848
line #849
line #850
line #851
line #852
line #853
line #854
line #855
line #856
line #857
line #858
line #859
line #860
line #861
line #862
line #863
line #864
line #865
line #866
line #867
line #868
line #869
line #870
line #871
line #872
line #873
line #874
line #875
line #876
line #877
line #878
line #879
line #880
line #881
line #882
line #883
line #884
line #885
line #886
line #887
line #888
line #889
line #890
line #891
line #892
line #893
line #894
line #895
line #896
line #897
line #898
line #899
line #900
line #901
line #902
line #903
line #904
line #905
line #906
line #907
line #908
line #909
line #910
line #911
line #912
line #913
line #914
line #915
line #916
line #917
line #918
line #919
line #920
line #921
line #922
line #923
line #924
line #925
line #926
line #927
line #928
line #929
line #930
line #931
line #932
line #933
line #934
line #935
line #936
line #937
line #938
line #939
line #940
line #941
line #942
line #943
line #944
line #945
line #946
line #947
line #948
line #949
line #950
line #951
line #952
line #953
line #954
line #955
line #956
line #957
line #958
line #959
line #960
line #961
line #962
line #963
line #964
line #965
line #966
line #967
line #968
line #969
line #970
line #971
line #972
line #973
line #974
line #975
line #976
line #977
line #978
line #979
line #980
line #981
line #982
line #983
line #984
line #985
line #986
line #987
line #988
line #989
line #990
line #991
line #992
line #993
line #994
line #995
line #996
line #997
line #998
line #999
line #1000
line #1001
line #1002
line #1003
line #1004
line #1005
line #1006
line #1007
line #1008
line #1009
line #1010
line #1011
line #1012
line #1013
line #1014
line #1015
line #1016
line #1017
line #1018
line #1019
line #1020
line #1021
line #1022
line #1023
line #1024
line #1025
line #1026
line #1027
line #1028
line #1029
line #1030
line #1031
line #1032
line #1033
line #1034
line #1035
line #1036
line #1037
line #1038
line #1039
line #1040
line #1041
line #1042
line #1043
line #1044
line #1045
line #1046
line #1047
line #1048
line #1049
line #1050
line #1051
line #1052
line #1053
line #1054
line #1055
line #1056
line #1057
line #1058
line #1059
line #1060
line #1061
line #1062
line #1063
line #1064
line #1065
line #1066
line #1067
line #1068
line #1069
line #1070
line #1071
line #1072
line #1073
line #1074
line #1075
line #1076
line #1077
line #1078
line #1079
line #1080
line #1081
line #1082
line #1083
line #1084
line #1085
line #1086
line #1087
line #1088
line #1089
line #1090
line #1091
line #1092
line #1093
line #1094
line #1095
line #1096
line #1097
line #1098
line #1099
line #1100
line #1101
line #1102
line #1103
line #1104
line #1105
line #1106
line #1107
line #1108
line #1109
line #1110
line #1111
line #1112
line #1113
line #1114
line #1115
line #1116
line #1117
line #1118
line #1119
line #1120
line #1121
line #1122
line #1123
line #1124
line #1125
line #1126
line #1127
line #1128
line #1129
line #1130
line #1131
line #1132
line #1133
line #1134
line #1135
line #1136
line #1137
line #1138
line #1139
line #1140
line #1141
line #1142
line #1143
line #1144
line #1145
line #1146
line #1147
line #1148
line #1149
line #1150
line #1151
line #1152
line #1153
line #1154
line #1155
line #1156
line #1157
line #1158
line #1159
line #1160
line #1161
line #1162
line #1163
line #1164
line #1165
line #1166
line #1167
line #1168
line #1169
line #1170
line #1171
line #1172
line #1173
line #1174
line #1175
line #1176
line #1177
line #1178
line #1179
line #1180
line #1181
line #1182
line #1183
line #1184
line #1185
line #1186
line #1187
line #1188
line #1189
line #1190
line #1191
line #1192
line #1193
line #1194
line #1195
line #1196
line #1197
line #1198
line #1199
line #1200
line #1201
line #1202
line #1203
line #1204
line #1205
line #1206
line #1207
line #1208
line #1209
line #1210
line #1211
line #1212
line #1213
line #1214
line #1215
line #1216
line #1217
line #1218
line #1219
line #1220
line #1221
line #1222
line #1223
line #1224
line #1225
line #1226
line #1227
line #1228
line #1229
line #1230
line #1231
line #1232
line #1233
line #1234
line #1235
line #1236
line #1237
line #1238
line #1239
line #1240
line #1241
line #1242
line #1243
line #1244
line #1245
line #1246
line #1247
line #1248
line #1249
line #1250
line #1251
line #1252
line #1253
line #1254
line #1255
line #1256
line #1257
line #1258
line #1259
line #1260
line #1261
line #1262
line #1263
line #1264
line #1265
line #1266
line #1267
line #1268
line #1269
line #1270
line #1271
line #1272
line #1273
line #1274
line #1275
line #1276
line #1277
line #1278
line #1279
line #1280
line #1281
line #1282
line #1283
line #1284
line #1285
line #1286
line #1287
line #1288
line #1289
line #1290
line #1291
line #1292
line #1293
line #1294
line #1295
line #1296
line #1297
line #1298
line #1299
line #1300
line #1301
line #1302
line #1303
line #1304
line #1305
line #1306
line #1307
line #1308
line #1309
line #1310
line #1311
line #1312
line #1313
line #1314
line #1315
line #1316
line #1317
line #1318
line #1319
line #1320
line #1321
line #1322
line #1323
line #1324
line #1325
line #1326
line #1327
line #1328
line #1329
line #1330
line #1331
line #1332
line #1333
line #1334
line #1335
line #1336
line #1337
line #1338
line #1339
line #1340
line #1341
line #1342
line #1343
line #1344
line #1345
line #1346
line #1347
line #1348
line #1349
line #1350
line #1351
line #1352
line #1353
line #1354
line #1355
line #1356
line #1357
line #1358
line #1359
line #1360
line #1361
line #1362
line #1363
line #1364
line #1365
line #1366
line #1367
line #1368
line #1369
line #1370
line #1371
line #1372
line #1373
line #1374
line #1375
line #1376
line #1377
line #1378
line #1379
line #1380
line #1381
line #1382
line #1383
line #1384
line #1385
line #1386
line #1387
line #1388
line #1389
line #1390
line #1391
line #1392
line #1393
line #1394
line #1395
line #1396
line #1397
line #1398
line #1399
line #1400
line #1401
line #1402
line #1403
line #1404
line #1405
line #1406
line #1407
line #1408
line #1409
line #1410
line #1411
line #1412
line #1413
line #1414
line #1415
line #1416
line #1417
line #1418
line #1419
line #1420
line #1421
line #1422
line #1423
line #1424
line #1425
line #1426
line #1427
line #1428
line #1429
line #1430
line #1431
line #1432
line #1433
line #1434
line #1435
line #1436
line #1437
line #1438
line #1439
line #1440
line #1441
line #1442
line #1443
line #1444
line #1445
line #1446
line #1447
line #1448
line #1449
line #1450
line #1451
line #1452
line #1453
line #1454
line #1455
line #1456
line #1457
line #1458
line #1459
line #1460
line #1461
line #1462
line #1463
line #1464
line #1465
line #1466
line #1467
line #1468
line #1469
line #1470
line #1471
line #1472
line #1473
line #1474
line #1475
line #1476
line #1477
line #1478
line #1479
line #1480
line #1481
line #1482
line #1483
line #1484
line #1485
line #1486
line #1487
line #1488
line #1489
line #1490
line #1491
line #1492
line #1493
line #1494
line #1495
line #1496
line #1497
line #1498
line #1499
line #1500
line #1501
line #1502
line #1503
line #1504
line #1505
line #1506
line #1507
line #1508
line #1509
line #1510
line #1511
line #1512
line #1513
line #1514
line #1515
line #1516
line #1517
line #1518
line #1519
line #1520
line #1521
line #1522
line #1523
line #1524
line #1525
line #1526
line #1527
line #1528
line #1529
line #1530
line #1531
line #1532
line #1533
line #1534
line #1535
line #1536
line #1537
line #1538
line #1539
line #1540
line #1541
line #1542
line #1543
line #1544
line #1545
line #1546
line #1547
line #1548
line #1549
line #1550
line

rectangle #0
line #2

[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.]

line #3

text

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