Earnings2Insights: Analyst Report Generation for Investment Guidance

Takehiro Takayanagi[1,2,3], **Tomas Goldsack**[4], **Kiyoshi Izumi**[1,2],
Chenghua Lin[5], **Hiroya Takamura**[3], **Chung-Chi Chen**[3]

[1]Simulacra Inc., [2]The University of Tokyo,
[3]National Institute of Advanced Industrial Science and Technology,
[4]The University of Sheffield, [5]University of Manchester

{takayanagi,izumi}@simulacra.co.jp,tgoldsack1@sheffield.ac.uk
chenghua.lin@manchester.ac.uk,takamura.hiroya@aist.go.jp, c.c.chen@acm.org

Abstract

We present **Earnings2Insights**, a shared task on generating actionable investment reports from earnings conference call (ECC) transcripts. Unlike traditional financial summarization or QA, the goal is decision support: systems must synthesize facts, highlight risks and opportunities, and support investors in making sound actions. The task required participants to produce reports based on ECC transcripts. In total, 45 teams registered, with 12 teams submitting reports and 9 submitting solution papers, spanning diverse agentic designs, retrieval-augmented methods, and data expansion strategies. Our evaluation consists of *human evaluation* and *automatic evaluation*. Results reveal a consistent divergence between systems that scored highly in automatic evaluations and those that most effectively supported human investment decisions, underscoring the limits of style- or reference-based comparisons in high-stakes financial report generation. We advocate human-centered, decision-oriented assessment as the primary lens, with automated signals serving as complementary diagnostics. We release task design, evaluation data, and scripts to catalyze research on decision-centric financial text generation.[1]

1 Introduction

With the advent of large language models (LLMs), researchers have increasingly explored their application in specialized professional domains. Beyond automatic text comprehension, LLMs now demonstrate promising abilities in analytical report generation, enabling new forms of decision support in high-stakes fields such as law, medicine, and finance (Goldsack et al., 2025). Financial decision-making is a particularly high-stakes domain, where inaccurate or misleading reports can directly impact markets and investor outcomes (Lai et al., 2023). Traditional NLP tasks in finance, such as information extraction (Chen et al., 2021a), question answering (Chen et al., 2021b; Liu et al., 2023), and summarization (Huang et al., 2024), have focused on factual accuracy. Recently, more and more focus has shifted to the **human side**, such as building financial advisor systems with LLMs (Takayanagi et al., 2025a,b). At the same time, producing actionable investment insights requires more than summarizing facts: systems must synthesize information, highlight risks and opportunities, and persuade investors to act (Huang et al., 2025).

The **Earnings2Insights shared task** is designed to evaluate the capability of LLMs to generate convincing investment reports from earnings call transcripts. Participants may approach the task in two ways: using only the raw transcript, or enriching the input with timestamp-aligned retrieval of relevant external information. A central challenge in financial report generation is evaluation. Prior studies have shown that comparing generated outputs with ground-truth answers via automatic metrics may be insufficient, and that current LLMs remain unreliable as evaluators (Chen et al., 2024; Goldsack et al., 2025). Inspired by decision-based evaluation frameworks (Takayanagi et al., 2025c; Huang et al., 2025), we instead assess systems by their ability to guide human investment decisions. Annotators are asked to make buy/hold/sell judgments based on the generated reports, and the correctness of these decisions serves as the primary evaluation metric.

This paper provides an overview of the Earnings2Insights shared task and dataset, summarizes the methods employed by participating teams, and evaluates their experiments. Through this, we aim to shed light on the current capabilities and limitations of LLMs in financial report generation, and to foster broader discussion on human-centered evaluation for decision-critical AI.

[1] https://github.com/TTsamurai/Earnings2Insights.git

2 Tasks and Dataset

The Earnings2Insights shared task evaluates the ability of large language models to generate actionable investment reports from earnings call transcripts. Unlike traditional summarization or QA tasks, the objective is not merely to condense information but to produce guidance that highlights risks, opportunities, and potential actions for investors. This setting mirrors real-world analyst workflows, where the value of a report lies in its ability to influence financial decisions rather than reproduce factual details alone.

We use earnings conference calls (ECCs) as our primary scenario. ECCs are quarterly events in which company executives present financial results and discuss their outlook with investors and analysts. ECCs play a central role in shaping market sentiment because they combine both quantitative disclosures (e.g., revenues, forecasts, margins) and qualitative signals (e.g., managerial tone, confidence, and forward-looking statements). Importantly, professional equity analysts routinely write analyst reports immediately after ECCs, making this setting particularly suitable for our task: it naturally links raw financial discourse to the generation of actionable investment insights.

In this shared task, we provide two complementary subsets of earnings call transcripts:

- **ECTSum Subset (40 transcripts)**
  This subset corresponds to the ECTSum dataset (Mukherjee et al., 2022). Each transcript is paired with a "ref" file representing its associated summary. Participants may choose whether or not to use these summaries as auxiliary supervision.

- **Professional Subset (24 transcripts)**
  This subset consists of transcripts that are aligned with professional analyst reports. Unlike the ECTSum subset, no reference summaries are provided to participants. Instead, the organizers will later compare system outputs against the analyst reports to assess alignment with professional standards.

In total, participants are required to generate reports for all 64 earnings calls across both subsets.

A total of 45 teams registered for the Earnings2Insights shared task, of which 12 teams submitted reports and 9 teams submitted solution papers.

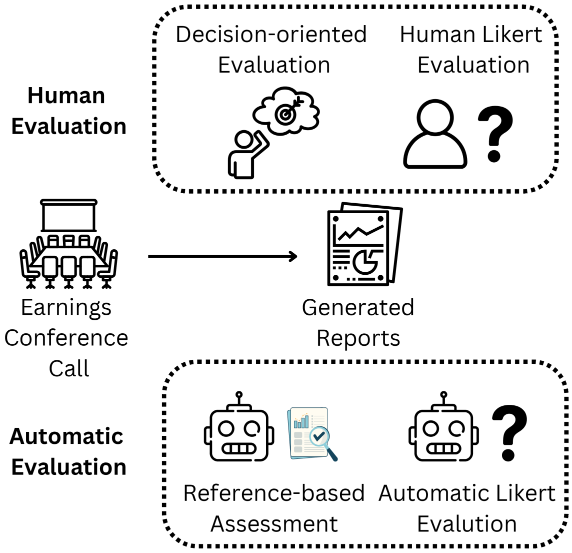

Figure 1: Evaluation framework consisting of human evaluation and automatic evaluation.

3 Evaluation

For evaluation, we conducted both human evaluation and automatic evaluation in order to capture complementary perspectives on system performance. Our evaluation framework is illustrated in Figure 1.

3.1 Human Evaluation

Human evaluation was designed to test whether the generated reports could effectively guide investment decisions. After reading each report, annotators were asked to make one of three decisions: *Buy* (expect the stock to go up), *Neutral* (uncertain), or *Sell* (expect the stock to go down). Ground-truth labels were derived from realized stock returns at three horizons: one business day (1bd), one week (5bd), and one month (20bd). These labels were coded as $+1$ for upward movements and -1 for downward movements. Neutral responses were excluded from the calculation, since they indicate uncertainty rather than a directional prediction. Accuracy was computed at each horizon as the proportion of correct predictions among all non-neutral responses, and an overall accuracy score was obtained by averaging across the three horizons.

In addition to directional accuracy, we also evaluated the perceived quality of the generated reports. Annotators rated each report on five criteria—clarity, logic, persuasiveness, readability, and usefulness—using a 7-point Likert scale. We report both the average score for each dimension and the overall mean across all five dimensions.

3.2 Human Evaluation Setup

For human evaluation, we used the Prolific platform.[2] We recruited 192 English native speakers residing in either the United Kingdom or the United States, each with a past task acceptance rate above 80%. Each crowdworker participated in one hour of evaluation, during which they made financial decisions based on a total of 12 generated reports. Consequently, every one of the 64 reports submitted by the 12 participating teams received independent judgments from three annotators. Participants were compensated at a rate of £8 per hour. In total, the study required 210 participants, amounting to a total cost of £1,680.

3.3 Automatic Evaluation

To complement the human evaluation, we also introduced automatic evaluation measures based on large language models. In particular, we adopted an "LLM-as-a-judge" framework (Gu et al., 2024), to provide pairwise and absolute quality judgments.[3] First, we measure the win rate against professional analyst reports. In this pairwise comparison, each system-generated report is compared directly with an analyst-written report, and the win rate reflects the proportion of cases in which the system report was judged superior, excluding ties. Second, we compute the average Likert score by aggregating the 1–7 ratings across the five qualitative dimensions described above. This provides a single summary indicator of report quality.

4 Methods

Overall, the participating teams adopted diverse agentic approaches, with many incorporating retrieval-augmented generation (RAG) and various data expansion strategies. This diversity illustrates the richness of methods explored for financial report generation.

SigJBS used a three-agent pipeline (extraction, reasoning, critique) to parse transcripts into key financial milestones, generate recommendations with risk analysis, and iteratively refine outputs for consistency and factuality (Sinha et al., 2025).

Jetsons combined writer agents with feedback agents in a ReAct-style loop (Yao et al., 2023), integrating structured financial data via Alpha Vantage to produce reports that balanced factual accuracy, risk coverage, and persuasiveness (Dakle et al., 2025).

Team	Average	Day	Week	Month
DKE	**0.581**	0.596	0.577	**0.570**
DataLovers	0.579	0.597	**0.611**	0.529
Jetsons	0.571	0.607	0.555	0.552
SigJBS	0.545	**0.609**	0.513	0.512
iiserb	0.537	0.576	0.558	0.477
PassionAI	0.537	0.588	0.557	0.466
Finturbo	0.524	0.504	0.568	0.500
Bgreens	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 across time horizons.

LangKG employed a cognitive reasoner framework, generating personalized reports tailored to investor profiles using a six-dimensional analysis and conviction scores for transparency (Prasanna and Su, 2025).

DataLovers orchestrated multiple analyst agents (finance, sentiment, strategy) whose outputs were merged into a structured report template. Their meta-prompting framework emphasized collaborative reasoning, implemented with a compact LLaMA model (Chatwal et al., 2025).[4]

iiserb modeled investment committee debates through a Structured Adversarial Synthesis framework, staging adversarial dialogues among bullish, bearish, and devil's advocate agents to refine logic and persuasiveness (Sadhu et al., 2025).

Bgreens mimicked the analyst–writer–editor workflow with multi-agent roles implemented via AutoGen (Wu et al., 2024). Iterative feedback improved consistency and readability, with experiments showing higher decision accuracy compared to single-agent baselines (Satapathy et al., 2025).

DKE built a retrieval-augmented debate system with five domain-specific analyst agents and a collaborative debate phase among trust, skeptic, and leader agents, synthesizing robust recommendations with confidence scores (Cai et al., 2025).

FinTurbo emphasized professional-style reports with structured data and visualization, combining charting, highlighting, writing, and editing

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

[3] We use gpt4.1 as our evaluator.

[4] https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct

Team	Average	Clarity	Logic	Persuasiveness	Readability	Usefulness
LangKG	**5.96**	**6.02**	**5.92**	5.90	**5.81**	**6.13**
Jetsons	5.90	6.00	5.89	5.81	**5.81**	6.01
DKE	5.74	5.71	5.89	**5.95**	5.17	5.98
SigJBS	5.67	5.76	5.68	5.59	5.61	5.72
SI4Fin	5.56	5.52	5.84	5.60	5.06	5.80
DataLovers	5.50	5.56	5.45	5.32	5.73	5.47
Bgreens	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 across five qualitative dimensions.

Team	ALS	WR
SI4Fin	**4.916**	0.956
LangKG	4.903	0.881
Jetsons	4.834	0.762
KrazyNLP	4.830	**0.962**
iiserb	4.807	0.930
DKE	4.803	0.783
Finturbo	4.625	0.169
SigJBS	4.597	0.526
Bgreens	4.575	0.615
bds-LAB	4.510	0.711
PassionAI	4.143	0.365
DataLovers	4.134	0.345

Table 3: Automatic evaluation results. ALS = Average Likert Score (1–7); WR = Win Rate vs. Analyst Report.

agents. They expanded the dataset by crawling additional transcripts to enable temporal RAG comparisons (Yang et al., 2025).

SI4Fin integrated external financial statements from Alpha Vantage with an AutoGen-based agentic framework (Wu et al., 2024), where analyst agents extracted trends (YoY, QoQ) and writers incorporated them into grounded reports (Tan et al., 2025).

5 Results

5.1 Human Evaluation Results

Table 1 reports the average accuracy of financial decisions made by annotators after reading the reports generated by each team. Accuracy is computed for one business day (Day), one week (Week), and one month (Month) horizons, with the overall average representing the mean of the three horizons.

Table 2 presents the average Likert scores for clarity, logic, persuasiveness, readability, and usefulness, as rated on a 7-point scale. We also report the overall mean score across the five criteria.

Overall, the results show noticeable variation across teams, with certain systems excelling in decision accuracy while others were rated more highly on subjective quality dimensions. This highlights the complementary nature of accuracy-based and human-centered evaluations in financial text generation.

5.2 Automatic Evaluation Results

In addition to human evaluation, we conducted automatic evaluation using an LLM-as-a-judge framework. Table 3 reports two measures: **ALS** (Average Likert Score), the average 1–7 rating across five dimensions (persuasiveness, logic, usefulness, readability, and clarity), and **WR** (Win Rate vs. Analyst Report), the proportion of pairwise comparisons in which a system-generated report was judged superior to a professional analyst report (ties excluded).

6 Discussion

The results reveal a key divergence between decision-oriented human evaluation and automatic evaluation based on win rates against professional analyst reports. Teams such as DKE and DataLovers scored highly in human evaluation—effectively supporting annotators' investment decisions—yet ranked lower in automatic evaluation. In particular, DataLovers' reports provided practical guidance but showed a notably low win rate. This suggests that automatic metrics fail to capture the true decision utility of generated texts. Prior studies indicate that amateur investors are often unpersuaded by professional analyst reports, whose language and logic can be inaccessible. Thus, benchmarking generated texts solely against professional reports is insufficient for assessing their usefulness in real decision-making (Takayanagi et al., 2025c).

Moreover, the divergence between human and automatic Likert-scale evaluations highlights risks in relying on LLMs as evaluators. While LLMs offer scalability and consistency, their judgments may not align with actual investor behavior. This reinforces the central motivation of the shared task: evaluation must remain grounded in human decision outcomes, with automatic methods serving as complements. Future work should therefore pursue hybrid evaluation schemes that integrate human judgment, domain-specific financial metrics, and scalable LLM-based assessments.

7 Conclusion

This paper presented the Earnings2Insights shared task, which evaluates the capability of large language models to generate actionable investment guidance from earnings call transcripts. Distinct from traditional summarization or QA, our setting targets human-centered decision support: systems must synthesize facts, surface risks and opportunities, and support investors toward sound actions. We released two complementary subsets (ECTSum and Professional), and attracted a diverse set of agentic methods from participating teams.

Our evaluation combined decision-oriented human assessment with an automatic "LLM-as-a-judge" protocol. Results revealed a consistent divergence: several systems that improved human decision accuracy did not necessarily score highly against professional analyst reports or in LLM-based judgments, and vice versa. These findings underscore a central lesson for high-stakes financial NLP: evaluation must remain grounded in human decision outcomes; automatic metrics are valuable but imperfect complements. In the future work, we envision hybrid evaluation protocols that integrate human decision accuracy, domain-specific financial measures, and calibrated, auditable LLM judgments.

We hope Earnings2Insights catalyzes sustained progress on decision-centric financial text generation. By releasing the task design, data splits, and evaluation scripts, and by documenting successful agentic and retrieval-augmented patterns, we aim to provide a practical foundation for research and deployment of human-centered advisory systems in finance.

References

Tianshi Cai, Guanxu Li, Nijia Han, Ce Huang, Zimu Wang, Changyu Zeng, Yuqi Wang, Jingshi Zhou, Haiyang Zhang, Qi Chen, Yushan Pan, Shuihua Wang, and Wei Wang. 2025. FinDebate: Multi-Agent Collaborative Intelligence for Financial Analysis. In *Proceedings of the FinNLP Workshop at EMNLP 2025*, Suzhou, China.

Pulkit Chatwal, Mann Bajpai, Priyanshu Deswal, Harish Pratap Singh, and Santosh Kumar Mishra. 2025. Meta Prompting for Analyst Report Generation: Turning Earnings Calls into Investment Guidance. In *Proceedings of the FinNLP Workshop at EMNLP 2025*, Suzhou, China.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021a. *From opinion mining to financial argument mining*. Springer Nature.

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.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021b. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Parag Dakle, Sai Krishna Rallabandi, Nikhi Kohli, Khyati Morparia, Ojas Raundale, and Preethi Raghavan. 2025. Jetsons at the FinNLP-2025 - Earnings2Insights: Persuasive Investment Report Generation Using Single And Multi-Agent Frameworks. In *Proceedings of the FinNLP Workshop at EMNLP 2025*, Suzhou, China.

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.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Jiantao Huang, Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2024. Numhg: A dataset for number-focused headline generation. In *LREC/COLING*.

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

Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 2023. Towards a science of human-ai decision making: An overview of design space in empirical human-subject studies. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 1369–1385.

Chuang Liu, Junzhuo Li, and Deyi Xiong. 2023. Tab-CQA: A tabular conversational question answering dataset on financial reports. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 196–207, Toronto, Canada. 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. ECT-Sum: 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*, pages 10893–10906, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shivika Prasanna and Hui Su. 2025. LangKG at the FinNLP 2025 - Earnings2Insights: Task-Adaptive LLMs To Generate Human-Persuasive Investment Reports. In *Proceedings of the FinNLP Workshop at EMNLP 2025*, Suzhou, China.

Saisab Sadhu, Biswajit Patra, and Tannay Basu. 2025. Structured Adversarial Synthesis: A Multi-Agent Framework for Generating Persuasive Financial Analysis from Earning Call Transcripts. In *Proceedings of the FinNLP Workshop at EMNLP 2025*, Suzhou, China.

Ranjan Satapathy, Raphael Liew, Joyjit Chattorj, Erik Cambria, and Rick Goh. 2025. From Earnings Calls to Investment Reports: Evaluating Role-based Multi-Agent LLM Systems. In *Proceedings of the FinNLP Workshop at EMNLP 2025*, Suzhou, China.

Gaurangi Sinha, Rajarajeswari Palacharla, and Manoj Balaji Jagadeeshan. 2025. Agentic LLMs for Analyst-Style Financial Insights: An LLM Pipeline for Persuasive Financial Analysis. In *Proceedings of the FinNLP Workshop at EMNLP 2025*, Suzhou, China.

Takehiro Takayanagi, Kiyoshi Izumi, Javier Sanz-Cruzado, Richard McCreadie, and Iadh Ounis. 2025a. Are generative ai agents effective personalized financial advisors? In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 286–295, New York, NY, USA. Association for Computing Machinery.

Takehiro Takayanagi, Masahiro Suzuki, Kiyoshi Izumi, Javier Sanz-Cruzado, Richard McCreadie, and Iadh Ounis. 2025b. Finpersona: An llm-driven conversational agent for personalized financial advising. In *European Conference on Information Retrieval*, pages 13–18. Springer.

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

Mingrui Tan, Yang Liu, Kun Gao, Fei Gao, and Yuting Song. 2025. SI4Fin at Earnings2Insights: LLM-Based Analyst Report Generation for Earnings Calls. In *Proceedings of the FinNLP Workshop at EMNLP 2025*, Suzhou, China.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, and 1 others. 2024. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*.

Weijie Yang, Junbo Peng, and Weijie Yang. 2025. Beyond Summaries: Multi-Agent Generation of Investment Reports with Text, Tables, and Charts. In *Proceedings of the FinNLP Workshop at EMNLP 2025*, Suzhou, China.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.