Agentic LLMs for Analyst-Style Financial Insights: An LLM Pipeline for Persuasive Financial Analysis

Gaurangi Sinha[1] **Rajarajeswari Palacharla**[1] **Manoj Balaji Jagadeeshan**[2,3]

[1]Department of Computer Science and Engineering, Texas A&M University
[2]Department of Computer Science and Engineering, Indian Institute of Technology, Kharagpur
[3]Hewlett Packard Enterprise (HPE), India

Correspondence: gaurangisinha@tamu.edu

Abstract

This paper presents our approach for the Earnings2Insights 2025 shared task, which focuses on generating a persuasive financial analysis report from earnings call transcripts. The FinNLP challenge required changing lengthy, unstructured earnings call text into concise, analyst-style insights and investment recommendations. We developed an approach, as described in this paper, that utilizes a multistage LLM-based pipeline to ensure both factual accuracy and narrative quality. First, we used a large language model (LlaMA3-70B) in an extractive summary step to capture key financial metrics and the details of the transcript guidance. We then fed these structured insights into a generative LLM to produce a comprehensive research report evaluating the company's performance, highlighting bullish/bearish signals, assessing risks, and providing clear long/short recommendations over short-term goals. To further enhance the quality of these summaries, we incorporate an LLM-driven self-evaluation loop. This strategy addresses the task criteria of persuasiveness, logic, usefulness, readability, and clarity. We (Team name: *SigJBS*), through our method in the official evaluation, achieved an average Likert score of 4.60 (out of 7) and a 52.6% win rate against professional analyst reports, demonstrating the effectiveness of the proposed approach in generating high-quality financial insights.

1 Introduction

Artificial Intelligence is transforming the way we work, automating repetitive tasks and even helping us make complex decisions. Yet despite these breakthroughs, many industries still haven't tapped into AI's full potential. Constraints around compute power, adapting models to specialized fields, and worries about reliability and safety often stand in the way.

In the financial sector, earnings call transcripts represent a critical source of information [1] for analysts, as they combine quantitative metrics with qualitative insights from corporate leadership. Numerous studies have explored the potential of large language models (LLMs) to generate investment recommendations from these transcripts. Yet the quality and persuasiveness of AI-generated reports remain below professional standards (Goldsack et al., 2025)(Hu et al., 2025). This gap is especially significant in the context of the Earnings2Insights shared task (Takayanagi et al., 2025a), which requires participants to generate investment guidance directly from earnings calls, with evaluation based on human investment decisions rather than traditional similarity-based metrics (Huang et al., 2025). In this work, we propose a agentic AI framework for investment report generation. Our approach employs three specialized agents: a summarization agent, which applies hierarchical fragmentation to extract structured financial milestones; a reasoning agent, which synthesizes these signals into investment theses and risk assessments; and a critique agent, which evaluates and refines candidate reports to ensure persuasiveness and decision relevance. By decomposing the task into modular stages, the system improves both the factual foundation and alignment with investor decision-making needs.

This study makes four key contributions:

- An agentic framework for financial decision support designed for the analysis of earnings call transcripts.
- The integration of retrieval-augmented generation and online search capabilities to improve contextual awareness in investment reasoning.
- A comparative evaluation of chunking strategies, highlighting the effectiveness of hierarchical chunking for context retention.
- A novel AI-based evaluation setup, where an

[1]Our code is available on our Github page SigJBS

Proceedings of The 10th Workshop on Financial Technology and Natural Language. EMNLP-2025, Suzhou, China

agent acts as a judge to assess quality and consistency of generated reports.

2 Related Work

Automated summarization and analysis of financial discourse have gained significant attention in recent years. Mukherjee et al. (2022) introduced ECTSum, a benchmark of 40 long-form earnings call transcripts paired with expert bullet-point summaries. Their work highlighted the challenge of distilling detailed Q&A dialogue into concise and factually consistent takeaways. Around the same time, Liu et al. (2022) released FINDSum, which comprises more than 21,000 annual reports with human-written summaries, and demonstrated how the joint modeling of narrative text and tabular data improves the extraction of key numeric facts. Chang et al. (2024) systematically explored book-length summarization with LLMs, highlighting hierarchical and multistage techniques that inspired our hierarchical chunk summarization agent (Chang et al., 2024).

More recently, large language models (LLMs) have been fine-tuned and evaluated for generating financial reports. BloombergGPT (Wu et al., 2023) and FinGPT (Wang et al., 2023) are two notable efforts to adapt general LLM architectures to finance-specific corpora, supporting tasks from question answering to narrative summarization. Yang et al. (2023) further demonstrated that a 65 billion-parameter model, InvestLM, when instructed according to analyst-style instructions, can produce investment notes of comparable quality to those of GPT-4 in expert evaluations. Takayanagi et al. (2025b) took this step further by demonstrating that GPT-4-generated stock commentaries can actually influence real investor decisions, underscoring both the power and responsibility of LLM-based analyses.

Despite these advances, ensuring numerical accuracy remains a hurdle. Standard metrics like ROUGE often miss errors in critical figures, prompting the SemEval 2024 NumEval challenge (Chen et al., 2024), which evaluated the model's ability to preserve and generate correct numerical values in tasks such as headline generation. In parallel, (Huang et al., 2025) proposed a decision-oriented evaluation framework: instead of measuring surface overlap, they judged summaries by their impact on model or human trading performance. This approach aligns closely with the goals of Earnings2Insights (Takayanagi et al., 2025a), where success is defined by whether a generated report leads readers to the right investment choices.

Our work builds on these strands by combining an extractive summarization stage, anchored in the ECTSum framework, with a generative LLM pipeline that produces full-blown analyst-style notes. Crucially, we adopt a decision-driven evaluation, asking annotators to make hypothetical trades based on our reports. In doing so, we hope to contribute both a practical method for high-fidelity financial analysis and a rigorous way to assess its real-world utility.

3 Methodology

3.1 Dataset

We adopt the ECTSum dataset (Mukherjee et al., 2022) as the basis for our experiments. In ECTSum, there are 40 earnings call transcripts, each accompanied by a reference summary that provides ground-truth information for milestone extraction. Additionally, the Professional subset contains 24 transcripts matched with professional analyst reports; only the raw transcripts are provided, and comparison with the professional reports is reserved for evaluation by the shared-task organizers. In total, all 64 transcripts are processed to extract key financial milestones and generate investment guidance using our agentic LLM framework.

3.2 Experiment Design

Our experimental procedure consists of two main steps, followed by an evaluation stage. In Step 1, the Summarization Agent extracts structured financial milestones from each transcript. We tested various prompt formulations to optimize extraction accuracy, and we selected a final prompt that offered consistent and comprehensive coverage of key financial events. In Step 2, the Reasoning Agent uses these extracted milestones to generate an investment recommendation for each company. After the recommendation is generated, an independent Critique Agent evaluates the draft report and assigns a confidence score reflecting the LLM's certainty in the recommendation's correctness and persuasiveness. Additionally, we manually spot-checked a subset of the generated reports to identify common trends or issues, which informed further prompt refinement for each agent.

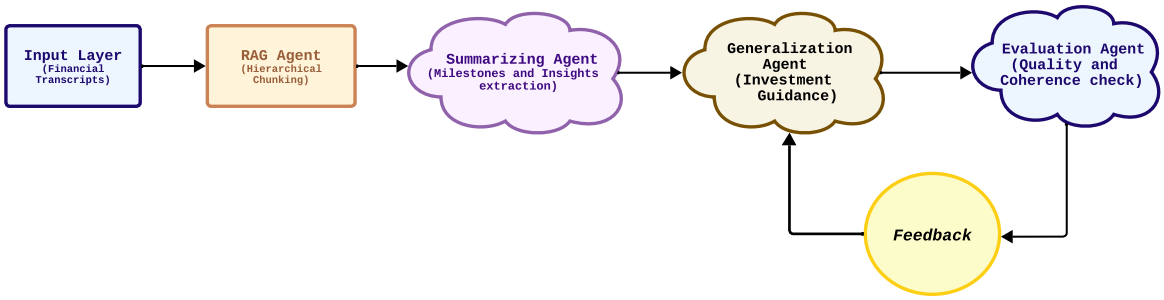

Figure 1: Flowchart of the experimental setup. Overview of the proposed agentic LLM pipeline for generating investment reports from earnings call transcripts. The system sequentially processes transcripts through RAG, Summarization, Generalization, and Evaluation agents, with quality feedback loops to ensure actionable, coherent, and persuasive analyst-style insights.

3.3 Framework

We propose an agentic large language model (LLM) pipeline to generate investment reports from earnings call transcripts. See Figure 1 for our proposed method and experimental setup. The system comprises three sequential agents: Summarization, Reasoning, and Critique, which work in tandem to extract key information, interpret it, and refine the final output.

3.3.1 Summarization Agent

This stage condenses lengthy transcripts into structured financial milestones. Hierarchical chunking is employed to divide transcripts into semantically coherent sections, thereby preserving contextual dependencies. We evaluated several chunking strategies, including fixed-length segmentation and sliding windows, and found that hierarchical chunking consistently provided superior coverage and contextual consistency. The structured output produced at this stage captures key financial attributes such as company name, fiscal period, revenue, earnings per share (EPS), guidance, dividends, and notable events. This structured representation reduces extraneous noise and enables more reliable downstream reasoning.

3.3.2 Reasoning Agent

This agent interprets the extracted milestones to generate actionable investment recommendations. Recommendation generation involves assigning investment stances (Long, Short, Hold), conviction levels (Low, Medium, High), and time horizons (1 day, 1 week, 1 month) based on the structured financial information. Risk and mitigation analysis is incorporated by identifying potential risk factors such as acquisition integration challenges or supply chain constraints and mapping them to corresponding mitigation strategies, thereby contextualizing the recommendations. Finally, timeline aggregation is applied when multiple transcript entries are available across quarters, enabling the system to capture longitudinal trends in company performance and investor guidance.

3.3.3 Critique Agent

The third agent refines candidate reports by employing an independent large language model as a judge. Candidate reports are evaluated according to criteria such as factual consistency, clarity, and persuasiveness. To improve quality, multiple prompt variations were tested, and feedback from the critique agent was incorporated to iteratively refine the reports until a satisfactory version was achieved. The final output consists of structured recommendations, key positives, time-specific performance drivers, and identified risk-mitigation factors.

3.4 Models Used

The framework primarily employs the LLaMA3 70B (llama3-70b-8192) (Grattafiori et al., 2024) model accessed via the Groq API[2]. The Summarization Agent leverages the model for milestone extraction, while the Reasoning Agent generates investment recommendations based on the structured data. The Critique Agent uses a separate instance of an LLM to evaluate report quality and provide iterative feedback.

3.5 Evaluation

The evaluation of system outputs was conducted through two primary phases: an internal automatic

[2] https://console.groq.com

evaluation for iterative development and the official shared task human evaluation for final assessment.

3.5.1 Internal Automatic Evaluation

During development, we employed an LLM-based critique agent to enable rapid iteration. This agent assessed each generated report based on key criteria including factual consistency (alignment with the source transcript), logical coherence (soundness of the argument from data to recommendation), and persuasiveness (clarity and strength of the investment thesis). This automated feedback loop was crucial for refining our prompts and improving the performance of the summarization and reasoning agents.

3.5.2 Official Shared Task Evaluation

The final ranking was determined by a human evaluation study. Annotators made ternary investment decisions (Long, Short, or Neutral) for three time horizons based on the generated reports. The primary ranking metric was decision accuracy, defined as the proportion of correct directional predictions to all non-Neutral decisions, averaged across the three horizons.

4 Results and Discussion

Our agentic framework was applied to all 64 earnings call transcripts from the provided ECTSum and Professional subsets. The system successfully generated structured analyst reports for each instance, comprising extracted financial milestones, a concrete investment recommendation (Long, Short, or Neutral), and a supporting rationale derived from the transcript data.

4.1 Official Human Evaluation Performance

The official evaluation, based on the accuracy of investment decisions made by human annotators after reading our reports, yielded the following results:

Average	1-Day	1-Week	1-Month
0.545	0.609	0.513	0.512

Table 1: Investment decision accuracy based on human evaluation.

As shown in Table 1, our framework achieved a mean decision accuracy of 0.545 in the official human evaluation, securing 4th place in the final shared task ranking. This result indicates that the investment decisions guided by our reports were correct 54.5% of the time on average across all evaluated time horizons. Performance was most robust at the one-day horizon (60.9% accuracy), suggesting that our method was particularly effective at identifying the immediate market catalysts and salient insights within the earnings calls. The accuracy across all horizons remained consistently above chance, demonstrating the practical utility of the system for short-term investment guidance.

In addition to decision accuracy, human evaluators assessed the reports on several qualitative dimensions using a 7-point Likert scale.

Metric	Score	Metric	Score
Clarity	5.76	Readability	5.61
Logic	5.68	Usefulness	5.72
Persuasiveness	5.59	Avg.	5.67

Table 2: Human Likert Ratings for Report Quality (1–7 Scale).

The human evaluation yielded strong qualitative ratings for our reports, with an overall average score of 5.67/7. Our submission, as shown in Table 2, received its highest scores in Usefulness (5.72) and Logic (5.68), indicating that the generated reports were found to be particularly actionable and well-reasoned for investment purposes.

4.2 Automatic Evaluation Correlation

The official automatic evaluation results, which employed an LLM-as-a-judge protocol, provide a preliminary assessment of report quality. Our submission achieved an average score of 4.597 on a 7-point Likert scale across several qualitative dimensions. In a comparative pairwise evaluation, the LLM judge preferred our generated reports over those written by professional financial analysts in 52.6% of instances. These automatic metrics suggest our framework produces outputs that are competitive with expert-authored content in terms of perceived quality and persuasiveness.

4.3 Error Analysis

To improve our agentic LLM pipeline, we conducted a manual review of generated reports and focused on three primary error categories:

1. **Hallucinations:** Occasionally the model invented figures or details not present in the source transcript. To address this, we reinforced our call to insist on 'using only the data provided', and added a post-generation check that flags any numeric value not appearing in the structured summary.

2. **Missing Fields:** Some required sections (e.g., the confidence score from the Critique Agent) were sometimes omitted. We revised the prompt to explicitly request every field and to output "n/a" when a value is unavailable, guaranteeing complete coverage.

3. **Formatting Drift:** Early outputs included extraneous phrases (e.g., "Here is the note:") or stray markdown characters. We enforced a strict output template in the prompt, 'output only the numbered headings and bullet points, with no extra text', which eliminated filler language and ensured a uniform, professional format.

After each prompt revision, we spot-checked a random subset of 20 reports to verify that hallucinations were reduced, all sections were present, and formatting was consistent. This iterative loop of *identify–revise–re-evaluate* produced stable improvements in factual fidelity, completeness, and style across our 64 earnings-call reports.

5 Limitations and Future Work

Our pipeline, even if it is effective, still faces challenges. It can hallucinate unsupported figures despite data-only prompts, and confidence scores from the Critique Agent are not consistently reported, reducing transparency. Furthermore, relying on zero-shot prompts without domain-specific fine-tuning can limit distinct financial reasoning.

For future work, we plan to fine-tune each agent on financial texts to remove hallucinations, integrate real-time financial data to ground analyses in current facts, and implement a more meticulous uncertainty estimate to accompany recommendations made by these models. These enhancements should improve the factual reliability and user trust of our earnings call reports.

6 Conclusion

Our agentic LLM pipeline, combining targeted extraction, reasoned recommendation, and automated critique, proved both practical and persuasive in the Earnings2Insights shared task, delivering above-chance decision accuracy (54.5% overall, 60.9% one day) and strong human ratings (5.67/7). By iteratively refining prompts and leveraging a self-evaluation loop, we minimized hallucinations and ensured consistency. Moving forward, we'll integrate richer financial context, explore dynamic prompt adaptation, and develop deeper critique agents to further boost both the precision and impact of automated investment reports.

References

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of llms. *Preprint*, arXiv:2310.00785.

Chung-chi Chen, Jian-tao Huang, Hen-hsen Huang, Hiroya Takamura, and Hsin-hsi Chen. 2024. SemEval-2024 task 7: Numeral-aware language understanding and generation. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1482–1491, Mexico City, Mexico. Association for Computational Linguistics.

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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yebowen Hu, Xiaoyang Wang, Wenlin Yao, Yiming Lu, Daoan Zhang, Hassan Foroosh, Dong Yu, and Fei Liu. 2025. Define: Decision-making with analogical reasoning over factor profiles. *Preprint*, arXiv:2410.01772.

Yu-Shiang Huang, Chuan-Ju Wang, and Chung-Chi Chen. 2025. Decision-oriented text evaluation. *Preprint*, arXiv:2507.01923.

Shuaiqi Liu, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. 2022. Long text and multi-table summarization: Dataset and method. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1995–2010, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and Pawan Goyal. 2022. Ectsum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. *Preprint*, arXiv:2210.12467.

Takehiro Takayanagi, Tomas Goldsack, Kiyoshi Izumi, Chenghua Lin, Hiroya Takamura, and Chung-Chi Chen. 2025a. 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.

Takehiro Takayanagi, Hiroya Takamura, Kiyoshi Izumi, and Chung-Chi Chen. 2025b. Can GPT-4 sway experts' investment decisions? In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 374–383, Albuquerque, New Mexico. Association for Computational Linguistics.

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

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. *Preprint*, arXiv:2303.17564.

Yi Yang, Yixuan Tang, and Kar Yan Tam. 2023. Investlm: A large language model for investment using financial domain instruction tuning. *Preprint*, arXiv:2309.13064.