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

Anonymous ACL submission

Abstract

Multimodal financial research report generation is challenging because it requires synthesizing heterogeneous evidence—market news, stock indicators, annual reports, and financial metrics—while producing forward-looking analysis under uncertainty and maintaining consistency between textual claims and supporting visualizations. Existing long-form generation methods primarily focus on text-only writing or single-source summarization, and therefore do not directly address this setting. We formulate `MM-FinRG` (Multi-Modal Chinese Financial Research Report Generation), a new task for generating Chinese financial research reports from multi-source market evidence. To tackle this task, we propose `FinRGAgents`, a multi-agent framework that combines dialectical reasoning for forming defensible core views with schema-grounded planning for multimodal report writing. We further construct `FinRG`, a benchmark of 1,500 reports covering 25 financial domains, and introduce a task-specific evaluation protocol spanning factuality, forward-looking reasoning, logical coherence, and vision-text consistency. Experiments show that `FinRGAgents` achieves an average LLM score of 3.53, surpassing the strongest baseline by 6.3%; ablation and human agreement analyses ($r = 0.75$) validate the effectiveness of the proposed design. Our code is available at https://anonymous.4open.science/r/FinRGAgents-B0D3

1 Introduction

Financial research reports are a cornerstone of capital markets, providing investors with forward-looking analysis, risk assessments, and investment guidance (Baker and Nofsinger, 2010). Crucially, producing such reports is not a summarization exercise: professional analysts engage in dialectical reasoning — constructing bull and bear arguments, stress-testing evidence across heterogeneous data sources, and resolving contradictions into defensible core views before committing to structured, visually enriched documents (Bennett and Gadlin, 2012). This process demands days of expert labor per report. Automating it without losing analytical depth is the central challenge of **multimodal financial research report generation** (`MM-FinRG`).

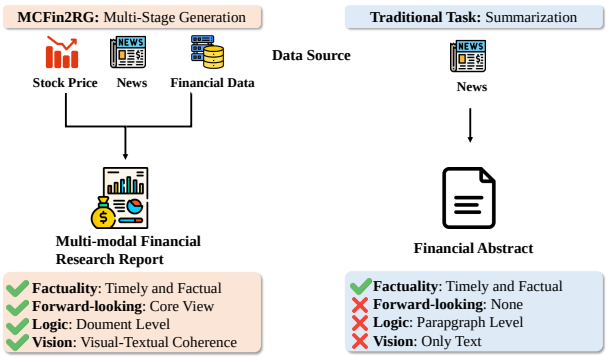

Figure 1: Comparison between our `MM-FinRG` task (left) and the traditional task (right). Our task processes heterogeneous multi-source inputs and generates multimodal, document-level financial research reports.

Existing approaches fall into two lines, both of which inadequately address `MM-FinRG`. **Financial generation methods** treat report creation as single-source summarization (Wang et al., 2024), reducing rich market evidence — news feeds, earnings filings, price time-series — to shallow abstracts that omit the inferential depth investors require. **General long-form and multi-agent methods** offer a broader toolkit: STORM (Shao et al., 2024) retrieves evidence and constructs outlines through perspective-driven questioning; LongWriter (Bai et al., 2024) decomposes generation into sub-tasks; FinRobot (Yang et al., 2024) orchestrates specialized financial agents for analysis pipelines. Yet these systems share a critical blind spot: they treat viewpoint formation as a step rather than a dialectical process, and none produces structured multimodal output in which textual claims are grounded

Dataset	Task	InputDataType	OutputDataType	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
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.

in co-generated visual evidence (Figure 1).

Critically, no existing method simultaneously addresses the three constituent challenges of MM-FinRG: **(C1) evidence synthesis** from heterogeneous sources (market news, stock indicators, annual reports, financial metrics); **(C2) dialectical viewpoint construction** that explicitly models argument–counterargument reasoning to produce balanced, forward-looking core views; and **(C3) schema-grounded multimodal generation** that maintains semantic consistency between textual analysis and supporting visualizations. Beyond methodology, **(C4)** there remains a scarcity of large-scale multimodal benchmarks for this task, and standard metrics (BLEU, PPL, BertScore) are demonstrably inadequate for capturing the factuality, reasoning depth, and visual coherence that define report quality (Table 1).

We propose FinRGAgents, a multi-agent collaborative framework that mirrors the analytical workflow of professional research teams. The system operates through three coordinated stages, each targeting a specific challenge: an Information Summarization stage (**C1**) that deploys specialized agents to extract and reconcile evidence from multi-source inputs; a Collaborative Reasoning stage (**C2**) in which an argument–counterargument mechanism drives two analyst agents to debate competing viewpoints and converge on a defensible core view and structured outline; and a Schema-Grounded Report Writing stage (**C3**) that generates multimodal reports from visual schemas, ensuring alignment between narrative claims and co-produced charts. Critically, the dialectical debate is not mere process scaffolding: replacing the multi-agent debate with a single strong chain-of-thought reasoner (e.g., DeepSeek-R1) yields measurably weaker forward-looking and logical scores, confirming that the gain is structural rather than model-dependent (Table 5).

To address **C4**, we construct FinRG, a large-scale benchmark of 1,500 multimodal financial research reports spanning 25 industry domains, with an average report length of 5,723 tokens — substantially exceeding existing financial generation datasets (Table 1). We further propose a four-dimensional LLM-as-judge evaluation protocol assessing **Factuality**, **Forward-looking reasoning**, **Logical coherence**, and **Vision–text consistency**. Human evaluation confirms its reliability, with Pearson and Spearman correlations of 0.75 and 0.80 against expert ratings — surpassing the 0.5–0.7 range reported for general text generation tasks (Zheng et al., 2025).

Our contributions are as follows:

- We formulate MM-FinRG, a new task requiring multi-source evidence synthesis, dialectical viewpoint reasoning, and schema-grounded multimodal generation, and propose FinRGAgents, a three-stage multi-agent framework that explicitly models each constituent challenge.

- We construct FinRG (1,500 reports, 25 domains) and introduce a task-specific evaluation protocol; human agreement analysis (Pearson = 0.75, Spearman = 0.80) validates its reliability and discriminative power.

- FinRGAgents achieves an average LLM score of 3.53 with GPT-5.1, surpassing the strongest baseline by a clear margin and improving over FinSight by 6.3%; ablation studies confirm that each architectural component — dialectical debate, role specialization, and visual schema verification — contributes significantly ($p < 0.01$) to overall performance.

2 Related Works

2.1 Report Generation

Report generation spans multiple domains with distinct objectives. In medicine (Chen et al., 2021; Wang et al., 2023), multimodal models generate image descriptions from clinical scans. In finance, prior work frames it as summarization: Wang et al. (2024) generates abstracts from news, and Goldsack et al. (2025) builds a multi-agent pipeline from earnings calls. Most recently, FinSight (Jin et al., 2025) introduces a Code Agent with iterative vision enhancement for multimodal report generation, yet relies on single-pass planning without dialectical reasoning. General long-form methods such as STORM (Shao et al., 2024) and LongWriter (Bai et al., 2024) extend coverage via retrieval and task decomposition, but neither addresses financial reasoning or multimodal output. No existing

Industry	News	Reports	Stock	Time Span
Internet (互联网)	3,757	122	1,800	2023.01 - 2024.01
Transportation (交通运输)	3,798	127	1,492	2023.01 - 2024.04
Finance (金融)	4,172	152	1,462	2023.01 - 2024.04
Utilities (公用事业)	2,142	108	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	115	1,326	2023.01 - 2024.07
Culture and Media (文化传媒)	3,415	138	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 8.

approach simultaneously handles heterogeneous multi-source inputs, dialectical viewpoint construction, and visually grounded document generation.

2.2 LLM-as-a-Judge

LLM-based evaluation has been explored in G-Eval (Liu et al., 2023) and related work (Chiang and Lee, 2023), showing that LLM-based scoring can approximate human judgment for open-ended generation. However, performance is task-dependent (Zheng et al., 2023), and recent frameworks targeting multimodal evaluation (Chen et al., 2024; Kwan et al., 2024) remain oriented toward general tasks. No dedicated framework exists for multimodal financial research reports, where assessment must jointly cover factual accuracy, forward-looking reasoning, logical coherence, and vision–text consistency.

2.3 Multi-Agent Collaboration

LLM-based multi-agent systems have demonstrated strong collaborative capabilities. ChatDev (Qian et al., 2024) decomposes development into specialized roles via chat chains; Hu et al. (2024) show that role assignment enables multi-perspective argumentative generation; and Ni and Buehler (2024) scale this with coordinated planning and critique loops. In finance, FinCon (Yu et al., 2024), FinRobot (Yang et al., 2024), and trading-oriented systems (Li et al., 2023b; Tsao, 2023) apply agent orchestration to analysis pipelines. Yet none models the dialectical argument–counterargument process central to professional financial analysis, nor couples it with schema-grounded multimodal generation.

3 FinRG Dataset

3.1 Dataset Processing

We build FinRG from publicly accessible Eastmoney financial disclosures and market information [1], retaining metadata and derived representations for research release subject to source terms. Reports pass a four-criterion quality filter (Appendix A.1), reducing the raw pool by 62%. We select the top 60 companies by market capitalization across 25 industries and use each target report date as the cutoff for all inputs to avoid look-ahead leakage. We then perform four preprocessing steps:

- News Data Collection: Collect news data related to the company from the EastMoney information page [2] for the past years, including news titles and main content. Implement preprocessing steps to filter out articles containing fewer than 20 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 OHLCV price. Candlestick charts are used as high-frequency visual supplements to analyze stock price fluctuations in recent months.
- Annual Report Processing: Obtain the latest annual report PDF document. To facilitate understanding by large models, use the Doc2x [3] parsing tool to convert the annual report into an md document, preserving picture, table structures, and organizing the content by sections.
- Financial Metrics Calculation: We extract raw financial data from the structured tables within the parsed annual reports. These figures are then used to calculate a comprehensive set of key performance indicators, including valuation ratios, profitability metrics, and growth indicators.

3.2 Data Statistics

The data presented in Table 2 from the FinRG dataset illustrates the diversity and volume of multimodal resources across various industries. The Internet and Finance sectors show high activity with substantial news articles and stock data, reflecting industry dynamics. Conversely, the Utilities and Construction sectors demonstrate lower figures, 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.

[1] https://www.eastmoney.com/
[2] https://so.eastmoney.com/news/s?keyword
[3] https://doc2x.noedgeai.com/

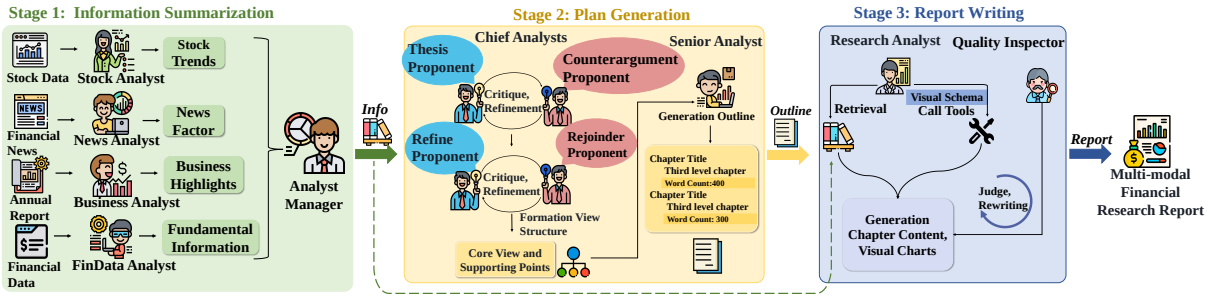

Figure 2: Overview of our framework. Upon receiving a company name, these agents conduct multi-source data organization, then form core insights through deliberative processes and develop a detailed outline plan. Following this outline, they execute a series of subtasks to produce a multimodal financial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

3.3 Dataset Comparison

Table 1 shows FinRG differs from existing benchmarks in three dimensions: multimodal input/output (vs. unimodal prior datasets), richer context (avg. prompt/report: 2,054/5,723 tokens, 5–54× longer than prior work), and broader domain coverage (25 domains vs. single-domain baselines), reflecting real-world complexity. Crucially, all reference reports are institutionally authored and pass a four-criterion quality filter (Section 4.1), ensuring reliable generation targets rather than noisy web text. Unlike prior plain-text datasets, FinRG reports contain co-generated visual charts, enabling visual-textual consistency evaluation.

4 FinRGAgents

We introduce FinRGAgents (Figure 2), a collaborative multi-agent framework based on AutoGen (Wu et al., 2023) for generating multimodal financial research reports. Inspired by real-world organizational structures, FinRGAgents integrates multiple "report agents" with distinct roles 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. Full agent role specifications are in Appendix B; complete system prompt templates and a worked debate example are in Appendix D. The framework employs structured intra/inter-stage coordination, cross-stage memory persistence, and adaptive deliberation, detailed in Appendix C.

4.1 Problem Formulation

We propose MM-FinRG, 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

The four-agent decomposition in this stage directly mirrors the division of labor in a sell-side research team, where distinct analytical perspectives are assigned to specialists rather than aggregated by a generalist. The Stock Analyst Agent applies industry-standard technical indicators—Moving Averages (MA), MACD, and KDJ—that are canonical in CFA-level technical analysis (Murphy, 1999) to identify price momentum and near-term reversal signals. The News Analyst Agent screens for material events (earnings surprises, regulatory changes, macroeconomic shifts) whose market impact is recognized in event-study literature (MacKinlay, 1997). The Business Analyst Agent distills the audited annual report into business highlights, revenue mix, and cash flow quality—mirroring the fundamental analysis layer that underpins equity research. The FinData Analyst Agent computes valuation and profitability ratios (P/E, ROE, etc.) on a trailing-twelve-month basis, grounding the report in the intrinsic-value framework standard in buy-side analysis.

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

where $\mathcal{A}_s$ denotes the `Stock Analyst Agent` (similarly for others) and $\bar{Info}$ the integrated multi-modal market information.

The `Analyst Manager Agent` resolves inter-source conflicts using a domain-motivated priority rule: audited annual-report data takes precedence over market-derived signals (stock trends $>$ news sentiment) when claims are contradictory—reflecting the professional norm that verified accounting data is more authoritative than market noise (Penman, 2013).

$$Info = \mathcal{A}_m(\bar{Info}) \tag{2}$$

4.3 Stage 2: Plan Generation

In professional investment research, a core view is not formed by a single analyst but stress-tested through an investment committee process where bull and bear cases are argued before a thesis is ratified (Jordan et al., 2014). We operationalize this practice: `Chief Analyst Agents` adopt opposing investment stances and engage in multi-round dialectical debate to develop the thesis and sub-arguments, as detailed in Section 4.5. Subsequently, the `Senior Analyst Agent` translates the ratified thesis into a structured report outline—mirroring the editorial role in a research department—with second-level headings, third-level headings, and visual schema specifications. The final report is drafted in Stage 3 based on this outline.

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

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

4.4 Stage 3: Multi-modal Report Writing

Adopting a **Code-as-Action** paradigm to ensure reproducibility, the `Research Analyst Agent` retrieves fine-grained knowledge from the first-phase dataset via RAG based on the outlines. It generates structured content while preventing information loss. Crucially, the `Quality Inspector Agent` conducts a dual-verification loop: assessing compliance with privacy standards and verifying *visual-textual consistency*. If factual mismatches or compliance issues are detected, the `Research Analyst Agent` triggers a revision cycle:

$$\begin{aligned} \tilde{r}_i^{(t)} &= \mathcal{A}_r(o_i, Info, \phi_i^{(t-1)}), \\ (r_i, \phi_i^{(t)}) &= \mathcal{A}_q(\tilde{r}_i^{(t)}, Info, \mathsf{Schema}_i) \end{aligned} \tag{4}$$

where $\tilde{r}_i^{(t)}$ denotes the draft of the i-th section at revision step t, r_i is the accepted section after quality inspection, and $\phi_i^{(t)}$ records feedback for the next revision cycle.

To strictly ground visualizations in factual data, the agent avoids direct pixel generation. Instead, it identifies sections requiring visualization and formulates a **Visual Schema** (JSON) that explicitly maps narrative data to graphical parameters:

$$\mathsf{Schema} \triangleq \begin{cases} \texttt{x_axis} \in \mathbb{R}^n \quad \text{s.t. } x_1 \leq \cdots \leq x_n \\ \texttt{y_axis} \in \mathbb{R}^n \\ \texttt{chart_type} \in \{\texttt{line}, \texttt{pie}, \texttt{bar}\} \end{cases} \tag{5}$$

where $\mathbb{R}^n$ denotes the real numbers, `y_axis` represents the measured quantitative values. See Appendix J for details.

Subsequently, the agent synthesizes executable Python code to invoke plotting tools within a **sandboxed environment**. This deterministic process ensures that every pixel in the chart is traceable to specific data points:

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

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

4.5 Core View Generation

Inspired by stakeholder theory (Freeman, 2010; Rohman, 1965), which posits that diverse stakeholders prioritize different facets of a problem, we simulate this diversity by orchestrating multiple specialized agents to collaboratively construct and refine a central perspective. Following Freeman's argumentative stakeholder framework (Freeman, 2011), we extend the chief analyst agent into three specialized variants: the `thesis proponent agent`, the `counterargument proponent agent`, and the `rejoinder proponent agent`. These agents operate in a pipeline to emulate the dialectical process of argumentation.

Initial Proposition. The `thesis proponent agent` ($\mathcal{A}_{tp}$) first proposes an initial major view v^m based on the writing prompt P and input information $Info$, and then generates a set of supporting views $(v_1, v_2, \ldots, v_n)$:

$$\begin{aligned} \mathcal{A}_{tp} &: P, Info \longmapsto v^m, \\ &: P, Info, v^m \longmapsto (v_1, v_2, \ldots, v_n) \end{aligned} \tag{7}$$

Method	Paradigm	Model	Automatic Evaluation			LLM Evaluation				
			BLEU-2	PPL	Distinct-4	Factuality	Forward-looking	Logical	Vision	Avg
E2E	🤖	Qwen3-32B	0.126	472.84	0.625	2.25	2.81	3.01	—	2.69
		GPT-5.1	0.234	340.42	0.784	2.93	3.03	3.21	—	3.06
CoT	🤖	Qwen3-32B	0.136	328.02	0.723	2.89	2.72	3.01	—	2.87
		GPT-5.1	0.235	318.87	0.827	3.10	2.98	3.35	—	3.15
LongWriter (Bai et al., 2024)	🤖	Qwen3-32B	0.306	312.32	0.735	2.85	2.86	3.21	—	2.97
		GPT-5.1	0.315	293.87	0.795	3.09	3.12	3.37	—	3.19
Dual-Agent	🤖🤖	Qwen3-32B	0.295	302.57	0.757	2.94	2.80	3.17	—	2.97
		GPT-5.1	0.312	283.23	0.842	3.11	3.01	3.22	—	3.11
STORM (Shao et al., 2024)	🤖🤖🤖	Qwen3-32B	0.307	316.26	0.776	2.92	2.83	3.05	—	2.93
		GPT-5.1	0.323	294.76	0.852	3.05	3.08	3.24	—	3.12
Facts2Insights (Goldsack et al., 2025)	🤖🤖🤖	Qwen3-32B	0.286	387.56	0.742	2.92	2.76	3.01	—	2.90
		GPT-5.1	0.302	323.12	0.794	3.13	3.08	3.23	—	3.15
FinRobot (Yang et al., 2024)	🤖🤖🤖	Qwen3-32B	0.328	295.46	0.825	2.55	2.42	2.76	2.62	2.58
		GPT-5.1	0.341	273.12	0.861	2.72	2.64	2.88	2.79	2.75
FinSight (Jin et al., 2025)	🤖🤖🤖	Qwen3-32B	0.339	302.63	0.828	2.73	2.98	3.14	3.29	3.04
		GPT-5.1	<u>0.349</u>	<u>235.62</u>	<u>0.867</u>	<u>3.13</u>	<u>3.24</u>	<u>3.32</u>	<u>3.59</u>	<u>3.32</u>
`FinRGAgents`	🤖🤖🤖	Qwen3-32B	0.340	287.85	0.852	3.01**	3.24**	3.26**	3.38**	3.22**
		GPT-5.1	**0.352**	**216.46**	**0.872**	**3.28****	**3.59****	**3.64****	**3.63****	**3.53****

Table 3: Overall performance of the LLM-powered financial report generation methods, encompassing single-agent(🤖), dual-agent(🤖🤖) and multi-agent(🤖🤖🤖) paradigms. The top scores are in **bold**, with the second highest <u>underlined</u>. Avg is the mean score across LLM metrics. Results are averaged over 5 runs. The symbol ** indicates statistically significant improvement ($p < 0.01$) over the best baseline under a paired t-test. Detailed discussion on metric stability is provided in Chapter 5.5.

Counterargument. To simulate critical stakeholder feedback, the `counterargument proponent agent` ($\mathcal{A}_{cp}$) generates overriding rebuttals r_i that directly challenge the validity or implications of each v_i, and then produces refined views $\bar{v}_i$:

$$\begin{aligned} \mathcal{A}_{cp} &: \{v_i\} \longmapsto r_i, \\ &: \{v_i, r_i\} \longmapsto \bar{v}_i \end{aligned} \tag{8}$$

Rejoinder. Recognizing that initial criticisms may still overlook assumptions or reasoning flaws, the `rejoinder proponent agent` ($\mathcal{A}_{rp}$) introduces undercutting rebuttals r_i^u to further refine the supporting views, resulting in $\tilde{v}_i$:

$$\begin{aligned} \mathcal{A}_{rp} &: \{\bar{v}_i\} \longmapsto r_i^u, \\ &: \{\bar{v}_i, r_i^u\} \longmapsto \tilde{v}_i \end{aligned} \tag{9}$$

View Recomposition and Agreement. Each supporting view v_i undergoes a two-stage refinement process:

$$v_i \xrightarrow{\mathcal{A}_{cp}} \bar{v}_i \xrightarrow{\mathcal{A}_{rp}} \tilde{v}_i \tag{10}$$

where $\bar{v}_i$ and $\tilde{v}_i$ denote the view after overriding and undercutting rebuttals, respectively. The final set of refined views $\{\tilde{v}_1, \ldots, \tilde{v}_n\}$ is then synthesized by the thesis proponent agent into an updated major view:

$$\mathcal{A}_{tp} : \{\tilde{v}_1, \tilde{v}_2, \ldots, \tilde{v}_n\} \longmapsto \bar{V}^m \tag{11}$$

This updated major view $\bar{v}^m$ represents the **implicit consensus** among agents: achieved not through explicit agreement, but through dialectical convergence, where conflicting perspectives are filtered and reconciled iteratively.

5 Experiments

5.1 Baselines

We evaluate `MM-FinRG` in a zero-shot setting, where `FinRGAgents` generate a multimodal 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.

We compare against the following baselines with different generation strategies: (1) **E2E LLM**: direct report generation; (2) **CoT LLM**: generates a brief plan then the report in one pass (Wei et al., 2022); (3) **LongWriter** (Bai et al., 2024): decomposes generation into subtasks written sequentially via AgentWrite. (4) **Dual-Agent**: one agent for outline, one for report writing. (5) **STORM** (Shao et al., 2024): a multi-agent system constructing outlines via retrieval and perspective-driven questioning. (6) **FinRobot** (Yang et al., 2024): an open-source multi-agent financial platform adapted for report generation. (7) **Facts2Insights** (Goldsack et al., 2025): a multi-agent framework for analyti-

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

Model Variant	Forward-looking	Logical
FinRGAgents (ours)	**3.59**	**3.64**
w/o Core View (Debate)	3.46	3.38
w/ Core View (DeepSeek-V3)	3.48	3.41
w/ Core View (DeepSeek-R1)	3.52	3.53

Table 5: LLM-based evaluation scores of different FinRGAgents variants. The full multi-agent design outperforms both the ablated (w/o Core View) and single-agent variant, especially in forward-looking reasoning.

Method Variant	Factuality		Forward-looking		Logical		Vision	
	Score	Δ	Score	Δ	Score	Δ	Score	Δ
FinRGAgents(full)	**3.28**	-	**3.59**	-	**3.64**	-	**3.63**	-
Role Specialization								
w/o Info. Summarization	2.94^{**}	−0.34	3.54^{**}	−0.05	3.48^{**}	−0.16	3.56	−0.07
w/o Stock Analyst	2.90^{**}	−0.38	3.43^{**}	−0.16	3.58	−0.06	3.53	−0.03
Collaborative Reasoning								
w/o Core View	3.22	−0.06	3.46^{**}	−0.13	3.38^{**}	−0.26	3.52	−0.11
w/o Senior Analyst	3.19	−0.08	3.52	−0.07	3.42^{**}	-0.22	3.58	-0.05
Verification Loop								
w/o Visual Schema	3.21	−0.07	3.32^{**}	−0.27	3.46	−0.18	3.26^{**}	−0.37
w/o Quality Inspector	3.18	−0.09	3.57	−0.02	3.61	−0.03	3.48^{**}	−0.15

Table 6: Ablation on component contributions. Δ denotes performance change vs. the full model; ** indicates significant drop ($p < 0.01$, paired t-test).

cal reports from earnings calls. (8) **FinSight** (Jin et al., 2025): multimodal report generation with a Code Agent and iterative vision enhancement. All models are based on GPT-5.1 and Qwen3-32B, with the temperature to 0.50, the top-k to 50, and the top-p set to 0.95. Specific parameter settings are detailed in Appendix G. Full details on how each baseline receives input are provided in Appendix F.

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. Traditional metrics are an insufficient assessment. Consequently, we apply four rubric-based metrics reflecting diverse quality aspects scored by Claude-Sonnet-4-6, then integrate them for a holistic evaluation. Detailed scoring criteria are in Appendix I.

We assess four dimensions (1–5): **Factuality** (accuracy of facts and sourcing), **Forward-looking** (depth and reliability of predictive analysis), **Logical** (coherence and data-to-insights progression), and **Vision** (text–chart consistency). Detailed scoring rubrics are in Appendix I.

Human Evaluation To provide a comprehensive analysis, we conducted a human evaluation involving nine finance postgraduates divided into three independent groups. Inter-annotator agreement reaches an average Cohen's Kappa of 0.759 (Table 10), indicating reliable annotation. Further details are in Appendix H.

5.3 Main Results

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

`FinRGAgents` Achieves State-of-the-Art Performance. `FinRGAgents` achieves strong performance across both automated and LLM-based evaluations. The GPT-5.1 variant attains an average LLM score of 3.53, surpassing the next-best system FinSight (3.32) by a clear margin; the Qwen3-32B variant (3.22) likewise outperforms all baselines. Compared to FinRobot, `FinRGAgents` improves the average score by 28.4% (2.75→3.53), a gain primarily attributed to its argument–counterargument mechanism. Consistent improvements across 25 domains (Figure 3) further confirm the versatility of our approach.

Our LLM-based evaluation demonstrates superior assessment capability. Unlike traditional metrics (e.g., BLEU-2, PPL), which often yield inconsistent or misleading trends, our framework captures deeper semantic and logical integrity. For example, FinRobot (GPT-5.1) shows strong automatic scores but weak LLM evaluation (2.75), while our FinRGAgents (Qwen3-32B) achieves a higher average (3.22) despite similar automatic scores. This indicates that traditional metrics overemphasize surface overlap, whereas our multi-dimensional framework offers a more reliable measure of financial report quality (see Section 5.5).

5.4 Ablation Study

To rigorously assess component contributions, we conducted a mechanism-based ablation study (Table 6). First, regarding **Role Specialization**, remov-

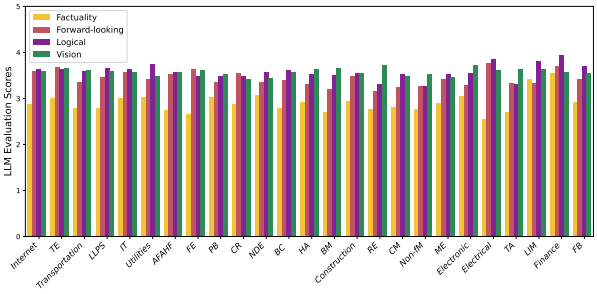

Figure 3: Evaluation results for FinRGAgents across 25 domains are presented using four dimensions: Factuality, Forward-looking, Logical and Vision.

Method	Evaluator	Baseline Wins	FinRGAgents Wins	Draw
E2E	Gemini-3-Pro	13.79%	85.28%	0.93%
	Human	8.24%	91.32%	0.44%
COT	Gemini-3-Pro	17.86%	81.73%	0.41%
	Human	10.56%	89.31%	0.13%
Dual-Agent	Gemini-3-Pro	23.52%	70.55%	5.93%
	Human	11.23%	87.23%	1.54%
FinRobot	Gemini-3-Pro	20.34%	73.21%	6.45%
	Human	12.57%	84.21%	3.22%
FinSight	Gemini-3-Pro	23.34%	70.85%	5.81%
	Human	16.57%	72.78%	10.65%

Table 7: Pairwise evaluation results comparing FinRGAgents against baselines under both Gemini-3-Pro and human evaluation.

ing the Stock Analyst caused the steepest drop in Factuality (Δ=-0.32^{**}), validating the necessity of specialized tools for technical analysis. Second, for **Collaborative Reasoning**, omitting the Core View debate significantly impaired Logical scores (Δ=-0.26^{**}), confirming that dialectical refinement fosters analytical depth. Finally, in the **Verification Loop**, removing Visual Schemas drastically degraded Vision performance (Δ=-0.37^{**}), while the Quality Inspector proved essential for maintaining multimodal consistency (Δ=-0.15^{**}). These significant drops (p<0.01) underscore the indispensability of our multi-agent architecture.

Isolating Architectural Contribution from Workflow Engineering. A natural concern is whether the performance gains of FinRGAgents stem from the multi-agent debate design itself, or simply from the additional workflow scaffolding (e.g., retrieval, outline generation, visual verification) that accompanies it. Table 6 already controls for this: the *w/o Core View* variant retains all three stages, all retrieval tools, and the full visualization pipeline—the sole removal is the dialectical debate. The resulting drops (Δ_{Forward}=-0.13^{**}, Δ_{Logical}=-0.26^{**}) therefore isolate the contribution of debate specifically.

To further stress-test this, Table 5 replaces the multi-agent debate with a single chain-of-thought reasoning model. Even with DeepSeek-R1, Forward-looking is 3.52 and Logical is 3.53, both below our full framework. This confirms that the gain is *structural*: multi-stakeholder dialectical reasoning produces more balanced and defensible views than a single model's internal deliberation, regardless of that model's reasoning capability.

5.5 Human Agreement Evaluation

Following (Li et al., 2023a), we conduct pairwise comparisons evaluated by both human annotators and Gemini-3-Pro. FinRGAgents achieves win rates exceeding 70% against all baselines under both evaluators (Table 7). Human–LLM rating correlations reach Pearson = 0.75 and Spearman = 0.80 on average (Table 4), surpassing the 0.5–0.7 range reported for general text generation (Zheng et al., 2025). Across 5 independent runs, metric standard deviations range from 0.083 to 0.101, and within-band sensitivity tests yield Pearson $r > 0.82$ across low/mid/high quality tiers, supporting stability and discriminability (Appendix K).

5.6 Analysis on Report Structure

RST tree depth analysis (Stede, 2016) shows that FinRGAgents generates reports with deeper and broader structural distributions than LongWriter, STORM, and FinSight baselines, corresponding to an 8.79% Logic score improvement (3.64 vs. 3.32). Full RST depth distributions are provided in Appendix L. **Computational Efficiency.** Generating a full report requires on average **356 seconds** (60× faster than human experts), at a cost of **$5** per report with GPT-5.1, reducible to **$0.15** with Qwen3-32B at minimal performance loss.

6 Conclusion

This paper introduces FinRGAgents, a multi-agent collaborative framework for generating multimodal financial research reports. Using LLM-powered agents and an argument-counterargument mechanism, FinRGAgents integrates fragmented analytical phases to refine core insights. To ensure rigorous assessment, we develop a novel evaluation paradigm, comprising the FinRG dataset and an LLM-as-Judge methodology. Experiments across 25 domains, including ablations and stability analyses, demonstrate the effectiveness of both the framework and evaluation paradigm.

Limitations

While FinRGAgents demonstrates strong performance, several limitations remain.

Visual consistency ceiling. Although our Visual Schema and Quality Inspector mechanism substantially reduces chart errors, complex multi-axis visualizations and mixed-unit charts remain challenging. The sandboxed code execution catches factual mismatches but cannot yet resolve ambiguous axis semantics without human intervention.

Incomplete data coverage. The current pipeline does not incorporate earnings call audio or real-time company announcements, which are important signals for institutional analysts. This may particularly affect forward-looking score accuracy for event-driven sectors such as technology and pharmaceuticals.

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.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. arXiv preprint arXiv:2402.04788.

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.

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.

Tomas Goldsack, Yang Wang, Chenghua Lin, and Chung-Chi Chen. 2025. From facts to insights: A study on the generation and evaluation of analytical reports for deciphering earnings calls. In Proceedings of the 31st International Conference on Computational Linguistics, pages 10576–10593, Abu Dhabi, UAE. Association for Computational Linguistics.

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.

Jiajie Jin, Yuyao Zhang, Yimeng Xu, Hongjin Qian, Yutao Zhu, and Zhicheng Dou. 2025. FinSight: Towards real-world financial deep research. arXiv preprint arXiv:2510.16844.

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.

Bradford D. Jordan, Thomas W. Miller, and Steven D. Dolvin. 2014. Fundamentals of Investments: Valuation and Management, 7 edition. McGraw-Hill Education, New York.

Wai-Chung Kwan, Xingshan Zeng, Yuxin Jiang, Yufei Wang, Liangyou Li, Lifeng Shang, Xin Jiang, Qun Liu, and Kam-Fai Wong. 2024. Mt-eval: A multi-turn capabilities evaluation benchmark for large language models. arXiv preprint arXiv:2401.16745.

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang 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.

A. Craig MacKinlay. 1997. Event studies in economics and finance. Journal of Economic Literature, 35(1):13–39.

John J. Murphy. 1999. Technical Analysis of the Financial Markets: A Comprehensive Guide to Trading Methods and Applications. New York Institute of Finance, New York.

Bo Ni and Markus J Buehler. 2024. Mechagents: Large language model multi-agent collaborations can solve mechanics problems, generate new data, and integrate knowledge. Extreme Mechanics Letters, 67:102131.

Stephen H. Penman. 2013. Financial Statement Analysis and Security Valuation, 5 edition. McGraw-Hill Education, New York.

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 Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15174–15186.

D Gordon Rohman. 1965. Pre-writing: The stage of discovery in the writing process. College Composition & Communication, 16(2):106–112.

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 Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6252–6278, Mexico City, Mexico. Association for Computational Linguistics.

Manfred Stede. 2016. Towards assessing depth of argumentation. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3308–3317, Osaka, Japan. The COLING 2016 Organizing Committee.

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.

Siyuan Wang, Bo Peng, Yichao Liu, and Qi Peng. 2023. Fine-grained medical vision-language representation learning for radiology report generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 15949–15956, Singapore. Association for Computational Linguistics.

Ziao Wang, Yunpeng Ren, Xiaofeng Zhang, and Yiyuan Wang. 2024. Generating long financial report using conditional variational autoencoders with knowledge distillation. IEEE Transactions on Artificial Intelligence.

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. Advances in neural information processing systems, 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. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. arXiv preprint arXiv:2308.08155.

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. arXiv preprint arXiv:2405.14767.

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. arXiv preprint arXiv:2407.06567.

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. arXiv preprint arXiv:2501.03936.

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. Advances in Neural Information Processing Systems, 36:46595–46623.

A Dataset Description

The basic information of the `FinRG` dataset is shown in Table 8.

Industry	News	Reports	Stock	Time Span	Industry	News	Reports	Stock	Time Span
Internet(互联网)	3,757	122	1,800	2023.01 - 2024.01	TE(交运设备)	4,058	231	1,250	2023.01 - 2024.01
Transportation(交通运输)	3,798	127	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	157	1,734	2023.01 - 2024.01	Utilities(公用事业)	2,142	108	1,101	2023.01 - 2024.06
AFAHF(农林牧渔)	3,358	124	1,342	2023.04 - 2024.06	FE(化石能源)	4,521	83	1,573	2023.02 - 2024.03
PB(医药生物)	4,312	143	1,384	2023.02 - 2024.03	CR(商贸零售)	2,683	125	1,263	2023.01 - 2024.01
NDE(国防与装备)	3,525	101	1,153	2023.01 - 2024.01	BC(基础化工)	3,782	212	1,282	2023.03 - 2024.03
HA(家电)	3,326	94	986	2023.03 - 2024.03	BM(建材)	4,252	123	1,452	2023.02 - 2024.02
Construction(建筑)	3,612	93	1,297	2023.02 - 2024.07	RE(房地产)	3,233	134	1,182	2023.02 - 2024.02
CM(文化传媒)	3,415	138	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	123	2,346	2023.02 - 2024.02	Electronic(电子设备)	4,273	119	1,963	2023.02 - 2024.02
Electrical(电气设备)	3,254	134	1,634	2023.03 - 2024.03	TA(纺织服装)	4,215	123	1,324	2023.05 - 2024.06
LIM(轻工制造)	3,123	87	972	2023.03 - 2024.04	Finance(金融)	4,172	152	1,462	2023.01 - 2024.04
FB(食品饮料)	3,591	115	1,326	2023.01 - 2024.07					

Table 8: Comprehensive overview of the full basic information for 25 domains in the `FinRG` dataset.

A.1 Quality Filter Criteria

We retain only reports that satisfy all four of the following criteria: (1) attributed to registered institutional contributors (as identified by Eastmoney's verified-author tag); (2) contain at least 2,000 characters of substantive analysis; (3) include at least one structured financial data table; (4) published within the target time window for each company. This filtering reduces the raw pool by approximately 62%, yielding a curated set of professionally authored reports suitable as generation targets.

B Agent Role Specifications

The description of the agent in `FinRGAgents` is depicted in the figures 4.

C Framework Protocol

Our system, `FinRGAgents`, is designed as a collaborative multi-agent framework. Each agent has a specialized role, and structured protocols govern their interactions to ensure coherence and reliability. The architecture is defined by three core principles: coordination, memory, and adaptive behavior.

Agent Coordination and Communication. The framework employs a structured collaboration design where agents interact through well-defined protocols:

- `Intra-Stage Coordination:` In the summarization stage, the `Analyst Manager Agent` aggregates outputs from specialist agents. It resolves inter-source conflicts using predefined heuristics, such as prioritizing signals from the annual report when market news and stock trends diverge (as formalized in Eq. 2).
- `Inter-Stage Communication:` Information is passed between stages via structured representations. For instance, the core views ($\overline{v}^m$) and report outline (O) are encoded in a structured format (e.g., JSON) to guide the downstream writing task (Eq. 4). This outline explicitly contains fields for visual schemas, directly instructing the `Research Analyst Agent` on where and what charts to generate (Eq. 5).

Information Flow and Memory. The system incorporates not only a global long-term memory but also short-term memory, and it supports cross-stage information persistence.

- `Long-term Memory:` The summarized information ($Info$) from Stage 1 is cached and serves as a long-term memory base, accessible by all downstream agents. For example, the `Chief Analyst Agent` references it during debate (Eq. 3), and the `Research Analyst Agent` retrieves from it during content generation (Eq. 4).
- `Feedback Loop:` A short-term memory mechanism is implemented through a feedback loop. The `Quality Inspector Agent` verifies the generated content $\tilde{r}_i$; if factual inconsistencies or compliance issues are found, it triggers a revision cycle by the `Research Analyst Agent`, as indicated by the loop symbol in Eq. 4.

Adaptive Deliberation. The system exhibits dynamic, deliberative behavior in two key areas:

- `Dialectical Reasoning:` Instead of a simple pipeline, the core view generation process involves a multi-turn debate among the `Chief Analyst Agents` with different stances. (Eq. 8-11). Arguments are dynamically proposed, challenged with rebuttals, and refined, allowing the system to explore multiple perspectives before converging on a final thesis.
- `Selective Visualization:` The `Research Analyst Agent` adaptively determines where a visualization is needed to support the text. It then selects the most appropriate chart type (e.g., line, bar, pie) based on the data's structure and the narrative's intent, ensuring strong visual-textual alignment (Eq. 5).

D Agent Prompt Templates

To support reproducibility, we provide the system prompts for the three Chief Analyst roles and the Quality Inspector, along with a condensed example of one complete dialectical debate round.

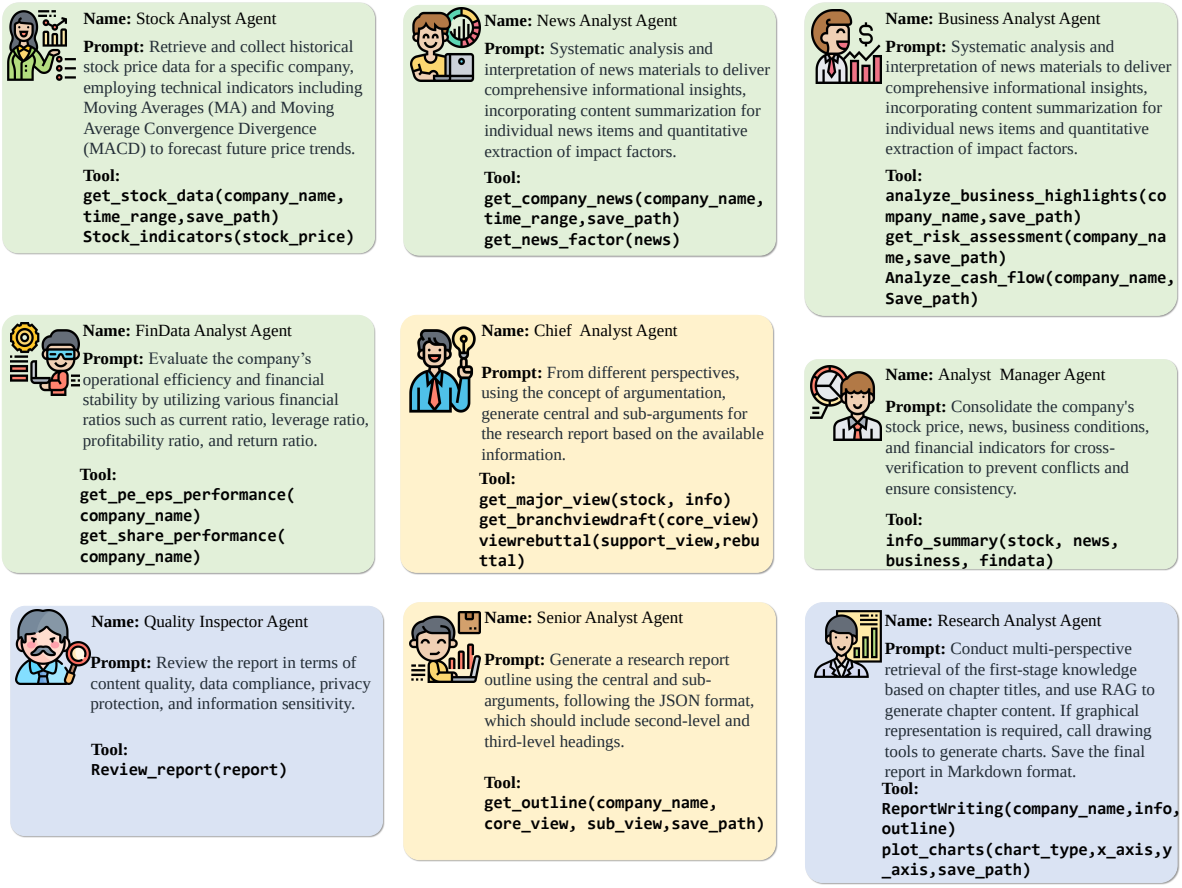

Figure 4: Detailed description of each agent in `FinRGAgents`

D.1 Chief Analyst Agent Prompts

Thesis Proponent Agent — System Prompt

You are a **senior equity analyst** responsible for forming the core investment thesis. Given the summarized market information `{Info}`, you must:

1. Propose one major investment view v^m (Buy / Hold / Sell with a 12-month price target rationale).
2. Generate **3–5 supporting sub-views** $(v_1, \ldots, v_n)$, each grounded in at least one data point from `{Info}`.

Output format (JSON): `{"major_view": "...", "supporting_views": ["v1", "v2", ...]}`

Counterargument Proponent Agent — System Prompt

You are a **devil's advocate analyst** tasked with stress-testing the thesis. For each supporting view v_i provided, you must:

1. Generate an **overriding rebuttal** r_i that directly challenges the validity or assumptions of v_i.
2. Produce a **revised view** $\bar{v}_i$ that integrates the challenge into a more cautious or nuanced position.

Constraint: every rebuttal must cite a specific risk factor or contradictory data point.
Output format (JSON): `{"rebuttals": [{"view_id": i, "rebuttal": "...", "revised_view": "..."}, ...]}`

Rejoinder Proponent Agent — System Prompt

You are a **senior strategist** performing a second-pass critique. For each revised view $\bar{v}_i$, identify any **hidden assumptions or reasoning flaws** not addressed by the first rebuttal, and produce:

1. An **undercutting rebuttal** r_i^u exposing the overlooked weakness.
2. A **further-refined view** $\tilde{v}_i$.

Stopping criterion: if consensus score > 0.9 (cosine similarity between $\bar{v}_i$ and $\tilde{v}_i$ embeddings), output `"converged": true` and skip further refinement.
Output format (JSON): `{"refinements": [{"view_id": i, "undercut": "...", "final_view": "...", "converged": bool}, ...]}`

D.2 Quality Inspector Agent Prompt

Quality Inspector Agent — System Prompt

You are a **compliance and consistency reviewer**. Given a draft report section $\tilde{r}_i$ and the source information `{Info}`, perform a dual check:

1. **Factual grounding**: Verify that every quantitative claim in $\tilde{r}_i$ can be traced to a specific entry in `{Info}`. Flag ungrounded claims.
2. **Visual–textual consistency**: If $\tilde{r}_i$ references a chart, verify that the chart's axes and data match the narrative description.

If any issue is found, return `"pass": false` with a structured list of violations. If all checks pass, return `"pass": true`. Maximum revision cycles: **3**. If issues persist after 3 cycles, flag for human review. Output format (JSON): `{"pass": bool, "violations": [{"type": "factual|visual", "description": "..."}]}`

D.3 Dialectical Debate: Condensed Example

The following illustrates one complete debate round for a single supporting view, using a simplified Pharmaceutical company case.

Debate Round Example (Pharmaceutical Sector)

Thesis Proponent v_1: *"The company's pipeline includes 3 Phase-III candidates; successful approval of any one would increase revenue by ∼30% based on comparable drug launches."*
Counterargument r_1: *"Phase-III approval rates in oncology average 58% (FDA, 2023). With 3 candidates, the probability of zero approvals is ∼7%, but the expected revenue impact is substantially lower than projected due to likely market competition."*
Revised view $\bar{v}_1$: *"Pipeline optionality provides upside, but projected 30% revenue growth is optimistic; a conservative 12–18% range is more defensible given competitive dynamics."*
Rejoinder r_1^u: *"The revised estimate still ignores launch timing risk: even approved drugs average 18 months from approval to peak sales, implying the revenue impact falls outside the 12-month investment horizon."*
Final view $\tilde{v}_1$: *"Pipeline candidates provide medium-term optionality (18–24 month horizon) rather than near-term catalysts; current valuation should be discounted accordingly."*
`converged: true` (similarity = 0.93)

E Broader Impact and Discussion

E.1 Cross-Market Generalizability

A primary concern in financial NLP is the dependency on language-specific resources. While `FinRG` is constructed using Chinese data (A-share market), the underlying cognitive architecture of `FinRGAgents` is fundamentally language- and market-agnostic.

The framework decouples the *Reasoning Layer* (Agents) from the *Data Source Layer*. Adapting our system to other markets (e.g., NASDAQ or LSE) requires only two localized modifications:

- **Data Interface Replacement:** Switching the data retrieval API from localized providers (e.g., Eastmoney) to global standards such as SEC EDGAR, Bloomberg Terminal, or Yahoo Finance.
- **Prompt Localization:** Translating the agent prompt instructions into the target language.

The core mechanisms—such as the argument-counterargument debate workflow and the visual-textual verification loop—transcend linguistic boundaries. This modular design ensures that `FinRGAgents` can be readily generalized to global financial analysis contexts without architectural overhaul.

E.2 Ethical Considerations and Risk Mitigation

Automated financial report generation carries inherent risks related to hallucinations and potential misuse. We explicitly define `FinRGAgents` as a *Decision Support System* (DSS) rather than an *Autonomous Trading Agent*. To ensure responsible deployment, we implement a three-tier safeguard:

- **Traceability:** The `Research Analyst` is constrained to generate content strictly based on retrieved evidence. The system enforces mandatory citation tracking, allowing human users to verify every claim against the original source documents.
- **Adversarial Verification:** The `Quality Inspector` acts as an adversarial filter, tasked with rejecting any claims that lack data support or exhibit logical fallacies. As evidenced by the Factuality ablation in Table 6, this verification loop significantly reduces hallucination rates.
- **Human-in-the-Loop:** We emphasize that the generated reports serve to accelerate human analysis, not replace it. The system identifies volatility and risks but leaves the final investment execution to human experts, ensuring accountability in high-stakes financial environments.

F Equal Information Access Across Baselines

To ensure a fair comparison, all baselines receive the identical input $\mathcal{I} = \{c, n, s, a, f\}$ defined in Section 4.1. For methods that do not natively support heterogeneous multimodal input (E2E, CoT, LongWriter, Dual-Agent, STORM), the four input components are serialized into a unified text prompt with explicit section markers (`[NEWS]`, `[STOCK]`, `[ANNUAL_REPORT]`, `[FINANCIALS]`), ensuring identical information access across all conditions. For retrieval-augmented baselines (STORM, Facts2Insights, FinRobot, FinSight), the same underlying corpus—the Stage-1 summarized information $Info$—serves as the retrieval source, so no baseline benefits from an exclusive knowledge advantage.

G Implementation Details and Reproducibility

To ensure the reproducibility of `FinRGAgents`, we provide comprehensive details regarding model configurations, agent-specific parameters, and the execution environment. The complete source code, including the Dockerized environment and preprocessing scripts, is available in the supplementary material and will be open-sourced upon acceptance.

G.1 Model Configurations

Our experiments distinguish generation backbones from evaluator models. Report generation utilizes two primary backbone models: **GPT-5.1** (via OpenAI API) and **Qwen3-32B** (deployed locally). Scalar rubric scoring is conducted with Claude-Sonnet-4-6 using the prompts in Appendix I, while pairwise preference evaluation is conducted with Gemini-3-Pro and human annotators. DeepSeek-V3 and DeepSeek-R1 are used only for the core-view ablation, where they replace the debate module while the rest of the report-generation pipeline is kept fixed. All API-based runs log the provider model identifier, access date, prompt template, and decoding parameters.

- **Context Window:** We utilize the 128k context window for GPT-5.1 to handle extensive retrieval documents. For Qwen3-32B, we adhere to a 64k context limit, employing a sliding window strategy for extremely long inputs.
- **Sampling Parameters:** To balance creativity and factual consistency, we set the temperature to 0.5, top-p to 0.95, and frequency penalty to 0.0.

G.2 Agent-Specific Parameters

Different agents require distinct resource allocations based on their roles. For instance, the `Stock Analyst` requires a larger context to process numerical time-series data, while the `Chief Analyst` requires strictly controlled output lengths for concise debate. Table 9 details these configurations.

Table 9: Detailed parameter specifications for each agent in the FinRGAgents framework.

Agent Role	Max Context	Max Output	Key Configuration / Tool Usage
Stock Analyst	8,192	2,048	Access to `calculate_MACD/KDJ`; Rolling window processing.
News Analyst	16,384	4,096	Semantic deduplication threshold set to 0.85 (cosine similarity).
FinData Analyst	8,192	2,048	Strict JSON output enforcement for financial tables.
Chief Analyst	32,768	2,048	**Debate Mechanism:** Max 3 rounds; Debate terminates if consensus > 0.9.
Thesis/Counter. Agents	16,384	1,024	Prompted for adversarial reasoning (Red-teaming mode).
Senior Analyst	8,192	1,024	Output format constrained to hierarchical JSON outline.
Research Analyst	128,000	4,096	**Code-as-Action:** Sandboxed Python execution for Matplotlib.
Quality Inspector	32,768	1,024	Visual-Textual Consistency Check; Feedback loop limit = 3.

G.3 Environment and Hardware

Software Stack. We implement the multi-agent framework using `AutoGen` (v0.2) and manage dependencies via Docker. The plotting environment utilizes `Matplotlib` (v3.8) and `Seaborn` (v0.13) within a secure sandbox to prevent arbitrary code execution risks.

Hardware Setup.

- **API-based Experiments:** Conducted on standard CPU instances, serving as the lightweight controller for GPT-5.1 calls.
- **Local Inference (Qwen3-32B):** Deployed on a high-performance cluster equipped with $4\times$ NVIDIA A100 (80GB) GPUs. We use `vLLM` for efficient inference serving, achieving an average throughput of 45 tokens/sec.

H Human Evaluation and Inter-Annotator Agreement

To ensure the reliability of our human evaluation, we recruited nine finance postgraduates and divided them into three independent groups, each consisting of three annotators. Outputs were anonymized and randomly shuffled before annotation, and annotators were not shown system identities. For scalar rating, each annotator received the generated report, the same source evidence available to the generator, and chart data or visual schemas when visual outputs were present. Each group independently rated

the assigned samples across the four rubric-based evaluation dimensions. For pairwise preference evaluation, annotators compared anonymized outputs from `FinRGAgents` and each baseline under the same evidence context. For each group, we report the average pairwise Cohen's Kappa among the three annotators; the final IAA is the mean across groups.

As summarized in Table 10, the average Kappa score reached 0.759. This level of consistency suggests that the scoring rubric is reliable for benchmarking.

Group	Cohen's kappa
1	0.731
2	0.746
3	0.802
Avg.	0.759

Table 10: Consistency analysis results showing the inter-annotator agreement (IAA) scores across different groups. The last row shows the average IAA scores for all groups.

I Prompts for LLM Evaluation

The following prompts were employed for LLM-based evaluation:

Factuality Instruction

[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.
[2.50,2.99]-The text contains frequent minor inaccuracies or occasional major omissions. Some data points (e.g., financial figures or timelines) may be misreported or inconsistently presented. Roughly half of the claims lack clear or verifiable sources, and the factual base is often fragmented or outdated. While the text attempts to present facts, it falls short in maintaining precision and consistent sourcing.
[3.00,3.49]-The text demonstrates basic factual accuracy with limited but noticeable inconsistencies. Most key facts are generally correct, but some supporting data may be outdated, weakly sourced, or oversimplified. Around 70% of factual claims are verifiable, with the remaining relying on secondary or unclear references. The content provides a reasonable overview but lacks nuance and evidentiary rigor in several areas.
[3.50-4.49]-The text demonstrates partial factual accuracy, but includes notable omissions or secondary errors. Some key statements lack clear sourcing or rely on questionable references. While basic facts are generally correct, supporting details may contain inaccuracies or ambiguities. The overall credibility is weakened by inconsistent verification across the content.
[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.]

ForwardLooking Instruction

[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.
[2.50,2.99]- The report provides a limited level of forward-looking analysis with projections that are general and lack analytical depth. While some future metrics are discussed (e.g., revenue growth or market share), the assumptions behind them are weak or unsubstantiated. The analysis lacks the rigor needed to support reliable decision-making.
[3.00,3.49]-The report presents a basic but structured forward-looking analysis. Forecasts are loosely tied to industry trends or strategies, with partially realistic assumptions. However, risk and scenario analysis are limited, and key external factors may be overlooked, reducing the analysis's overall robustness.
[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 cred-

ibility of future predictions, incorporating macroeconomic, industry, and company-specific factors. Point out any significant weaknesses or strengths.]

Logical Instruction

[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.
[2.50,2.99]-The report demonstrates a basic level of coherence, with a recognizable structure such as an introduction, body, and conclusion. Arguments follow a logical sequence but often lack sufficient elaboration. While data backs some conclusions, the reasoning tends to be superficial or generic.
[3.00,3.49]-The report presents a reasonably clear and structured analysis, with sections and arguments that are easy to follow. Key conclusions are generally supported by data, and logical flow is maintained across most parts of the report. However, the depth of reasoning is limited, and the narrative may overlook complexities.
[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.]

Vision Consistency Instruction

[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.
[2.50,2.99]-The text content and the charts exhibit partial alignment. Charts reflect some of the key data points mentioned in the text (e.g., revenue trends or categorical distributions), but important contextual or comparative elements are missing.
[3.00,3.49]-The text content and the charts demonstrate general consistency. However, the charts do not fully convey the complex relationships discussed in the text, which limits their effectiveness in supporting the argument.
[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:
Vision 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.]

J Visual Schema Generation

The visual schema is generated by GPT-5.1 based on the following instructions:

Visual Schema Generation Prompt

Objective: Extract key information and data points from the background knowledge and current chapter to first determine if the chapter is suitable for adding visualizations that can be represented as bar charts, line graphs, or pie charts.

Judge Criteria
1. Summary of Background Knowledge:\n - Summarize the background knowledge related to this chapter. This should include fundamental concepts of the field, important data sources, and any known information relevant to the content of the current chapter. This information will help identify important data points within the chapter.
2. Chapter Title Analysis:\n - Based on the chapter title, infer the possible themes and content it may cover. This helps predict the key topics and types of data that will be discussed in the chapter.
3. Key Information Identification:\n - List the 3-5 most important information points in the chapter. These should be directly linked to the chapter title and background knowledge and should support subsequent data visualization.
4. Data Point Extraction:\n - From the chapter content, identify specific data points related to the key information. This can include:\n - Specific values and statistics (suitable for bar and line charts).\n - Time-series data (suitable for line charts).\n - Category and proportion data (suitable for pie charts).
5. Visualization Recommendations:\n - Based on the extracted data points, recommend the most appropriate type of chart and briefly explain the rationale. If the chapter is not suitable for adding a chart, state the reason. For example: \n - "The chapter data is not suitable for visualization because it lacks specific data points or the data type does not

support the aforementioned charts." \n - "A bar chart is suitable because the data involves comparisons across multiple categories." \n - "A line chart is suitable because the data represents trends over time." \n - "A pie chart is suitable because the data reflects the relationship between parts and the whole."

Output Format Requirements

Visualization Chart Type and X-axis/Y-axis Data: - Based on the background knowledge, provide the most appropriate visualization chart type along with the corresponding X-axis and Y-axis data and the chart title in JSON format. For example:
{
"$chart_type$" : "bar",
"x_axis" : ["$content_1$", "$content_2$"],
"y_axis" : [num_1, num_2],
"$caption$" : "$chart_title_content$"
}
Strictly adhere to the JSON format without changing the field names, ensuring that `x_axis` is a list of temporal, numeric, or categorical labels and `y_axis` is a numeric list of the same length.

K Evaluation Stability and Fine-Grained Analysis

Stability Analysis. We randomly selected 25 reports (one per industry) and scored them across five independent evaluation runs. As shown in Table 11, the population standard deviations range from 0.083 to 0.101, with the narrowest variation for `Forward-looking` (σ=0.083).

Run	Factuality	Forward-looking	Logical	Vision
Round 1	3.00	3.52	3.60	3.51
Round 2	3.15	3.55	3.58	3.65
Round 3	3.13	3.40	3.65	3.72
Round 4	3.10	3.59	3.75	3.63
Round 5	3.30	3.65	3.85	3.79
Mean	**3.14**	**3.54**	**3.69**	**3.66**
Std. Dev. (σ)	**0.097**	**0.083**	**0.101**	**0.094**

Table 11: LLM evaluation scores across five independent runs (25 reports). Reported σ values are population standard deviations across runs.

Fine-Grained Sensitivity. To verify that small score differences (e.g., 3.1 vs. 3.3) reflect meaningful quality distinctions, we tested within-band sensitivity using 15 reports per quality tier (low/mid/high). Pearson correlations between LLM scores and human ratings are strong and significant across all bands ($r = 0.85, 0.82, 0.88$; $p < 0.01$), as shown in Table 12.

Quality Band	Score Range	Std.Dev. (σ)	Pearson
Low-Quality	1.5 – 2.5	0.05	0.85
Mid-Quality	3.0 – 3.4	0.04	0.82
High-Quality	4.0 – 5.0	0.03	0.88

Table 12: Within-band stability (σ) and sensitivity (Pearson r vs. human) across quality tiers.

L RST Tree Depth Analysis

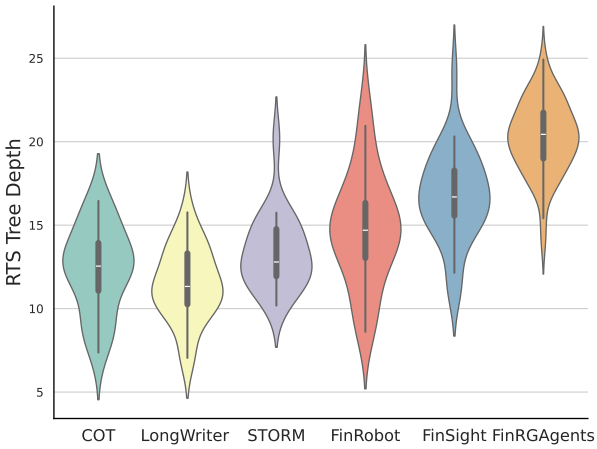

Figure 5: Distribution of RST tree depth of generated financial reports across methods. `FinRGAgents` produces reports with deeper and broader structural distributions, corresponding to an 8.79% improvement in Logic scores over FinSight (3.64 vs. 3.32).