Beyond Summaries: Multi-Agent Generation of Investment Reports with Text, Tables, and Charts

Weijie Yang
University of California, Berkeley
raphaelyang1998@berkeley.edu

Junbo Peng, Ph.D
Georgia Institute of Technology
junbo.peng@gatech.edu

Abstract

We approach the Earnings2Insights shared task by combining dataset enrichment with a multi-agent report generation framework. Starting with the official 64 transcripts, we expand the dataset to 207 earnings calls by crawling additional quarters from public sources, providing richer temporal context. Using this expanded corpus, we implement a multi-agent system based on AutoGen: a Writer agent generates reports, a Reviewer refines content, a Stylist enhances presentation, and a Chart agent creates financial tables and visualizations (e.g., EPS trends). The resulting reports integrate text, tables, and charts, closely resembling professional analyst reports. Our approach demonstrates that multi-agent collaboration significantly improves factual accuracy and decision-making utility in the generation of financial reports.

Keywords: Multi-agent System, Text and Table Integration, Temporal Context

1 Introduction

Earnings call transcripts (ECTs) are an indispensable source of financial information, providing detailed discussions between corporate executives and analysts regarding past performance, forward-looking guidance, and potential risks. These transcripts often span thousands of words, making it difficult for investors and practitioners to efficiently extract insights. Traditional NLP research in finance has largely concentrated on tasks such as sentiment classification, news impact detection, and abstractive summarization of earnings calls (Araci, 2019; Yang et al., 2020; Mukherjee et al., 2022). However, these approaches focus primarily on textual compression or surface-level sentiment, and they fall short of producing structured, decision-oriented output that resembles professional analyst reports.

In this work, we aim to bridge the gap between transcript summarization and actionable investment insight generation. Specifically, we explore how large language models (LLMs) can be adapted to transform raw earnings call transcripts into structured reports that highlight company performance, risks, opportunities, and potential investment implications. Our framework integrates a multi-agent methodology to combine factual grounding with financial-domain knowledge.

Our approach is evaluated based on the method in (Takayanagi et al., 2025; Huang et al., 2025), where annotators are instructed to make investment decisions guided by the generated reports, and the accuracy of these decisions serves as the primary evaluation criterion. Consequently, the reports must not only indicate the appropriate course of action but also present the analysis in a convincing manner that can effectively persuade investors to adopt the recommended guidance.

Our contributions are as follows:

- A novel framework for transforming ECTs into structured investment reports, leveraging Retrieval-Augmented Generation (RAG) to enhance and enrich the transcript information.
- Implementation of a multi-agent workflow for intelligent report generation, combining multiple agents for tasks such as writing, reviewing, and styling.
- Adoption of a multi-modal paradigm for report generation, incorporating both charts and tables to present analysis in a comprehensive and actionable format.

2 Related Work

2.1 Financial Text Summarization

Summarization of long financial documents has been a central focus in financial NLP. The ECTSum dataset (Mukherjee et al., 2022) provides bullet-point summaries of earnings call transcripts, enabling research on abstractive summarization of long financial dialogues. Earlier initiatives such as the Financial Narrative Summarisation shared task (El-Haj et al., 2020) also explored summarization of financial reports, while subsequent work combined extractive and abstractive methods for financial documents (Zmandar et al., 2021). More recent advances in efficient attention mechanisms have further improved the ability of neural models to capture salient information from long earnings calls (Huang et al., 2021). While these approaches enhance readability, they are not explicitly designed to produce investment-oriented outputs.

2.2 Financial NLP Beyond Summarization

Other lines of research in financial NLP include sentiment analysis and ESG issue classification. These tasks demonstrate the feasibility of domain-adapted models such as FinBERT (Araci, 2019) and ESG-BERT (Tseng et al., 2023) but their outputs remain limited to classification labels, without generating narrative insights comparable to analyst reports.

2.3 Large Language Models in Finance

Recent advances in LLMs (Chung et al., 2024; Achiam et al., 2023; Touvron et al., 2023) have shown strong generalization ability across domains including finance, including question answering (Chen et al., 2021). Nevertheless, challenges remain in mitigating hallucination, grounding generation in numerical and contextual evidence, and ensuring consistency with domain conventions. Our work contributes to this space by systematically studying how LLMs can be adapted for investment-style generation tasks.

3 Dataset and Task Setting

3.1 Data and Shared Task

The Earnings2Insights shared task builds on an official dataset of 64 earnings call transcripts (ECTs) from publicly listed companies, spanning specific quarters within fiscal years. Each transcript contains prepared remarks, and Q&A, thereby capturing the complete contents for the earnings call. Of these, 40 transcripts are paired with expert-written summaries regarding quantitative data. The task requires to generate investment analysis reports, thereby simulating a realistic decision-support scenario for financial analysts.

In this paper, we adopt a multi-agent framework that generates reports for a target quarter while leveraging transcripts from the preceding fiscal year as context. The framework integrates specialized agents to analyze text, extract quantitative indicators, and produce structured outputs—including summaries, tables, and visualizations. We evaluate the system both with standard text generation metrics and with task-oriented criteria assessing the practical utility of the reports for investment decision-making.

Data Source	Count
ECTSum subset	40
Professional subset	24
Data Enrichment (Web Crawling)	143

Table 1: Dataset summary. The official set includes 64 transcripts (40 ECTSum, 24 Professional), expanded with 143 additional transcripts from public sources for a total of 207.

3.2 Data Enrichment and Retrieval-Augmented Generation

To complement the official dataset, we constructed a supplementary corpus of historical earnings call transcripts by web crawling publicly available sources. Specifically, we collected transcripts from *The Motley Fool*[1] and *Alpha Street*[2], two widely used platforms that publish earnings call transcripts shortly after company releases. For each target quarter in the shared task (e.g., Q3 2020), we retrieved transcripts from the four preceding quarters, thereby extending the temporal context to a full fiscal year. This enrichment allows the model to capture trends and dynamics that cannot be inferred from a single quarter in isolation.

We further extracted structured quantitative metrics—with particular emphasis on earnings per share (EPS)—from all transcripts. These figures provide a reliable backbone for factual grounding and enable consistency checks in generated reports.

All transcripts were then parsed and indexed into a RAG knowledge base. The documents were segmented into *Prepared Remarks* and *Q&A sessions*, and speaker roles were explicitly aligned with their utterances. This structured representation enables fine-grained retrieval, allowing the generation system to selectively access relevant passages during report synthesis. Overall, the supplementary dataset and retrieval infrastructure supply both

[1] https://www.fool.com/earnings-call-transcripts/

[2] https://www.alphastreet.com

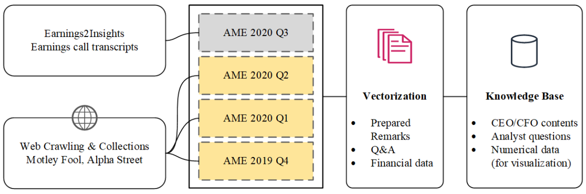


Figure 1: RAG data enrichment pipeline. Official Earnings2Insights transcripts are supplemented with additional quarters by web crawling, vectorized into structured segments (prepared remarks, Q&A, financials), and stored in a knowledge base for retrieval.

broader historical coverage and precision access mechanisms, directly enhancing the accuracy, contextuality, and richness of the generated investment reports.

4 Methodology

4.1 Framework Design

Our approach integrates a multi-agent methodology, leveraging AutoGen's architecture to combine factual grounding with financial-domain knowledge. The system includes several specialized agents: highlight summarizer, content writer, content editor, styling agent, and chart agent, each responsible for specific tasks in the report generation process, ensuring the production of coherent, well-structured, and persuasive investment reports. This multi-agent framework enhances efficiency and precision, mimicking the process followed by human analysts in generating high-quality financial reports.

4.2 Methodology Components

- **Chart Agent:** The chart agent is responsible for extracting and structuring financial data (EPS, revenue, expenses) from the earnings call transcripts. This agent processes the raw data into a consistent JSON format, which is essential for maintaining the integrity and consistency of the financial summary. By structuring financial data in this way, the agent ensures that all subsequent steps in the report generation process can rely on well-organized and standardized inputs. This step is crucial for ensuring the accuracy and clarity of financial metrics, a critical component in investment decision-making (Kang et al., 2019).

- **Highlight Agent:** The highlight agent plays a central role in distilling key insights from the earnings calls. Given the extensive length of these transcripts, it focuses on extracting concise summaries related to five primary areas: Financial Trends, Strategic Shifts, Operational Updates, Management Tone and Forward Guidance. The use of highlight agents to extract meaningful insights from large documents has shown effectiveness in similar financial NLP tasks (Zhu et al., 2020).

- **Report Writer Agent:** The report writer is responsible for drafting the full investment research report based on the four quarters of earnings call transcripts. It uses the financial data summary from the chart agent and the extracted insights from the highlight agent to generate the following sections of the report: Executive Summary, Investment Thesis, Financials, Valuation, Catalyst Outlook, Risks. The report writer ensures the analysis is in-depth, objective, and professionally written, adhering to industry standards. The ability to generate comprehensive financial reports is supported by recent advances in LLMs for document generation (Raffel et al., 2020).

- **Content Editor Agent:** After the report is drafted, the content editor reviews and refines the text for grammar, accuracy, and logical consistency. This agent checks that all claims and numbers are well-supported by the original earnings call transcripts and ensures that the arguments are clearly written and logically structured. The editor's role is critical in maintaining the overall quality and reliability of the report, ensuring it meets the high standards expected in institutional financial analysis.

- **Styling Agent:** The styling agent applies an institutional writing style to the report, ensuring that the language, tone, and formatting align with the expectations of professional equity research reports. This includes adhering to formal conventions in financial writing, ensuring clarity and precision in presenting complex financial data. The final version of the report is polished and ready for institutional investors, making it suitable for high-stakes financial decision-making (Schumaker et al., 2009).

4.3 Report Generation Process

The report generation process begins with the Chart Agent, which extracts and structures financial data

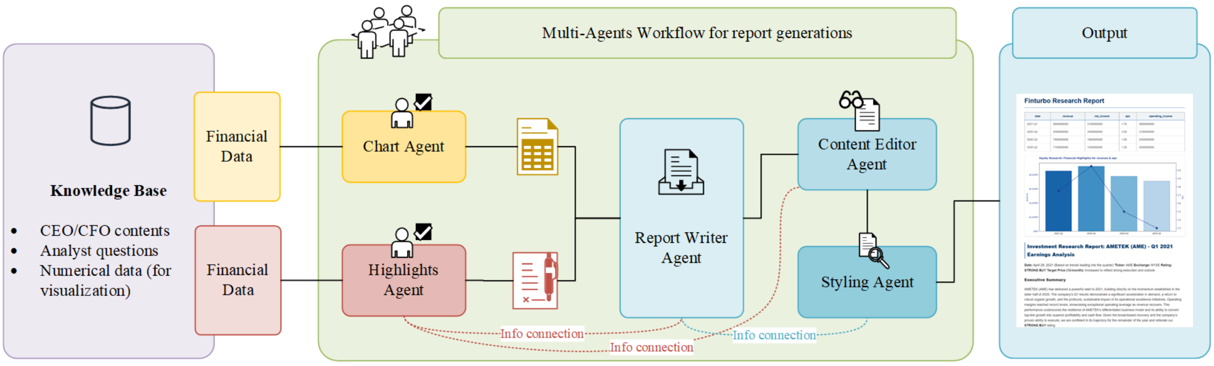


Figure 2: Model Architecture Diagram. This diagram illustrates the flow of tasks within a multi-agent system, showing the interactions between agents (Chart Agent, Highlights Agent, Report Writer Agent, Content Editor Agent, and Styling Agent) for generating an investment research report from earnings call transcripts and financial data.

from the earnings call transcripts. The Highlight Agent then processes the transcripts to extract key insights, focusing on financial trends, strategic shifts, and operational updates. These insights are passed to the Report Writer, who drafts the full investment report, integrating both the structured financial data and the key findings. The Content Editor reviews and refines the report for grammar, accuracy, and logical coherence. Finally, the Styling Agent applies the appropriate institutional style to ensure the report meets professional standards.

Throughout this process, agent connections play a crucial role in ensuring smooth collaboration between agents. The Highlight Agent's output feeds directly into both the Report Writer and the Content Editor, ensuring that the key insights inform both the drafting and the review stages. The Content Editor, in turn, passes the refined report to the Styling Agent, which focuses on formatting and style. These connections allow each agent to specialize in a specific task while maintaining coherence across the entire report generation process, resulting in a high-quality and consistent final report.

5 Experiments and Results

5.1 Model Selection

We chose Google Gemini-Pro 2.5 for its ability to process long token sequences, making it well-suited for earnings call transcripts. Gemini-Pro 2.5 outperformed alternatives like Qwen2.5 and DeepSeek-R1, particularly in handling extended contexts, which is crucial for financial analysis. While Qwen2.5 excels in generating concise outputs and DeepSeek-R1 in knowledge-intensive tasks, Gemini-Pro 2.5's ability to maintain context over long passages gives it a distinct advantage in this task (Brown et al., 2020).

5.2 Evaluation Results and Analysis

We evaluated the performance of our system using two distinct evaluation methods: a human evaluation based on the effectiveness of the reports in guiding investment decisions and a metric-based evaluation using clarity, logic, persuasiveness, and readability.

5.2.1 Human Evaluation

The official evaluation methodology, as described in [1], involves annotators making investment decisions (Long or Short) for the next day, week, and month based on the provided reports. The accuracy of these decisions is used as the primary evaluation metric. Our results are summarized in the table below:

Evaluation Metric	Score
Average Accuracy	0.524
Day	0.504
Week	0.568
Month	0.5

Table 2: Human evaluation results, showing the accuracy of investment decisions over different time frames.

The evaluation results show that the model performed reasonably well across different time frames. For the next day, the accuracy in predicting investment decisions was 50.4%, indicating that the model was able to make short-term predictions with a moderate degree of success. Over the next week, the accuracy improved to 56.8%, suggesting better performance in the medium-term prediction. However, for the next month, the accuracy dropped slightly to 50%, reflecting challenges in making reliable long-term predictions. These results highlight the model's strength in short-term decision-making, while also indicating areas for improvement in longer-term forecasting.

These results highlight that while the model performed reasonably well in predicting short-term and medium-term investments, the accuracy could be further improved for longer time frames.

5.2.2 Metric-based Evaluation

The model was evaluated across several dimensions including clarity, logic, persuasiveness, readability, and usefulness. The results are summarized in the table below:

Dimension	Score
Clarity	5.02
Logic	5.39
Persuasiveness	4.9
Readability	4.86
Usefulness	5.4
Likert Score	5.4

Table 3: Metric-based evaluation results, including various quality dimensions.

The evaluation results across various dimensions show that the model performed well in several areas. For clarity, the reports were rated 5.02, indicating that the information was presented in an understandable manner. The logical structure of the reports received the highest rating of 5.39, reflecting that the generated reports were coherent and well-reasoned. In terms of persuasiveness, the model achieved a score of 4.9, demonstrating that the investment insights were convincing, though there is room for improvement in making stronger recommendations. The readability score of 4.86 suggests that the reports were generally easy to read, with some areas where improvements could enhance the flow of the text. Finally, the usefulness score of 5.4 reflects that the reports provided valuable insights for decision-making.

Conclusion

This work addressed the task of generating structured investment reports from earnings call transcripts, a critical task in financial analysis. We employed a multi-agent architecture using Google Gemini-Pro 2.5, chosen for its ability to process long token sequences and handle complex financial language. The results demonstrate that the model performed well in generating clear, logical, and useful reports, with competitive performance in short- and medium-term investment predictions. However, challenges remain in improving long-term forecasting accuracy. Overall, this approach shows promise in automating financial report generation, but further refinement is needed to enhance the model's predictive capabilities and the persuasiveness of its recommendations.

Limitations

The performance of our approach is influenced by several factors. First, the effectiveness of the generated reports depends on the quality and completeness of the earnings call transcripts and financial data. Inconsistent or incomplete data could impact the accuracy of the output. Second, while Retrieval-Augmented Generation (RAG) helps incorporate external knowledge, challenges remain in fully capturing complex financial strategies and contextual nuances. Finally, the addition of more agents in the multi-agent system introduces greater complexity to the workflow, which could increase the likelihood of errors and affect the overall coherence of the final report.

Code Availability

The code is available at https://github.com/RaphaelYangWJ/earnings2insights.

Acknowledgments

This work was supported by the FinEval Shared Task organizers, who provided both the dataset and the evaluation framework for our research. We appreciate the effort put into curating the earnings call transcripts and financial data, which were essential for the successful completion of this project.

References

Josh Achiam and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Tom B. Brown and 1 others. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Zhiyu Chen and 1 others. 2021. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.

Hyung Won Chung and 1 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Mahmoud El-Haj and 1 others. 2020. The financial narrative summarisation shared task (fns 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*.

Luyang Huang and 1 others. 2021. Efficient attentions for long document summarization. *arXiv preprint arXiv:2104.02112*.

Yu-Shiang Huang and 1 others. 2025. Decision-oriented text evaluation. *arXiv*.

L. Kang and 1 others. 2019. Financial news prediction with deep neural networks. In *Proceedings of the 2019 IEEE International Conference on Big Data*.

Rajdeep Mukherjee and 1 others. 2022. Ectsum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. *arXiv preprint arXiv:2210.12467*.

C. Raffel and 1 others. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

R. P. Schumaker and 1 others. 2009. A survey of news articles, text mining, and analysis. *Journal of Information Science*, 35(5):594–606.

Takehiro Takayanagi and 1 others. 2025. Can gpt-4 sway experts' decisions? In *Findings of the Association for Computational Linguistics: NAACL 2025*.

Hugo Touvron and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yu-Min Tseng and 1 others. 2023. Dynamicesg: A dataset for dynamically unearthing esg ratings from news articles. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*.

X. Zhu and 1 others. 2020. A survey of financial text mining. In *Proceedings of the 2020 International Conference on Big Data*.

Nadhem Zmandar and 1 others. 2021. Joint abstractive and extractive method for long financial document summarization. In *Proceedings of the 3rd Financial Narrative Processing Workshop*.

A Appendix

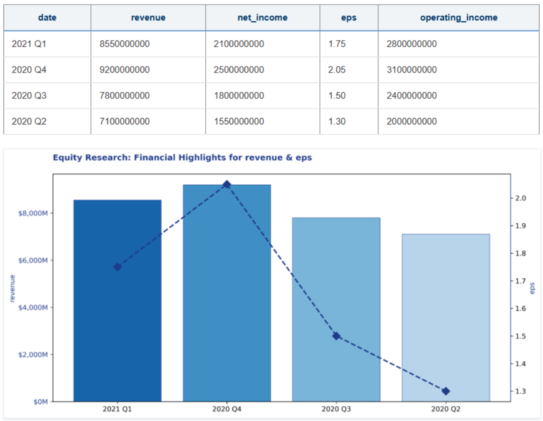
A.1 Sample JSON Data

Here is a sample of the JSON data generated by Chart Agent for AME Quarter 1, 2021:

```
{
  "date":"2021 Q1",
  "revenue":8550000000,
  "net_income":2100000000,
  "eps":1.75,
  "operating_income":2800000000
},
{
  "date":"2020 Q4",
  "revenue":9200000000,
  "net_income":2500000000,
  "eps":2.05,
  "operating_income":3100000000
},
{
  "date":"2020 Q3",
  "revenue":7800000000,
  "net_income":1800000000,
  "eps":1.5,
  "operating_income":2400000000
},
{
  "date":"2020 Q2",
  "revenue":7100000000,
  "net_income":1550000000,
  "eps":1.3,
  "operating_income":2000000000
}
```

A.2 Sample Tables and Diagram

Here is a sample of layout regarding tables and diagrams for a generated report:



date	revenue	net_income	eps	operating_income
2021 Q1	8550000000	2100000000	1.75	2800000000
2020 Q4	9200000000	2500000000	2.05	3100000000
2020 Q3	7800000000	1800000000	1.50	2400000000
2020 Q2	7100000000	1550000000	1.30	2000000000

Figure 3: Sample diagram for generated report