From Earnings Calls to Investment Reports: Evaluating Role-based Multi-Agent LLM Systems

Ranjan Satapathy[1], Raphael Liew[2], Joyjit Chattoraj[1], Erik Cambria[2], Rick Siow Mong Goh[1]
[1]Institute of High Performance Computing (IHPC),
Agency for Science, Technology and Research (A$*$ STAR), Singapore,
[2]Nanyang Technological University, Singapore
Correspondence: satapathy_ranjan@a-star.edu.sg

Abstract

This paper presents a novel multi-agent framework leveraging LLMs for automated financial analysis and investment report generation from earnings call transcripts. Traditional financial analysis struggles with increasing volumes of unstructured data. We propose a collaborative multi-agent system that mimics professional analyst team structures through role specialization. Our framework employs three specialized agents: Analyst, Writer, and Editor, that collaborate through structured workflows with tool support for financial data retrieval and sentiment analysis. Through extensive human evaluation on the Prolific platform, we demonstrate that our system achieves good accuracy in guiding financial decisions, placing it competitively among twelve evaluated systems. The system scores high on human quality assessment, with particularly strong performance in usefulness, indicating practical value for investment decision-making. In automatic evaluation, our system outperforms professional analyst reports most of the time, validating its competitive quality. Our findings provide empirical evidence that role-based agent collaboration offers a balanced approach to AI-generated financial analysis, demonstrating stable performance that prioritizes practical utility over surface-level report quality.

1 Introduction

Entity engagement and investment target prioritization have become increasingly critical for institutional investors navigating dynamic financial markets. This process heavily relies on financial analysis, which has traditionally depended on manual examination of structured data such as balance sheets (Loughran and McDonald, 2016). However, the exponential growth of unstructured data sources - including earnings calls, patent filings, and social media sentiment has rendered traditional approaches inefficient in capturing real-time market insights (Du et al., 2024b).

Early deep learning solutions attempted to automate parts of this process through sentiment analysis models and neural networks for risk forecasting (Loughran and McDonald, 2016). However, these approaches face significant challenges with data drift, where gradual or rapid changes in the input data distribution cause a degradation of model performance (Lu et al., 2018). Market dynamics such as regulatory changes, emerging sectors, and macroeconomic shocks alter data distributions, requiring costly retraining with substantial computational resources and labor intensive data labeling (Alzubaidi et al., 2021).

Recent advances in generative AI, particularly LLMs, offer a compelling alternative. Pretrained on massive diverse corpora, LLMs can interpret complex contextual relationships without task-specific retraining (Du et al., 2024a). They excel at capturing subtle linguistic cues for nuanced sentiment analysis and have demonstrated strong performance in summarization, question answering, and market sentiment prediction (Yang et al., 2024a). Domain-specialized models like BloombergGPT (Wu et al., 2023) and FinGPT (Yang et al., 2023) further demonstrate the benefits of adapting general-purpose models to financial text.

However, existing systems predominantly adopt single agent paradigms in which one LLM handles the entire analysis pipeline. While effective for narrow applications, these frameworks struggle with hallucinations causing factual inaccuracies (Kang and Liu, 2023) and incomplete coverage when tackling complex tasks such as investment reporting. Moreover, current financial AI frameworks lack realistic organizational modeling, failing to replicate the structured workflows and division of labor characteristic of professional analyst teams (Yu et al., 2023). In this paper, we address such limitations by exploring multi-agent LLM systems for financial analysis and investment report generation.

We design and implement a collaborative agent framework powered by GPT-4.1 that analyzes financial textual data and produces well-informed investment recommendations. Our key contributions are as follows:

- A novel multi-agent framework that mimics professional analyst team structures through specialized role assignment
- Evidence that agent collaboration with iterative feedback significantly improves report quality and factual accuracy

2 Related Work

2.1 Explainability and Interpretability in Financial AI

The deployment of AI systems in financial contexts demands not only accuracy but also transparency and interpretability, particularly given regulatory requirements and the high-stakes nature of investment decisions. Recent work has comprehensively examined the landscape of explainable AI (XAI) in finance (Yeo et al., 2025b), highlighting the critical need for systems that can provide faithful and interpretable explanations for their outputs.

The challenge of generating trustworthy explanations from LLMs has received considerable attention. Yeo et al. (2025a) demonstrate through activation patching that natural language explanations from LLMs may not always faithfully represent their internal decision-making processes, raising important questions about the reliability of single-agent systems that lack verification mechanisms. This finding directly motivates our multi-agent approach, where the Editor agent serves as an additional layer of validation for the explanations and reasoning provided by other agents.

Interpretability concerns extend beyond individual model outputs to the reasoning processes themselves. Jie et al. (2024c) examine how interpretable reasoning explanations from prompted LLMs actually are, finding significant variability in explanation quality. Our multi-agent framework addresses this through role specialization: the Analyst agent provides data-grounded explanations, while the Editor ensures these explanations maintain logical consistency and clarity.

The extraction of interpretable rationales from financial text presents unique challenges. Jie et al. (2024b) propose semi-supervised approaches for extractive rationalization, which aligns with our Analyst agent's function of identifying and extracting key financial metrics and insights from earnings transcripts. Similarly, Ong et al. (2023) introduce aspect-based sentiment analysis for explainable finance, demonstrating that decomposing financial analysis into specific aspects (similar to our agent specialization) improves both performance and interpretability. The self-improvement capabilities of LLMs through knowledge detection (Jie et al., 2024a) suggest potential enhancements to our framework. While our current implementation uses fixed agent roles, future iterations could incorporate self-training mechanisms where agents learn from successful report generations to refine their specialized capabilities.

Finally, the challenge of structuring unstructured financial data, as addressed by Sun et al. (2024) in the context of ESG reports, parallels our task of converting free-form earnings call transcripts into structured investment reports. Their information extraction techniques could be integrated into our Analyst agent to enhance its ability to systematically extract and organize financial information.

These works collectively underscore that explainability and interpretability are not merely desirable features but essential requirements for financial AI systems. Our multi-agent framework inherently promotes explainability through its transparent workflow: each agent's contribution is distinct and auditable, the iterative refinement process is traceable, and the use of external tools for validation provides grounded explanations for financial claims.

2.2 Earnings Call Transcript Analysis

Researchers have explored transcript data for a variety of downstream tasks. For example, Sawhney et al. (Sawhney et al., 2021) examined bias in multimodal EC analysis for volatility prediction, while Keith and Stent [2] modelled analyst decision-making using semantic features of EC discourse. These studies highlight the predictive power of managerial language and financial context in shaping market outcomes.

Post-EC, two types of reports typically surface: journalistic summaries, which summarises headline figures and key takeaways into concise narratives, and analytical (equity research) reports, which offer a considerably more extensive evaluation of financial performance, managerial tone, and strategic implications for investment strategies (Vipond, 2024) (AlphaSense, 2025).

Although previous research has focused on automating journalistic summary (Mukherjee et al., 2022), automatic generation of analytical reports remains underexplored. By automating this complex output, we could significantly reduce an analyst's workload to allow timely dissemination of insights to investors, and improve the overall scalability of equity research. Hence, this gap motivates the exploration of emerging AI methods, such as Generative AI, to transform earnings call transcripts into structured, actionable equity research reports.

2.3 Generative AI in Financial Analysis

The financial sector has witnessed growing adoption of generative AI for analyzing complex textual documents. Recent studies demonstrate LLMs' strong performance in summarization, question answering, and sentiment extraction from corporate earnings calls, 10-K filings, and analyst briefings (Yang et al., 2024a; Chowdhery et al., 2023; Touvron et al., 2023). These models identify subtle language cues correlating with market movements, often outperforming human analysts in specific prediction scenarios (Hu et al., 2018).

Domain-adapted models further illustrate the benefits of financial corpus training. BloombergGPT (Wu et al., 2023) achieves state-of-the-art performance in sentiment analysis and entity recognition, while FinGPT (Yang et al., 2023) demonstrates that open-source fine-tuned models can rival proprietary approaches on financial NLP benchmarks. However, hallucination remains a primary concern, where models fabricate plausible but inaccurate statements—particularly problematic in high-stakes settings where small factual errors can mislead investors (Kang and Liu, 2023). This motivates research into retrieval-augmented generation (RAG) that constrains LLM outputs with reliable external data (Gao et al., 2024).

2.4 Single-Agent AI Systems

AI agents extend LLMs into autonomous, goal-directed entities that operate more like human workers (Park et al., 2023; Sumers et al., 2023). These systems incorporate planning capabilities for multi-step actions, memory mechanisms for context maintenance, and tool use for accessing external resources (Parisi et al., 2022). This design enables agents to retrieve data through APIs, compute metrics, and generate fact-grounded reports rather than relying on speculative language (Yu et al., 2023).

Recent frameworks have operationalized these concepts successfully. GPT-Engineer demonstrated LLM-driven software generation (Qian et al., 2023), while Toolformer showed that LLMs can self-learn API usage (Schick et al., 2023). In finance, FinMem introduced layered memory architecture enhancing trading agents' decision-making (Yu et al., 2023). However, single-agent systems face limitations as task complexity increases, remaining vulnerable to hallucinations and struggling with self-error correction without external feedback (Darwish et al., 2025).

2.5 Multi-Agent Systems and Role Specialization

To address single-agent limitations, recent research explores multi-agent systems where multiple LLMs interact under role-specific instructions (Hong et al., 2024; Qian et al., 2023; Du et al., 2025). The premise draws on collective intelligence: specialized agent groups can outperform single generalist models when roles and workflows are well-defined (Zhang et al., 2024; Salve et al., 2024; Lu et al., 2023; Yang et al., 2024b). Several frameworks demonstrate this approach, e.g., MetaGPT (Hong et al., 2024) and ChatDev (Qian et al., 2023). Each role activates domain-relevant behavior in the underlying LLM, producing more reliable outputs than single-agent prompting. Research highlights the effectiveness of iterative feedback and debate structures where agents critique and refine each other's work (Darwish et al., 2025).

Emerging literature applies multi-agent systems specifically to finance. Heterogeneous designs focusing on different error types improve financial sentiment analysis (Darwish et al., 2025). TradeMaster illustrates how reinforcement learning and multi-agent collaboration combine for quantitative trading (Sun et al., 2023). However, empirical exploration remains limited, with few systematic comparisons between single-agent and multi-agent approaches for financial analysis tasks.

3 Methodology

3.1 System Architecture Overview

We developed a multi-agent collaborative system (Figure 1) which is powered by OpenAI's GPT-4.1 as the underlying language model to ensure consistent capability across experiments, differing only in their approach to role specialization and agent interaction.

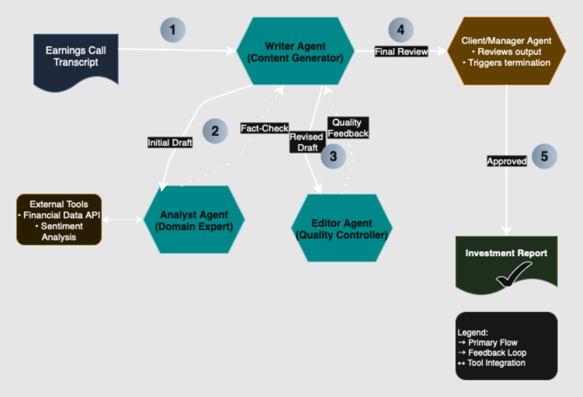


Figure 1: Multi-agent architecture for Investment Guidance

3.2 Multi-Agent Configuration

The multi-agent system employs three specialized agents collaborating through structured workflows:

Agent	Responsibilities
Analyst	Extracts financial data, performs fact-checking, calculates ratios, conducts sentiment analysis using external tools
Writer	Drafts and revises investment reports incorporating Analyst data and Editor feedback, maintains professional tone and structure
Editor	Reviews drafts for accuracy, completeness, and readability, provides actionable feedback for revisions

Table 1: Multi-agent system roles and responsibilities

Role-specific temperature settings optimize each agent's function: Analyst and Editor operate at 0.2-0.3 for maximum accuracy, while Writer uses 0.7 for natural language fluency. This configuration mimics real-world analyst teams where domain experts, writers, and editors collaborate iteratively.

3.3 Orchestration Framework

We implemented both systems using Microsoft AutoGen[1], an open-source framework for multi-agent LLM applications. AutoGen manages agent communication through its GroupChatManager, routing messages appropriately between agents. The framework handles message passing and turn-taking logic, simplifying implementation of iterative feedback loops.

Algorithm 1 Multi-Agent Workflow

```
**Input:** Earnings call transcript $T$, Instructions $I$
Writer creates initial draft $D_0$ from $T$
Analyst validates $D_0$ against external data:
    Extract metrics, calculate ratios
    Call external APIs for validation
    Generate structured feedback $F_A$
Writer revises draft: $D_1 = \text{revise}(D_0, F_A)$
**repeat**
    Editor reviews $D_i$ for quality
        Check factual consistency
        Evaluate completeness and clarity
        Generate editorial feedback $F_E$
    Writer produces $D_{i+1} = \text{revise}(D_i, F_E)$
**until** Editor approves or max iterations reached

Client validates final draft $D_n$
**Output:** Investment report $D_n$
```

3.4 Dataset and Preprocessing

We used the official Earnings2Insights(Takayanagi et al., 2025)dataset released for the shared task. The dataset consists of 64 earnings call transcripts drawn from two subsets:

- **ECTSum subset (40 transcripts)**: aligned with the ECTSum benchmark, where each folder includes both a transcript and a reference summary. Participants may optionally leverage these summaries.
- **Professional subset (24 transcripts)**: matched with professional analyst reports. Only the transcripts are provided to participants; comparisons with analyst reports are conducted later by the organizers.

The transcripts were distributed in Markdown format, already structured with speaker metadata and sections (e.g., management remarks, Q&A). Since the files were ready for direct LLM ingestion, no additional preprocessing was required. Each transcript was processed by multi-agent systems to generate reports in JSON format.

3.5 External Tools and APIs

Both configurations access two specialized tools:

- **historicalFinancialData(ticker, year, quarter):** Retrieves quarterly metrics (EPS, revenue, cash flow, balance sheet) from Alpha

[1] https://github.com/microsoft/autogen

Vantage for year-over-year and quarter-over-quarter comparisons

- **analyzeMarketSentiment(ticker, year, quarter):** Collects news articles published within 30 days before the earnings call, ensuring realistic temporal constraints matching real analyst workflows

3.6 Evaluation Framework

3.6.1 Automatic Evaluation

Our evaluation follows the official shared task protocols from the Earnings2Insights competition (Takayanagi et al., 2025) where reports were evaluated automatically using an LLM-based judge following standardized guidelines:

- **Average Likert Score**: mean 1–7 rating across Persuasiveness, Logic, Usefulness, Readability, and Clarity.

- **Win Rate vs Analyst Report**: pairwise comparison against professional analyst reports, where win rate = Wins ÷ (Wins + Losses).

3.6.2 Human Evaluation

The organizers also conducted a large-scale human evaluation with 210 participants on the Prolific platform (176 retained after attention checks). Each participant reviewed 12 reports and two measures were collected:

- **Accuracy of financial decisions**: fraction of correct Buy/Neutral/Sell predictions, evaluated at day, week, and month horizons, then averaged.

- **Human Likert Scores**: 7-point scores for clarity, logic, persuasiveness, readability, and usefulness.

Together, these evaluations provide a rigorous test of system performance. Automatic scoring offers a scalable baseline, while human evaluation captures how well reports can actually guide and persuade investors in practice. This dual framework ensures that the final rankings reflect both the formal quality of the report and the impact of real-world decision making.

4 Discussion

Our experiments demonstrate that multi-agent collaboration significantly enhances the quality of AI-generated financial analysis. The multi-agent system achieved the highest financial decision accuracy (58.1%) among automated approaches. The human evaluation reveals interesting patterns in perceived quality versus actual utility. While some systems scored higher on individual Likert dimensions, our multi-agent approach achieved a balanced performance across all metrics, with particularly strong scores in logic (5.89) and persuasiveness (5.95). The correlation between Likert scores and decision accuracy suggests that report clarity and logical structure directly impact investment decision quality.

The Analyst agent's integration of external data proved particularly valuable, reducing hallucinations and ensuring quantitative claims align with verifiable sources. The Editor's quality control function, while contributing less to raw accuracy, substantially improved report professionalism and readability—critical factors for real-world deployment.

4.1 Comparison with Human Analysts

While our model achieves 52.2% decision accuracy, placing it in the middle tier of evaluated systems, several fundamental distinctions from human analysis remain:

- Limited ability to incorporate non-textual market signals or conduct primary research

- Absence of industry-specific intuition developed through years of experience

- Difficulty identifying subtle management communication patterns that experienced analysts recognize

- Inability to leverage professional networks for channel checks or proprietary information

The best-performing systems in our evaluation achieved average accuracy around 58%, with none exceeding 60%, suggesting a practical ceiling for current LLM-based approaches when relying solely on earnings transcript analysis. This performance gap underscores that AI systems should augment rather than replace human analysts. Our system's balanced scores for logic (5.61/7) and usefulness (5.74/7) indicate that while the system may not achieve top-tier decision accuracy, it produces reports that provide valuable foundational analysis for human refinement.

5 Conclusion

This paper presented a novel multi-agent framework for automated financial analysis that demonstrably improves investment decision-making. Through rigorous human evaluation with 176 participants, we showed that our multiagent system achieves 58. 1% accuracy in guiding financial decisions. The system also received strong quality ratings, with scores of 5.89/7 for logical structure and 5.95/7 for persuasiveness, indicating that structured LLM collaboration can address critical limitations of monolithic approaches.

Our contributions advance the field of financial NLP by providing empirical evidence for multiagent superiority in complex analytical tasks. Ultimately, we envision multi-agent systems becoming integral to institutional investment processes, enhancing human decision-making while preserving the judgment and intuition that remain uniquely human contributions to financial analysis. The human evaluation results demonstrate that role specialization not only improves technical metrics but also translates to better investment outcomes, a critical validation that is often missing from AI research. The relationship between the quality dimensions of the report and the accuracy of the decisions provides actionable insights for the design of the system, suggesting that the logical structure and persuasiveness are key factors in generating actionable financial intelligence.

- Our model produces functionally useful reports that prioritize actionable insights over surface-level quality.
- The multi-agent orchestration may be creating overly balanced perspectives that excel at weekly horizons but miss immediate market signals.
- The gap between automatic (4.575) and human (5.49) Likert scores indicates our system's value is better recognized by human readers than automated evaluators.

Acknowledgments

This research/project is supported by the Ministry of Education, Singapore under its MOE Academic Research Fund Tier 2 (MOE-T2EP20123-0005) and by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore. The authors would like to thank the annotators for their careful work in reviewing our reports and providing scores.

References

AlphaSense. 2025. Equity research reports: A complete guide. https://www.alpha-sense.com/resources/equity-research-guide/. AlphaSense.

Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. 2021. Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1):1–74.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Ahmed M Darwish, Ehab A Rashed, and Gasser Khoriba. 2025. Mitigating llm hallucinations using a multi-agent framework. *Information*, 16(7):517.

Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria. 2024a. An evaluation of reasoning capabilities of large language models in financial sentiment analysis. In *IEEE Conference on Artificial Intelligence*, pages 189–194, Singapore.

Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria. 2024b. Financial sentiment analysis: Techniques and applications. *ACM Computing Surveys*, 56(9):220.

Kelvin Du, Yazhi Zhao, Rui Mao, Frank Xing, and Erik Cambria. 2025. A retrieval-augmented multi-agent system for financial sentiment analysis. *IEEE Intelligent Systems*, 40(2):15–22.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, and 1 others. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. International Conference on Learning Representations, ICLR.

Ziniu Hu, Weiqing Liu, Jiang Bian, Xuanzhe Liu, and Tie-Yan Liu. 2018. Listening to chaotic whispers: A deep learning framework for news-oriented stock trend prediction. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 261–269.

Yeo Wei Jie, Teddy Ferdinan, Przemyslaw Kazienko, Ranjan Satapathy, and Erik Cambria. 2024a. Self-training large language models through knowledge detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15033–15045.

Yeo Wei Jie, Ranjan Satapathy, and Erik Cambria. 2024b. Plausible extractive rationalization through semi-supervised entailment signal. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5182–5192.

Yeo Wei Jie, Ranjan Satapathy, Rick Goh, and Erik Cambria. 2024c. How interpretable are reasoning explanations from prompting large language models? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2148–2164.

Haoqiang Kang and Xiao-Yang Liu. 2023. Deficiency of large language models in finance: An empirical examination of hallucination. *arXiv preprint arXiv:2311.15548*.

Tim Loughran and Bill McDonald. 2016. Textual analysis in accounting and finance: A survey. *Journal of accounting research*, 54(4):1187–1230.

Jie Lu, Anjin Liu, Fan Dong, Feng Gu, Joao Gama, and Guangquan Zhang. 2018. Learning under concept drift: A review. *IEEE Transactions on Knowledge and Data Engineering*, 31(12):2346–2363.

Pan Lu, Baolin Peng, Hao Cheng, Michel Galley, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, and Jianfeng Gao. 2023. Chameleon: Plug-and-play compositional reasoning with large language models. *arXiv preprint arXiv:2304.09842*.

Rajdeep Mukherjee, Abhinav Bohra, Ananya Banerjee, Sopan Khosla Sharma, Madhav Hegde, Asif Ekbal Shaikh, Saurabh Shrivastava, Kaushik Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 2022. Ectsum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Keane Ong, Wihan Van Der Heever, Ranjan Satapathy, Erik Cambria, and Gianmarco Mengaldo. 2023. Finxabsa: Explainable finance through aspect-based sentiment analysis. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 773–782. IEEE.

Aaron Parisi, Yao Zhao, and Noah Fiedel. 2022. Talm: Tool augmented language models. *arXiv preprint arXiv:2205.12255*.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Chen Qian, Xin Cong, Chenglong Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*.

Aniruddha Salve, Saurabh Attar, Mandar Deshmukh, Shashank Shivpuje, and Ujjwal A Mitra. 2024. A collaborative multi-agent approach to retrieval-augmented generation across diverse data. *arXiv preprint arXiv:2412.05838*.

Ravinder Singh Sawhney, A. Aggarwal, and Rajiv Ratn Shah. 2021. An empirical investigation of bias in the multimodal analysis of financial earnings calls. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

Theodore R Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L Griffiths. 2023. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*.

Shuo Sun, Molei Qin, Wentao Zhang, Haochong Xia, Chuqiao Zong, Jie Ying, Yonggang Xie, Lingxuan Zhao, Xinrun Wang, and Bo An. 2023. Trademaster: A holistic quantitative trading platform empowered by reinforcement learning. *Advances in Neural Information Processing Systems*, 36:59047–59061.

Zounachuan Sun, Ranjan Satapathy, Daixue Guo, Bo Li, Xinyuan Liu, Yangchen Zhang, Cheng-Ann Tan, Ricardo Shirota Filho, and Rick Siow Mong Goh. 2024. Information extraction: Unstructured to structured for esg reports. In *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 487–495. IEEE.

Takehiro Takayanagi, Tomas Goldsack, Kiyoshi Izumi, Chenghua Lin, Hiroya Takamura, and Chung-Chi Chen. 2025. Earnings2Insights: Analyst Report Generation for Investment Guidance. In *Proceedings of the FinNLP Workshop at EMNLP 2025*, Suzhou, China. Overview paper for the Earnings2Insights shared task (FinEval) at FinNLP 2025.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

T. Vipond. 2024. Equity research report: A recommendation to buy, sell, or hold shares of a public company. `https://corporatefinanceinstitute.com/resources/valuation/equity-research-report/`. Corporate Finance Institute.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023.

Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.

Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023. Fingpt: Open-source financial large language models. *arXiv preprint arXiv:2306.06031*.

Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Shaochen Zhong, Bing Yin, and Xia Hu. 2024a. Harnessing the power of llms in practice: A survey on chatgpt and beyond. *ACM Transactions on Knowledge Discovery from Data*, 18(6):1–32.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024b. Large language models for automated open-domain scientific hypotheses discovery. In *Proceedings of ACL*, pages 13545–13565.

Wei Jie Yeo, Ranjan Satapathy, and Erik Cambria. 2025a. Towards faithful natural language explanations: A study using activation patching in large language models. In *In Proceedings of EMNLP*.

Wei Jie Yeo, Wihan Van Der Heever, Rui Mao, Erik Cambria, Ranjan Satapathy, and Gianmarco Mengaldo. 2025b. A comprehensive review on financial explainable ai. *Artificial Intelligence Review*, 58(6):1–49.

Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Denghui Zhang, Rong Liu, Jordan W Suchow, and Khaldoun Khashanah. 2023. Finmem: A performance-enhanced llm trading agent with layered memory and character design. *arXiv preprint arXiv:2311.13743*.

Wentao Zhang, Lingxuan Shen, Jingwei Yang, Haoran Zhu, Zhangyang Chen, Dongyan Sun, Jiazheng Zhang, Xiaodong Li, and Yuxuan Zhang. 2024. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. *arXiv preprint arXiv:2402.18485*.

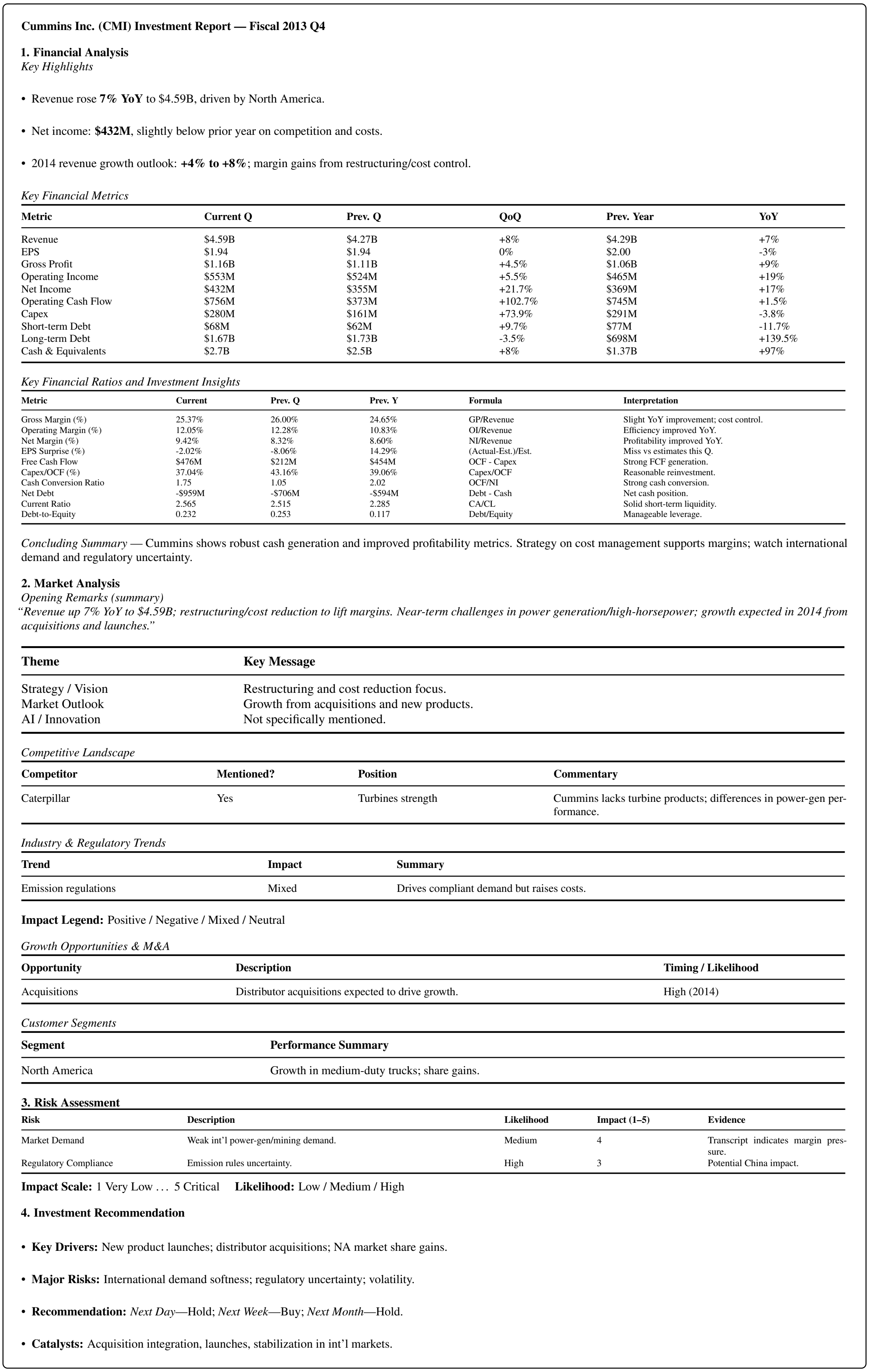
Appendix

A Agent Initialization prompts

This section shares agent initialization prompts that we have used in the experiments.

Agent	Responsibilities
Writer	You are the Writer. Draft the investment report and revise it based on other agents' feedback. Do not rewrite from scratch unless asked; make targeted edits. Always return the updated full report only (Markdown). Populate sections from the transcript; use tables for structured data; replace placeholders when Analyst/Editor provide updates.
Analyst	You are the Analyst. Fact-check and correct financial metrics and ratios using tool data (e.g., @historicalfinancialdata(ticker,year,quarter)). Compute QoQ/YoY, fill missing values with N/A, and provide: (1) FINANCIAL ANALYSIS UPDATE, (2) FINANCIAL RATIOS UPDATE, (3) KEY HIGHLIGHTS UPDATE, and (4) FINANCIAL ANALYSIS SUMMARY UPDATE. Fetch news via @analyzemarketsentiment and supply a complete News Sentiment section or instruct omission if none. Hand off to Editor after updates.
Editor	You are the Editor. Review Sections 1–3 for completeness, clarity, structure, and consistency (tables, legends, formatting). Ensure Analyst updates and sentiment are integrated; remove any internal notes/placeholders. Produce an INVESTMENT RECOMMENDATION FEEDBACK block containing: Key Drivers, Major Risks, and Buy/Hold/Sell calls for Next Day/Week/Month with data-backed justifications and Catalysts. The Writer must update Section 4 accordingly.
Client	You are the Client/Investor. Review the latest Writer report against the checklist (sections, metrics, ratios, highlights, summary, risks, editor feedback integration, formatting). If all checks pass, reply TERMINATE; otherwise list failed checks with exact fixes required.

Table 2: Agent initialization prompts.



Cummins Inc. (CMI) Investment Report — Fiscal 2013 Q4

1. Financial Analysis
Key Highlights

- Revenue rose **7% YoY** to $4.59B, driven by North America.
- Net income: **$432M**, slightly below prior year on competition and costs.
- 2014 revenue growth outlook: **+4% to +8%**; margin gains from restructuring/cost control.

Key Financial Metrics

Metric	Current Q	Prev. Q	QoQ	Prev. Year	YoY
Revenue	$4.59B	$4.27B	+8%	$4.29B	+7%
EPS	$1.94	$1.94	0%	$2.00	-3%
Gross Profit	$1.16B	$1.11B	+4.5%	$1.06B	+9%
Operating Income	$553M	$524M	+5.5%	$465M	+19%
Net Income	$432M	$355M	+21.7%	$369M	+17%
Operating Cash Flow	$756M	$373M	+102.7%	$745M	+1.5%
Capex	$280M	$161M	+73.9%	$291M	-3.8%
Short-term Debt	$68M	$62M	+9.7%	$77M	-11.7%
Long-term Debt	$1.67B	$1.73B	-3.5%	$698M	+139.5%
Cash & Equivalents	$2.7B	$2.5B	+8%	$1.37B	+97%

Key Financial Ratios and Investment Insights

Metric	Current	Prev. Q	Prev. Y	Formula	Interpretation
Gross Margin (%)	25.37%	26.00%	24.65%	GP/Revenue	Slight YoY improvement; cost control.
Operating Margin (%)	12.05%	12.28%	10.83%	OI/Revenue	Efficiency improved YoY.
Net Margin (%)	9.42%	8.32%	8.60%	NI/Revenue	Profitability improved YoY.
EPS Surprise (%)	-2.02%	-8.06%	14.29%	(Actual-Est.)/Est.	Miss vs estimates this Q.
Free Cash Flow	$476M	$212M	$454M	OCF - Capex	Strong FCF generation.
Capex/OCF (%)	37.04%	43.16%	39.06%	Capex/OCF	Reasonable reinvestment.
Cash Conversion Ratio	1.75	1.05	2.02	OCF/NI	Strong cash conversion.
Net Debt	-$959M	-$706M	-$594M	Debt - Cash	Net cash position.
Current Ratio	2.565	2.515	2.285	CA/CL	Solid short-term liquidity.
Debt-to-Equity	0.232	0.253	0.117	Debt/Equity	Manageable leverage.

Concluding Summary — Cummins shows robust cash generation and improved profitability metrics. Strategy on cost management supports margins; watch international demand and regulatory uncertainty.

2. Market Analysis
Opening Remarks (summary)
"Revenue up 7% YoY to $4.59B; restructuring/cost reduction to lift margins. Near-term challenges in power generation/high-horsepower; growth expected in 2014 from acquisitions and launches."

Theme	Key Message
Strategy / Vision	Restructuring and cost reduction focus.
Market Outlook	Growth from acquisitions and new products.
AI / Innovation	Not specifically mentioned.

Competitive Landscape

Competitor	Mentioned?	Position	Commentary
Caterpillar	Yes	Turbines strength	Cummins lacks turbine products; differences in power-gen performance.

Industry & Regulatory Trends

Trend	Impact	Summary
Emission regulations	Mixed	Drives compliant demand but raises costs.

Impact Legend: Positive / Negative / Mixed / Neutral

Growth Opportunities & M&A

Opportunity	Description	Timing / Likelihood
Acquisitions	Distributor acquisitions expected to drive growth.	High (2014)

Customer Segments

Segment	Performance Summary
North America	Growth in medium-duty trucks; share gains.

3. Risk Assessment

Risk	Description	Likelihood	Impact (1–5)	Evidence
Market Demand	Weak int'l power-gen/mining demand.	Medium	4	Transcript indicates margin pressure.
Regulatory Compliance	Emission rules uncertainty.	High	3	Potential China impact.

Impact Scale: 1 Very Low … 5 Critical **Likelihood:** Low / Medium / High

4. Investment Recommendation

- **Key Drivers:** New product launches; distributor acquisitions; NA market share gains.
- **Major Risks:** International demand softness; regulatory uncertainty; volatility.
- **Recommendation:** *Next Day*—Hold; *Next Week*—Buy; *Next Month*—Hold.
- **Catalysts:** Acquisition integration, launches, stabilization in int'l markets.

Figure 2: Full example of a report generated with all agents.