Meta Prompting for Analyst Report Generation: Turning Earnings Calls into Investment Guidance

Pulkit Chatwal[*] **Mann Bajpai** **Priyanshu**
Harish Pratap Singh **Santosh Kumar Mishra**
Rajiv Gandhi Institute of Petroleum Technology, Jais, India

Abstract

This paper presents our participation in the shared task *Earnings2Insights: Analyst Report Generation for Investment Guidance* at FinNLP @ EMNLP-2025. We develop a large language model (LLM)-based system with agentic prompting, where the model assumes the role of multiple analysts (financial, sentiment, strategic) to generate structured investment reports across day-, week-, and month-level horizons. A self-reflection module is further employed to enhance factual grounding and reduce hallucinations.

In the official evaluation, our system (**Team DataLovers**) ranked **2nd** in financial decision accuracy with average scores of **0.579**, **0.597**, **0.611**, and **0.529** (overall, day, week, and month). Human evaluation placed us **6th**, with average Likert ratings of **5.50** (clarity), **5.56** (logic), **5.45** (persuasiveness), **5.32** (readability), and **5.73** (usefulness), yielding an overall mean of **5.47**. These results highlight the effectiveness of our prompting strategy in producing reports that are both decision-oriented and persuasive, while also revealing challenges in achieving top human evaluation scores.

1 Introduction

The surge of large language models (LLMs) has transformed numerous domains by enabling machines to process, summarize, and generate human-like text with remarkable fluency. In financial contexts, however, generating reliable and actionable insights remains a major challenge due to the complexity, volatility, and domain-specific nature of financial discourse. The shared task *Earnings2Insights: Analyst Report Generation for Investment Guidance*, organized at EMNLP 2025, provides a benchmark for this emerging area by evaluating systems on their ability to convert earnings call transcripts into structured, investment-oriented analyst reports. Our work presents a systematic exploration of meta-prompting strategies, highlighting how carefully designed instructions can guide LLMs toward producing coherent, faithful, and decision-supportive reports. Through our participation, we aim to shed light on the potential and limitations of LLM-driven financial text generation.

2 Related Work

The intersection of financial analysis and large language models (LLMs) has become an active research area, with a growing emphasis on *multi-agent systems* for complex decision-making.

2.1 Multi-Agent Frameworks in Finance

Several works leverage LLM-based multi-agent systems for financial applications. Jajoo et al. (2025) introduce **MASCA**, a hierarchical framework for credit assessment that integrates contrastive learning and signaling game theory. Park (2024) propose a collaborative agent system for anomaly detection in stock markets, improving interpretability of alerts on the S&P 500 index. Beyond specific tasks, An et al. (2024) present **FinVerse**, an autonomous agent system with extensive API integration and code execution, while Yang et al. (2024) develop **FinRobot**, an open-source platform that formalizes a “Financial Chain-of-Thought” to democratize financial reasoning.

2.2 Surveys and Methodological Advances

Survey efforts further consolidate these directions. Ding et al. (2024) review LLM-powered trading agents, highlighting their architectures and evaluation challenges. Jadhav and Mirza (2025) synthesize 84 studies on LLMs in equity markets, categorizing applications such as forecasting, sentiment analysis, and portfolio management. Methodological advances outside finance also provide inspiration: Shen et al. (2025) show how textual feed-

[*] Corresponding Author: pulkitchatwal@gmail.com

back loops improve role-based multi-agent coordination in software engineering, while Li et al. (2024) present a general workflow for LLM-based MAS across domains.

2.3 Evaluation Paradigms

Traditional metrics for comparing generated analyses against ground truth have been criticized as insufficient for decision-making tasks (Goldsack et al., 2025; Chen et al., 2024). Recent work instead promotes *decision-oriented evaluation*, where generated texts are assessed by their influence on human judgment. Takayanagi et al. (2025) examine whether GPT-4 can sway expert decisions, while Huang et al. (2025) formalize decision-oriented text evaluation. Following this line, we adopt an evaluation setting where annotators make investment choices based on generated reports, emphasizing persuasiveness and actionability rather than surface-level similarity.

Positioning. While prior studies focus on trading, anomaly detection, or credit assessment, our work addresses the underexplored task of *investment guidance from earnings calls*, combining agentic prompting with reflective mechanisms and decision-oriented evaluation.

3 Problem Statement

Earnings call transcripts contain rich but unstructured information about a company's financial performance, management outlook, and market guidance. While human analysts can interpret these transcripts to produce actionable investment reports, this process is time-consuming, costly, and prone to subjective biases. The central problem addressed in this shared task is the development of automated systems that can transform raw transcripts into structured, coherent, and decision-oriented reports.

The key challenge lies in balancing multiple requirements: (*i*) ensuring factual accuracy and faithfulness to the source text, (*ii*) generating analyses that align with downstream financial decision-making (e.g., LONG/SHORT predictions across different time horizons), and (*iii*) producing outputs that are clear, logical, persuasive, and useful for human readers.

This problem is of practical importance, as inaccurate or uninformative reports may mislead investors and undermine trust in automated financial analysis. Therefore, the task provides not only an opportunity to benchmark natural language generation systems under realistic conditions, but also to advance methods for building reliable, interpretable, and actionable AI-driven financial assistants.

4 Methodology

In this section, we describe the resources and techniques employed in developing our system for the *Earnings2Insights* shared task. Specifically, we thoroughly outline the dataset characteristics, the chosen model architecture, the elaborate prompting strategy developed, and the comprehensive meta-prompting framework that meticulously guided the report generation process to ensure accuracy and coherence.

4.1 Dataset

The shared task organizers provided two primary subsets of earnings call transcripts to facilitate diverse and comprehensive system training and evaluation:

- **ECTSum subset**: 40 transcripts paired with reference summaries ("ref" files) from the ECTSum dataset (Huang et al., 2025). Use of these summaries was optional for participants.

- **Professional subset**: 24 transcripts matched to professional analyst reports. Only transcripts were accessible to participants; comparison to analyst reports was managed by the organizers downstream.

Consequently, all participating teams were mandated to generate detailed reports for each of the 64 earnings calls, ensuring full coverage of the dataset and enabling thorough performance comparison.

4.2 Model

For this task, we employed the **Meta LLaMA 3.2-1B Instruct** (Grattafiori et al., 2024) model, a recent instruction-tuned large language model released by Meta AI. Although relatively lightweight containing only 1 billion parameters, it is specifically designed to follow complex, multi-step instructions and reason deeply over structured and semi-structured texts, making it highly efficient for resource-constrained deployment scenarios while still maintaining state-of-the-art performance. Its instruction tuning, together with alignment through human feedback, enables the model to effectively handle domain-specific summarization and complex analytical tasks, notably without the need for additional fine-tuning or retraining. This carefully maintained balance between computational efficiency and advanced reasoning capability makes LLaMA 3.2-1B Instruct an ideal backbone choice for reliably generating highly structured and consistent financial reports in our system.

4.3 Prompting Strategy

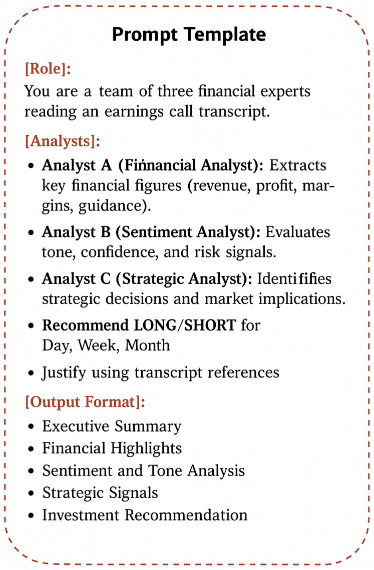
Prompt engineering played an absolutely central role in ensuring the generation of both accurate and thoroughly decision-oriented analyst reports. Initial experimental tests with vanilla prompting immediately highlighted noteworthy limitations in ensuring factual grounding and enforcing structured, logical reasoning. To successfully address these issues, we strategically adopted a robust multi-step prompting strategy that intricately integrates deep financial discourse understanding with rigorous structured report generation.

4.4 Meta-Prompting Framework

To significantly enhance reasoning consistency and overall reliability, we designed an innovative meta-prompting (Hou et al., 2022) framework inspired by principles of multi-agent collaboration and distributed cognition. The system simulates a collaborative team of three specialized financial experts—each focusing on quantitative analysis, nuanced sentiment evaluation, and strategic interpretation respectively. By explicitly defining individual expert roles and precisely specifying the required structured output format, our framework effectively guides the model toward producing reports that are coherent, factually faithful, and sharply focused on investment decision-making.

4.5 Prompt Template Illustration

Figure 1 and Figure 2 vividly illustrate the systematic design of our structured multi-agent prompt. The prompt template shown in Figure 1 defines explicit analyst roles, clearly specifies the required structured output, and thoughtfully incorporates diverse contextual information such as company introduction details, recent news reports, financial performance metrics, and stock price movement data. This meticulous design ensures that the generated investment reports are comprehensive in scope, factually consistent throughout, and explicitly decision-oriented. In parallel, the accompanying workflow diagram (Figure 2) graphically depicts the systematic process by which the system processes an earnings call transcript. Specifically, the three analysts extract complementary insights focusing respectively on financial data, sentiment signals, and strategic factors, which are then synthesized and consolidated into a unified investment recommendation. This modular setup not only enhances interpretability but also enforces strict domain-specific rigor and analytical precision.



Prompt Template

[Role]:
You are a team of three financial experts reading an earnings call transcript.

[Analysts]:

- **Analyst A (Fińnancial Analyst):** Extracts key financial figures (revenue, profit, margins, guidance).
- **Analyst B (Sentiment Analyst):** Evaluates tone, confidence, and risk signals.
- **Analyst C (Strategic Analyst):** Identififies strategic decisions and market implications.
- **Recommend LONG/SHORT** for Day, Week, Month
- Justify using transcript references

[Output Format]:

- Executive Summary
- Financial Highlights
- Sentiment and Tone Analysis
- Strategic Signals
- Investment Recommendation

Figure 1: Structured multi-agent prompt template with role definitions, task breakdown, and output format.

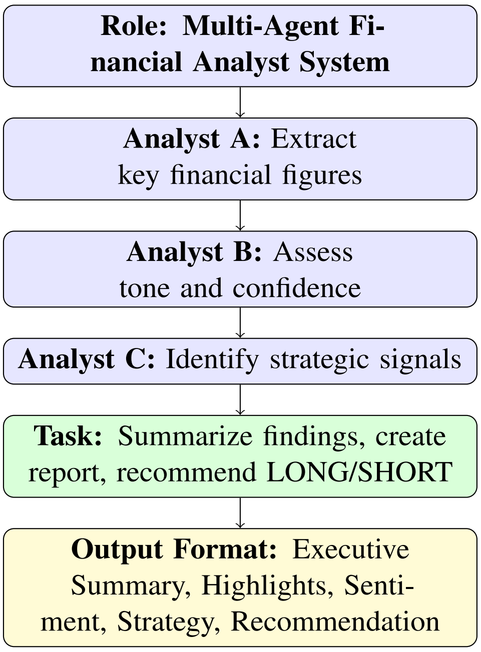

Role: Multi-Agent Financial Analyst System
Analyst A: Extract key financial figures
Analyst B: Assess tone and confidence
Analyst C: Identify strategic signals
Task: Summarize findings, create report, recommend LONG/SHORT
Output Format: Executive Summary, Highlights, Sentiment, Strategy, Recommendation

Figure 2: Workflow of the multi-agent prompt showing analyst roles, sequential tasks, and structured output.

5 Results

5.1 Evaluation Setup

The shared task employed both automatic and human evaluation. While participants could explore automatic metrics and LLM-based evaluations, the official ranking was determined solely by human judgments. Annotators were recruited via the Prolific platform (210 participants, with 34 excluded for failed attention checks). Each annotator reviewed 12 reports and made investment decisions (LONG/SHORT) for the next day, week, and month. Final system performance was scored by the average accuracy of these decisions across the three horizons. In addition, human raters assessed reports on five qualitative dimensions: clarity, logic, persuasiveness, readability, and usefulness, using a 7-point Likert scale.

5.2 Results

Our system, demonstrated strong empirical performance in the shared task. We ranked **second** out of twelve teams in terms of average financial decision accuracy, achieving an overall score of **0.579**. This reflects robust predictive alignment across daily (0.597), weekly (0.611), and monthly (0.529) investment horizons. On qualitative aspects, our reports received an average Likert score of **5.47** out of 7 from human judges, with particularly strong ratings for clarity (5.56) and readability (5.73).

Tables 1 and 2 present a breakdown of scores across all participating teams.

Team	Avg.	Day	Week	Month
DKE	0.581	0.596	0.577	0.570
Our Result	**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 1: Average accuracy of financial decisions made by participants after reading the reports generated by each team, across daily, weekly, and monthly horizons.

Team	Avg.	Cl.	Lo.	Per.	Read.	Use.
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
Our Result	**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 2: Average Likert scores (7-point scale) of generated reports across five qualitative dimensions: clarity, logic, persuasiveness, readability, and usefulness.

6 Conclusion

In this work, we presented our system for the *Earnings2Insights* shared task, focusing on generating insightful analyst reports from earnings call transcripts to aid investment decisions. Our approach, leveraging the **Meta LLaMA 3.2 1B Instruct** model with structured prompting and iterative refinement, achieved competitive results across quantitative and qualitative metrics.

Our team, ranked second out of twelve in financial decision accuracy. This ranking reflects the system's strong ability to produce reports that effectively guide human annotators toward correct investment choices (LONG or SHORT), demonstrating robust alignment with actionable financial outcomes, especially in capturing near-term market signals and predictive insights.

Qualitatively, our reports earned strong Likert ratings on a 7-point scale across clarity, logic, persuasiveness, readability, and usefulness. These ratings highlight the reports' balanced quality, indicating high accessibility, logical coherence, and practical value for readers, as driven by our meta-prompting framework.

These outcomes demonstrate the value of domain-specific reasoning in LLMs for faithful financial analysis. However, challenges like hallucinations and explainability persist. Future efforts will integrate external knowledge and enhance grounding for more trustworthy systems.

Acknowledgments

We thank the organizers of the Earnings2Insights shared task for their efforts in curating the dataset, designing the evaluation protocol, and fostering a collaborative environment for advancing financial text generation research.

References

Siyu An, Qin Li, Junru Lu, Di Yin, and Xing Sun. 2024. Finverse: An autonomous agent system for versatile financial analysis. *arXiv preprint arXiv:2406.06379*.

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.

Han Ding, Yinheng Li, Junhao Wang, and Hang Chen. 2024. Large language model agent in financial trading: A survey. *arXiv preprint arXiv:2408.06361*.

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, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yutai Hou, Hongyuan Dong, Xinghao Wang, Bohan Li, and Wanxiang Che. 2022. MetaPrompting: Learning to learn better prompts. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3251–3262, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

Aakanksha Jadhav and Vishal Mirza. 2025. Large language models in equity markets: Applications, techniques, and insights. *Techniques, and Insights (March 15, 2025)*.

Gautam Jajoo, Pranjal A Chitale, and Saksham Agarwal. 2025. Masca: Llm based-multi agents system for credit assessment. *arXiv preprint arXiv:2507.22758*.

Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinagearth*, 1(1):9.

Taejin Park. 2024. Enhancing anomaly detection in financial markets with an llm-based multi-agent framework. *arXiv preprint arXiv:2403.19735*.

Ming Shen, Raphael Shu, Anurag Pratik, James Gung, Yubin Ge, Monica Sunkara, and Yi Zhang. 2025. Optimizing llm-based multi-agent system with textual feedback: A case study on software development. *arXiv preprint arXiv:2505.16086*.

Takehiro Takayanagi, Hiroya Takamura, Kiyoshi Izumi, and Chung-Chi Chen. 2025. 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.

Hongyang Yang, Boyu Zhang, Neng Wang, Cheng Guo, Xiaoli Zhang, Likun Lin, Junlin Wang, Tianyu Zhou, Mao Guan, Runjia Zhang, and 1 others. 2024. Finrobot: An open-source ai agent platform for financial applications using large language models. *arXiv preprint arXiv:2405.14767*.