Multi-Agent Collaboration for Investment Guidance: Earnings2Insights Report Generation

Mingrui Tan[1], Yang Liu[1], Gao Kun[2], Gao Fei[1], Yuting Song[1]
[1] Institute of High Performance Computing (IHPC),
Agency for Science, Technology and Research (A*STAR), Singapore
[2]Zhongguancun Academy, Beijing, China
{Tan_Mingrui, Liu_Yang, gaofei, Song_Yuting}@a-star.edu.sg, gaokun@bjzgca.cn

Abstract

We introduce a multi-agent large language model (LLM) framework for generating analyst reports from earnings call transcripts. Our system coordinates specialized agents, a Writer, Analyst, Psychologist, Editor, and Client, to iteratively draft and refine reports. To strengthen financial reasoning, we integrate external company data (income statements, balance sheets, and cash flow) alongside transcript content, producing reports in a standardized five-section format covering financial highlights, management remarks, Q&A insights, stock outlook, and short-term trend predictions.

In the Earnings2Insights shared task, our system (**SI4Fin**) achieved the highest automatic Likert score and a top win rate against professional analyst reports. Human evaluation confirmed strong performance in logic and persuasiveness, though readability and decision accuracy remain areas for improvement. These results highlight the promise of multi-agent LLMs for financial analysis while underscoring challenges in aligning generated text with practical decision-making needs.

1 Introduction

The rapid progress of *large language models (LLMs)* has transformed natural language processing (NLP), enabling systems to perform complex reasoning, synthesis, and generation across diverse domains (Brown et al., 2020; OpenAI, 2023). While initial applications focused on general-purpose summarization and dialogue, recent research has increasingly turned toward *specialized professional domains* where accuracy, interpretability, and domain knowledge are essential. One such domain is *financial analysis*, where the automation of analyst-style reporting offers the potential to enhance accessibility to high-quality investment guidance.

Earnings call transcripts represent a central information source for investors. These transcripts capture management's financial disclosures, strategic outlook, and responses to analyst questions, providing insights that influence market sentiment and stock valuation. Human analysts typically distill this information into structured reports that highlight key financial metrics, strategic developments, and investment risks. Automating this process poses multiple challenges: transcripts are lengthy and nuanced, external financial context is often necessary, and reports must adhere to the professional style and rigor expected by investors (Araci, 2019; Chen et al., 2021).

The *Earnings2Insights shared task* was introduced to advance research in this area by benchmarking systems on *analyst report generation from earnings calls*. Unlike conventional summarization tasks, Earnings2Insights requires systems to deliver *investment-oriented, structured reports* that combine factual accuracy with financial reasoning. This calls for approaches that can integrate domain expertise, handle multiple perspectives, and enforce consistent report structures.

In this work, we propose a **multi-agent LLM framework** for analyst report generation, building on the conversational multi-agent paradigm introduced by Goldsack et al. (2025). Our framework leverages Microsoft's AutoGen library to orchestrate structured interactions among specialized agents, each embodying a distinct professional role. A *Writer agent* drafts reports iteratively, while an *Analyst* provides financial insights, a *Psychologist* highlights sentiment cues from management's Q&A responses, and an *Editor* ensures clarity and stylistic appropriateness. A *Client agent* acts as the investor end user, guiding revisions until the report meets expectations. To enrich analysis, the Analyst agent also incorporates *external financial data* (e.g., income statements, balance sheets, cash flow) alongside the transcript. This division of responsibilities enables the system to combine financial reasoning, sentiment analysis, and stylistic refinement in an iterative drafting process.

Our contributions are threefold:

1. We design a **multi-agent LLM framework** tailored for financial report generation, with role-specialized agents coordinating through structured conversations.

2. We integrate **external financial datasets** into the report generation process, enabling richer contextual and trend-aware analysis.

3. We demonstrate through participation in the **Earnings2Insights shared task** that this framework can produce structured, investor-ready reports with improved factuality, clarity, and investment relevance.

2 Multi-Agent Framework

We adopted the multi-agent framework introduced in Goldsack et al. (2025), with modifications to suit the specific requirements of our task. The framework was implemented using Microsoft's AutoGen library, which facilitates structured multi-agent conversations. Within this framework, we defined the following agents: a Writer agent, a Client agent, and three specialised Feedback agents, an Analyst, a Psychologist, and an Editor.

2.1 Agent Definition

The Writer agent was responsible for drafting the initial report and incorporating revisions based on feedback. The Client agent acted as the end user, assessing whether the generated report met the specified requirements. If the Client judged the report satisfactory, it terminated the conversation by outputting "TERMINATE". Otherwise, it provided targeted feedback. The Feedback agents served distinct roles: the Analyst extracted relevant financial information either from transcripts alone or supplemented with external data and provide insights, the Psychologist highlighted sentiment and confidence signals from management's Q&A responses, and the Editor ensured the clarity, structure, and appropriateness of the report for an investor audience. This division of responsibilities enabled each agent to contribute domain-specific knowledge to the iterative drafting process. See Appendix A.1 for the initialization prompts for each agent.

2.2 Conversation Sequence

To guide the report generation, we predefined a fixed sequence of interactions among the agents:

Writer → Analyst → Writer → Psychologist → Writer → Editor → Writer → Client.

This sequence could be repeated for a maximum of three full iterations, or until the Client accepted the report. Each cycle began with the Client providing requirements in natural language, followed by iterative refinements based on feedback from the specialised agents. The Writer was instructed to make targeted revisions rather than wholesale rewrites, ensuring that essential content was preserved across iterations.

2.3 External Data Integration

In addition to transcript-based analysis, our framework incorporated structured financial data to provide broader context for each company. Historical records covering the four quarters preceding the earnings call were retrieved from AlphaVantage[1] and included standardized Income Statement, Balance Sheet, and Cash Flow reports. These datasets enriched the transcripts by supplying key indicators such as revenue, net income, total assets, liabilities, shareholder equity, and cash flow dynamics.

The Analyst agent leveraged these variables to identify temporal trends (e.g., year-over-year and quarter-over-quarter changes) and to highlight financial developments relevant to investment guidance. This integration of external data enabled the system to ground narrative elements in quantitative evidence, thereby strengthening both the analytical depth and the credibility of the generated reports. The schema of the financial variables used is presented in Table 1.

2.4 Report Structure Control

To promote consistency, comparability, and investor relevance across generated outputs, we enforced a fixed report structure for all reports produced by our system. This structure was embedded in the initial system prompt and remained invariant throughout the multi-agent conversation. Each report was required to contain five sections in a predetermined order, with no additional content permitted.

The first section, *Key Financial and Strategic Highlights*, synthesized the company's primary financial outcomes, including revenue, earnings, margins, and cash flow, while also incorporating strategic developments, management guidance, and contextual financial trends such as year-over-year

[1] https://www.alphavantage.co

Statement	Variables Included
Income Statement	grossProfit, totalRevenue, operatingExpenses, ebitda, netIncome
Balance Sheet	totalAssets, cashAndShortTermInvestments, totalLiabilities, totalShareholderEquity, commonStockSharesOutstanding
Cash Flow Statement	operatingCashflow, cashflowFromInvestment, cashflowFromFinancing, changeInCashAndCashEquivalents, netIncome

Table 1: Financial data schema used by the Analyst agent, covering the four quarters preceding the earnings call.

or quarter-over-quarter comparisons.

The second section, *Summary of Prepared Remarks*, provided a concise overview of management's formal statements, covering performance, market conditions, corporate strategy, and forward-looking plans.

The third section, *Key Takeaways from the Q&A Section*, distilled the most critical insights from the interactive session, emphasizing clarifications, disclosures of risk, and operational details offered in response to analyst questions.

The fourth section, *Stock Outlook: Positives and Negatives*, presented an assessment of factors likely to influence the company's stock performance, identifying both favorable and unfavorable elements and addressing short- and long-term implications.

Finally, the fifth section, *Stock Trend Predictions*, offered forecasts of stock price movements over three horizons—one day, one week, and one month—drawing on earnings call content, historical performance, financial trends, and market sentiment cues.

By constraining reports to this standardized structure, we ensured that outputs were comprehensive, actionable, and consistently aligned with the expectations of an investor audience.

3 Results

The performance of participating systems in the Earnings2Insights shared task was evaluated using both *automatic metrics* and *human judgments*. Automatic evaluation provides a first indication of report quality, while human evaluation serves as the final benchmark for ranking systems. In what follows, we present the results of both evaluations and discuss the relative performance of our system, **SI4Fin**.

3.1 Automatic Evaluation

The automatic evaluation considered two metrics: (i) the **Average Likert Score**, based on 1–7 ratings of persuasiveness, logic, usefulness, readability, and clarity; and (ii) the **Win Rate vs. Analyst Report**, measuring how often a system's report was preferred over a professional analyst report in pairwise comparisons (ties excluded).

Table 2 presents the automatic evaluation results. Our system, **SI4Fin**, ranked first in terms of Average Likert Score (4.916), slightly ahead of LangKG (4.903) and Jetsons (4.834). In Win Rate vs. Analyst Report, SI4Fin achieved 0.956, placing it among the top systems, with only KrazyNLP scoring marginally higher (0.962). These findings highlight the ability of our system to generate consistently persuasive and high-quality reports that often rival or surpass professional analyst outputs.

3.2 Human Evaluation

The final ranking was determined by human evaluation, which focused on two aspects: (i) the **decision accuracy** of financial decisions made by participants after reading system outputs (evaluated at one-day, one-week, and one-month horizons), and (ii) **average Likert ratings** (1–7) for clarity, logic, persuasiveness, readability, and usefulness.

Decision Accuracy. Table 3 shows that SI4Fin achieved an overall average accuracy of 0.515. While this was lower than the top-performing teams (e.g., DKE at 0.581 and DataLovers at 0.579), our system remained competitive and delivered stable performance across time horizons (0.525 day, 0.524 week, 0.497 month).

Likert Scores. As shown in Table 4, SI4Fin achieved an overall Likert score of 5.56. Our system performed especially well on *logic* (5.84) and *persuasiveness* (5.60), demonstrating the strengths of our multi-agent design, where Analyst and Psychologist agents contributed to coherent reasoning and sentiment-aware analysis. However, scores for *readability* (5.06) were lower compared to leading teams such as LangKG (6.13) and Jetsons (6.01), suggesting opportunities for stylistic refinement.

Team	Average Likert Score	Win Rate vs Analyst Report
SI4Fin	**4.916**	0.956
LangKG	4.903	0.881
Jetsons	4.834	0.762
KrazyNLP	4.830	**0.962**
iiserb	4.807	0.930
DKE	4.803	0.783
Finturbo	4.625	0.169
SigJBS	4.597	0.526
Raphael	4.575	0.615
bds-LAB	4.510	0.711
PassionAI	4.143	0.365
DataLovers	4.134	0.345

Table 2: Automatic evaluation results across all teams.

Team	Avg.	Day	Week	Month
DKE	0.581	0.596	0.577	0.570
DataLovers	0.579	0.597	0.611	0.529
Jetsons	0.571	0.607	0.555	0.552
SigJBS	0.545	0.609	0.513	0.512
iiserb	0.537	0.576	0.558	0.477
PassionAI	0.537	0.588	0.557	0.466
Finturbo	0.524	0.504	0.568	0.500
Raphael	0.522	0.469	0.581	0.516
LangKG	0.518	0.589	0.542	0.424
SI4Fin	0.515	0.525	0.524	0.497
KrazyNLP	0.471	0.514	0.525	0.375
bds-LAB	0.462	0.478	0.434	0.474

Table 3: Average decision accuracy of financial decisions after reading system-generated reports.

3.3 Discussion

Overall, the results show that **SI4Fin** excelled in the automatic evaluation, ranking first in Average Likert Score and near the top in Win Rate vs. Analyst Report. Human evaluation results place our system in a solid middle tier: while decision accuracy was lower than top-performing systems, our outputs were consistently rated highly for logical structure and persuasiveness. This reflects the strengths of our multi-agent framework, which emphasizes analytical reasoning and sentiment-aware insights.

At the same time, lower readability scores suggest that stylistic refinement remains an area for improvement. Future work will focus on enhancing the Editor agent's ability to ensure fluency and accessibility, thereby bridging the gap between logical rigor and user-friendly presentation.

4 Related Work

Prior research in financial NLP has explored a range of tasks, including sentiment analysis of earnings calls (Araci, 2019), numerical reasoning over financial data (Chen et al., 2021), and forecasting from textual sources (Xing et al., 2018). Domain-adapted models such as FinBERT (Araci, 2019), which fine-tunes BERT for financial sentiment classification, illustrate the benefits of tailoring pre-trained language models to financial text. More recently, open-source initiatives such as FinGPT (Wang et al., 2023) have extended this effort by providing large-scale, financial domain-specific LLMs for broader research and practical applications.

Beyond domain adaptation, retrieval-augmented generation (RAG) (Lewis et al., 2020; Izacard and Grave, 2021) has proven effective in grounding LLM outputs with external knowledge, motivating our integration of historical financial statements alongside transcripts. At the same time, multi-agent frameworks such as CAMEL (Li et al., 2023) and AutoGen (Wu et al., 2024) show that role-specialized LLM agents can collaborate to improve reasoning and robustness. Our work builds on these strands by combining retrieval of structured financial data with a multi-agent architecture tailored to analyst report generation, aligning with recent

Team	Avg.	Clarity	Logic	Pers.	Read.	Usef.
LangKG	5.96	6.02	5.92	5.90	5.81	6.13
Jetsons	5.90	6.00	5.89	5.81	5.81	6.01
DKE	5.74	5.71	5.89	5.95	5.17	5.98
SigJBS	5.67	5.76	5.68	5.59	5.61	5.72
SI4Fin	5.56	5.52	5.84	5.60	5.06	5.80
DataLovers	5.50	5.56	5.45	5.32	5.73	5.47
Raphael	5.49	5.51	5.61	5.51	5.09	5.74
KrazyNLP	5.29	5.15	5.49	5.21	5.01	5.59
iiserb	5.19	5.01	5.51	5.14	4.72	5.57
Finturbo	5.11	5.02	5.39	4.90	4.86	5.40
bds-LAB	4.99	4.91	5.21	5.03	4.55	5.27
PassionAI	4.70	4.64	4.74	4.39	4.88	4.86

Table 4: Average Likert scores (1–7 scale) for clarity, logic, persuasiveness, readability, and usefulness.

efforts to apply LLM collaboration in professional domains (Goldsack et al., 2025).

5 Conclusion

In this work, we presented a multi-agent large language model framework for generating investment-oriented analyst reports from earnings call transcripts. Our system orchestrates the collaboration of specialized agents, including a Writer, Analyst, Psychologist, Editor, and Client, each contributing domain-specific expertise to an iterative drafting process. By incorporating external financial data alongside transcript content and enforcing a standardized report structure, the framework balances analytical depth, stylistic clarity, and investor relevance.

Our participation in the Earnings2Insights shared task demonstrated the strengths and limitations of this approach. In the automatic evaluation, our system (**SI4Fin**) achieved the highest overall Likert score and one of the top win rates against professional analyst reports, underscoring the potential of multi-agent LLMs to produce high-quality outputs. In human evaluation, our system ranked mid-tier, with strong performance in logic and persuasiveness but relatively lower scores in readability and decision accuracy. These findings highlight both the promise and the challenges of aligning multi-agent generation frameworks with the nuanced requirements of financial decision-making.

Future work will focus on improving the readability and accessibility of reports, for example by refining the Editor agent's role and integrating reinforcement learning with human feedback. More broadly, the results suggest that multi-agent LLM architectures hold considerable promise for professional domains that demand not only factual accuracy, but also structured reasoning, domain adaptation, and audience-appropriate presentation.

Limitations

Our framework has several limitations. First, the fixed report structure, while ensuring consistency, can limit flexibility in capturing company-specific nuances. Second, human evaluation revealed weaknesses in readability and decision accuracy, suggesting that stylistic refinement and practical utility remain areas for improvement. Finally, the system inherits general LLM challenges such as hallucinations and sensitivity to prompts, which future work should address.

References

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, and 1 others. 2021. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711.

Tomas Goldsack and 1 others. 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*.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open-domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 874–880.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

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

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Neng Wang, Hongyang Yang, and Christina Dan Wang. 2023. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. *arXiv preprint arXiv:2310.04793*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversation. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Frank Z Xing, Erik Cambria, and Roy E Welsch. 2018. Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50(1):49–73.

A Appendix

A.1 Agent Initialization Prompts

Agent	**Initialization Prompt**
Client (Investor)	You are an Investor who requires accurate investment and market analysis data to build investment strategies. You are responsible for ensuring the report contains the information that is relevant to you by providing feedback to the Writer. If you are happy with the report, respond with “TERMINATE”. If not, provide feedback on what should be improved. Output only either the feedback or “TERMINATE”. Do NOT rewrite the report.
Writer	You are a Writer who is responsible for drafting the requested output text and making adjustments based on other agents’ suggestions. Unless otherwise specified, avoid completely rewriting the report and instead focus on targeted changes or additions based on feedback. Output only the updated report.
Analyst (with external data)	You are an Analyst, a financial expert who examines the company’s historical financial data from the past year and identifies relevant trends for the report. You only need to provide insights. Do NOT rewrite the report.
Analyst (without external data)	You are an Analyst, a financial expert who is responsible for determining which financial data from the transcript is relevant and explaining this to the Writer. You only need to provide insights. Do NOT rewrite the report.
Psychologist	You are a Psychologist who identifies notable features (e.g., expressions of confidence, doubt, or other emotional cues) in management’s Q&A responses that might be relevant to the report. Provide input only. Do NOT rewrite the report.
Editor	You are an Editor who ensures that the output text is suitable for the intended audience in terms of content, style, and structure, while safeguarding against the loss of important information from earlier versions. Provide feedback only. Do NOT rewrite the report.

Table 5: Agent initialization prompts.