LangKG at the FinNLP 2025 - Earnings2Insights: Task-Adaptive LLMs To Generate Human-Persuasive Investment Reports

Shivika Prasanna
shivika.prasanna@fmr.com

Hui Su
hui.su@fmr.com

Abstract

In this paper, we address the challenge posed by the FinNLP 2025 shared task on Earnings2Insights: Analyst Report Generation for Investment Guidance, with our two-stage framework system. Success of these generated reports is measured on the correctness and persuasion of human investors for investment decisions across different time frames and has been evaluated using automated metrics and human evaluation. Our system comprises of two stages that incorporates a sophisticated analysis of investment-centric sentiments and personalities from the call transcripts and leveraging this information with a comprehensive cognitive reasoning framework to generate carefully curated, accurate and persuasive reports using LLMs for human-decision making. Our approach ranked 1st out of 12 teams on the human-evaluation average Likert Score and 2nd on the automated-evaluation average Likert Score, demonstrating competitive performance.

1 Introduction

Earnings Call Transcripts are rich in technical financial information that can be time-consuming for investors to parse quickly. Creating quality and correct investment analysis reports that can be convincing for decision-making requires human experts that can prevent scaling. Automatic generation of investment research reports from earnings call transcripts presents a fundamental paradigm shift in natural language processing (NLP) evaluation. Traditional financial NLP research has primarily focused on information extraction, summarization and sentiment analysis of earnings calls (Huang et al., 2025). Recent advances in large language models (LLMs) have proven to be increasingly promising in financial analysis tasks, especially with automatic analytical report generation (Goldsack et al., 2024). However, these approaches often optimize for content accuracy or similarity to reference summaries rather than persuasive effectiveness required for real-world investment decision-making.

The Earnings2Insights shared task (Takayanagi et al., 2025a) introduces a novel evaluation methodology where annotators are asked to make investment decisions based solely on generated reports, with correctness measured by actual investment outcomes across different time frames of 1 day, 1 week and 1 month. This evaluation paradigm reflects the recognition that traditional metrics may not be meaningful enough for financial analysis tasks and that the current LLMs are not yet completely suitable to serve as judges for investment guidance quality or correctness.

Previous studies on evaluation have demonstrated limitations of automatic evaluation in financial text generation. (Chen et al., 2024) highlighted challenges in numerical-aware language understanding and generation while (Goldsack et al., 2024) specifically addressed the gap between factual analysis and insightful report generation for earnings calls. The evaluation method adopted by the Earnings2Insights shared task has been proposed by (Takayanagi et al., 2025b) which demonstrated the potential for AI-generated content to influence expert decision-making. (Huang et al., 2025) introduced decision-oriented text evaluation emphasizing success through decision-making effectiveness over content similarity. (Mukherjee et al., 2022) contributed the ECTSum dataset for earnings call summarization, establishing important benchmarks for financial transcript processing.

We further breakdown the objectives of the task into several unique challenges and address them in our two-stage framework:

- Generated reports must be both analytically and psychologically sound to be persuasive;
- These reports must include necessary information to provide correct and actionable guidance across multiple investment time frames;

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

- Success of the generated reports depends on understanding various investor personalities and decision-making patterns to curate them to be suitable for a diverse range of profiles;
- Absence of ground truth poses a fundamental challenge for us to optimize LLMs to generate reports with human-like decision recommendations.

We recognize the importance of accurately communicating correct insights that would resonate with different investor profiles. Unlike approaches that cater to a generic investor population uniformly, our system considers how different investors such as growth, risk-aware and other types would interpret the same earnings information differently. Because in the real-world scenario, there can be several different personalities, we do not limit the large language model to the list of investor profiles, but instead allow it to identify the similarities and differences across these profiles when generating recommendations and evidences for decision-making. We introduce several key innovations in our proposed two-stage framework:

- Enhanced investment-centric sentiment classification system with 8 granular, distinct categories that capture management confidence levels and question dodge behavior from the transcripts;
- Systematic integration of investor personality considerations that guides report generation to address concerns and priorities of different investor types;
- Comprehensive and sophisticated 6-dimensional analysis framework covering financial performance, business fundamentals, risk assessment, forward outlook, Environmental, Social and Governance (ESG) considerations and personal factors;
- Explicit conviction scoring with reasoning and position sizing recommendations tailored to different investment time frames.

The remainder of this paper is organized as: Section 2 dives deeper into our two-stage framework; Section 3 describes our experiments and Section 4 presents the results and analysis of our approach as well as the evaluation results from the shared task. We conclude our work in Section 5.

2 Methodology

Our two-stage framework addresses the fundamental challenge of generating accurate reports that drive profitable human investment decisions through systematic analysis and psychological considerations. It combines two fundamental and distinct steps to achieve optimized reports that are highly human-persuasive.

To derive explicit and implicit information from the call transcripts, we first employed an enhanced investment-centric sentiment analysis and information extraction process. For each question-answer (Q/A) pair within the call transcript, the model, which acts as a '*Data Extractor*', first identifies the most suitable sentiment out of 8 possible defined categories, as described in Section 2.1. It then evaluates the confidence within the sentiment and identifies all key phrases that contributed to the sentiment classification for the Q/A pair. The model also identifies the tone of the speaker (avoidance or dodging) and provides a score with an interpretation. Finally, the model provides the interpretation on the investment signal for each exchange, which is crucial for the next stage in the framework.

The second stage focuses on personality- and analysis-driven investment report generation. The summary created in the previous stage and several different types of personalities along with their descriptions (Section 2.3) are fed as input to the model, which in this stage acts as a '*Cognitive Reasoner*', to generate reports. We employ a multi-perspective analysis that considers how different investor personality types would interpret the same earnings information, ensuring the reports address diverse investment philosophies and provide an unbiased, evidence-based recommendation. This analysis is guided by application of our 6-dimensional (6D) framework covering various aspects described in Section 2.2. The framework generates conviction scoring with percentages for clear interpretation that quantify uncertainty. Reports provide recommendations for 1-day, 1-week and 1-month, recognizing that short-term price movements substantially differ from long-term performance drivers. Throughout this process, the framework ensures error avoidance by identifying common analytical errors such as optimism bias, making recommendations without considering investor suitability, and strictly adhering to the framework, ensuring the generated reports maintain analytical rigor, correctness (no hallucinations) and practical applicability with persuasion for human decision-making.

2.1 Investment-Centric Sentiment Classification System

In the first stage of our system, we developed an 8-category sentiment classification specifically designed for earnings call investment analysis as follows:

- Bullish: Strong positive, growth accelerating, beating expectations
- Optimistic: Moderately positive, things improving, meeting expectations
- Cautious: Uncertain but not negative
- Neutral: Balanced, no strong directional signals
- Concerned: Worried tone, challenges mentioned, defensive responses
- Bearish: Negative outlook, problems acknowledged, guidance cuts
- Evasive: Avoiding questions, deflecting, noncommittal answers
- Confident: Stronger than Optimistic and Bullish, shows conviction in guidance

The sentiment classification for each Q/A pair provides a deterministic confidence level on a scale of 1 (least confidence) to 5 (highly confident), key phrases that triggered the classification and the score, a question dodge score on a scale of 0 (direct response) to 3 (completely evasive) and an investment signal interpretation highlighting what this would mean for stock performance.

2.2 Structured Analysis Framework

We employ a comprehensive framework to further analyze the call transcripts to generate the recommendation reports. Traditional earnings call analysis can often miss the holistic perspective required to make an investment decision. Our framework ensures systematic coverage of these 6 factors that can influence the investment outcomes.

- Financial Performance: Captures the quantitative performance.
  - Revenue trends, growth rates, profitable margins
  - Beat/miss versus guidance, expectations
  - Key performance indicators
- Business Fundamentals: Evaluates competition, management track record, management quality.
  - Market position, competitive advantages
  - Management quality, strategic vision
  - Growth catalyst, expansion opportunities
- Risk Assessment: Identifies regulatory threats, vulnerabilities, or any factors that could derail investment thesis.
  - Industry and company-specific risks
  - Operational challenges
  - Market volatility factors
- Forward Outlook: Evaluates guidance credibility based on management's historical accuracy and identifies any types of catalysts.
  - Management guidance and expectations
  - Industry trends affecting future performance
  - Short and medium-term price drivers
- Personal Factors: Growth investors and value investors analyzing identical earnings calls can reach different conclusions.
  - Risk tolerance
  - Financial situation
  - Investment goals
  - Length of time to hold investment before needing funds
- Conviction Scoring: To allow investors to make evidence-based decisions, instead of only recommending Long or Short, this framework also quantifies the uncertainty and translates it into actionable interpretation.
  - 90-100%: Multiple bullish signals, minimal risks, clear catalysts
  - 70-89%: Strong thesis, but some uncertainty/risks present
  - 50-69%: Mixed signals, requires smaller position sizing
  - 30-49%: Weak thesis, high uncertainty, avoid or minimal exposure
  - 0-30%: Strong negative signals, consider short position

With this structure framework, we addressed common analytical failures such as overlooking competitive threats or providing guidance without conviction levels for position sizing. Each dimension addresses a specific cognitive bias that affect both human analysts and automated systems.

2.3 Investor Personality Integration

Every investor has certain preferences when making an investment decision. Our system takes this into account through prompt engineering to consider multiple perspectives and identify common and different personality elements that can affect an investment decision. Specifically, we prompt the 'Cognitive Reasoner' model to analyze (1) how different personality types would interpret the same information; (2) common traits that tie suitable investors together; (3) identification of investor types beyond angel, venture, personal, institutional and crowdfunding, that would be best suited for the investment; (4) balanced reasoning that synthesizes multiple viewpoints. Through this approach, we achieve a more balanced recommendation, rather than an overly or under optimistic one.

2.4 Error Prevention

We address over- or under-optimism by implementing a systematic error identification and avoidance. The instructions given to the model include avoiding overoptimism bias from confident management tone, distinguishing between genuine guidance and reality catch-up, questioning margin expansion claims without clear drivers and flagging evasive responses to specific questions.

3 Experiments

3.1 Dataset & Task Setup

The dataset provided in the Earnings2Insights shared task consists of 64 earnings call transcripts. Forty transcripts are ECTSum transcripts with reference summaries, and 24 are Professional subset transcripts without reference summaries. Each transcript consists of prepared remarks and Q/A pairs. The transcript follows the typical structure of a call transcript, following a conversational format between the speaker and an audience. The primary objective in this shared task is to generate investment reports for all 64 transcripts that can convince human evaluators to make profitable and correct trading decisions. Each report has been evaluated using automated metrics such as Likert Score (average score between 1-7 Likert ratings of Persuasiveness, Logic, Usefulness, Readability, and Clarity) and Win Rate vs Analyst Report (average score showing how often the report outperformed a professional analyst report in a pairwise comparison), and human evaluation using Likert Score and Average Accuracy (across three different time frames of 1-day, 1-week and 1-month) that measures the accuracy of the investment decisions.

For our experiments, we utilize GPT-4o (Hurst et al., 2024), due to its high speed, low latency and high accuracy, for this task. We set the temperature to 0.5 for both 'Data Extractor' and 'Cognitive Reasoner', with number of maximum tokens as 4096 and top_p parameter set to 0.95.

3.2 Implementation Details

Our framework implements a novel Task-Adaptive Model Utilization that utilizes the cognitive processing of LLMs, for the two stages. We use 'task-adaptive' to denote prompt-based adaptation of the model to different tasks or roles. Through this, we recognize the importance of matching model operation to specific cognitive demands rather than applying uniform processing.

The model first acts as a 'Data Extractor' to extract key implicit and explicit information from the transcripts by carefully analyzing the prepared remarks and each Q/A pair. We implement rigid prompt engineering that forces systematic analysis of each Q/A exchange through predefined analytical steps such as speaker role identification, tone analysis, sentiment classification, evidence extraction and output format for stage two of the framework. The high constraint setup prevents the model from hallucinations and inconsistency.

Then, the model transitions its role to a 'Cognitive Reasoner' to perform flexible reasoning, leveraging its sophisticated inference capabilities for multi-perspective synthesis. We reduce structural constraints while introducing cognitive frameworks that guide comprehensive analysis. The model synthesizes different viewpoints during reasoning. This approach generates more nuanced reports that addresses diverse investor concerns, providing curated recommendations for different investor profiles. Prompts are available in Appendix A.

The dual-mode architecture we employ balances precision with creativity. Extraction tasks require high accuracy and persuasive report generation demands flexible reasoning with psychological insight. By optimizing each stage for specific tasks, our framework achieves both analytical precision and human persuasiveness, shown in Section 4.

Table 1: LangKG Performance Summary

Metrics	Measure	Score
Human Evaluation	Day	0.589
	Week	0.542
	Month	0.424
	Average	0.518
Human Likert Score (1-7)	Clarity	6.02
	Logic	5.92
	Persuasiveness	5.90
	Readability	5.81
	Usefulness	6.13
	Average	5.96
Automated	Average Likert Score (1-7)	4.903
	Win Rate vs Analyst Report	0.881

4 Results

We report the automated and human evaluation results in Table 1. Our results demonstrate the success of the approach using human and automated Likert Score. While there is still scope of improvement on accuracy of the model-generated reports, on the leader board, we rank 1^{st} out of 12 teams in the human average Likert Score, 4^{th} for the 1-day time frame in the human average accuracy, 2^{nd} in automated average Likert Score and 4^{th} in Win Rate vs Analyst report metrics.

5 Conclusion

We present a Task-Adaptive LLM framework that addresses the fundamental challenge of generating investment reports optimized for human decision-making, as posed by the FinNLP Earnings2Insights: Analyst Report Generation for Investment Guidance shared task. Our proposed two-stage approach that combines enhanced investment-centric sentiment analysis with investor personalities and investment analysis framework demonstrates that successful financial NLP requires understanding both linguistic signals and psychological factors that drive investment decisions, along with financial (investment) knowledge. Our key contribution lies in the dual-mode cognitive architecture where, in the first stage the model acts as a 'Data Extractor' using 8 sophisticated sentiments and in the second stage, as a 'Cognitive Reasoner' using investor personality modeling and our 6D analysis framework. Our experiment results validate our methodology, with our approach ranking 1^{st} out of 12 teams in the human average Likert Score, 2^{nd} in the automated average Likert Score. On financial accuracy, our methodology shows promise by ranking 4^{th} in the 1-day time frame in human evaluation.

Ethics Statement

This research has been conducted for academic purposes. Investments may be risky and investors are advised to use their discretion, instead of solely relying on the output of the models.

References

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.

Tomas Goldsack, Yang Wang, Chenghua Lin, and Chung-Chi Chen. 2024. From facts to insights: A study on the generation and evaluation of analytical reports for deciphering earnings calls. *arXiv preprint arXiv:2410.01039*.

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

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, and 1 others. 2022. Ectsum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. *arXiv 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.

A Appendix A

We present the 'Data Extractor' and 'Cognitive Reasoner' prompts here.

Data Extractor Prompt

```
You are an expert at analyzing sentiments from text.
## [INPUT]

Earnings Call Transcript: $earnings_transcript

## [TASK]

1. Carefully read through the entire transcript which details the speakers for an investment company
     followed by a Question/Answer session where the speakers are posed questions surrounding
     investing in their company or their product.
2. Identify the key points in the prepared remarks and the transcript and create a summary.
    - For all numbers, dates, amounts, etc, only include what is given in the input.
    - Do not round off, do not change any values.
3. For each Q/A pair:
    Step 1: Identify who is asking the question, do not change names
    Step 2: Identify who is being asked the question, do not change names
    Step 3: Carefully analyze the tone of the answer
    Step 4: Generate the most suitable sentiment based on your analysis
3. Sentiment must be classified into:
    - **BULLISH**: Strong positive, growth accelerating, beating expectations
    - **OPTIMISTIC**: Moderately positive, things improving, meeting expectations
    - **CAUTIOUS**: Uncertain but not negative, "Wait and See" tone
    - **NEUTRAL**: Balanced, no strong directional signals
    - **CONCERNED**: Worried tone, challenges mentioned, defensive responses
    - **BEARISH**: Negative outlook, problems acknowledged, guidance cuts
    - **EVASIVE**: Avoiding questions, deflecting, non-committal answers
    - **CONFIDENT**: Stronger than OPTIMISTIC and BULLISH, shows conviction in guidance.

## [OUTPUT]

Strictly generate a JSON object in the following format. Do not add any markdown or strings before
     or after the JSON object.

{
    "prepared_remarks_summary": <Summarize the transcript>,
    "sentiment": {
        [
            "From": <Who is asking?>,
            "To": <Who is being asked?>,
            "Question": <List question>,
            "Answer": <List answer>,
            "Sentiment": <Specify sentiment>,
            "Confidence": <1-5 scale within that sentiment>
            "Key Phrases": <Specific words or phrases that indicated this sentiment>,
            "Question Dodge Score": <0-3 scale, 0=direct response, 3=completely evasive>,
            "Signal": <What this means for stock performance, investment implication>
        ],
        [...]
    }
}
```

Cognitive Reasoner Prompt

```
You are a professional financial analyst with expertise in investments.

## [TASK]
Generate an investment report from the financial earnings call transcript summary that will convince
     human investors to make profitable and correct buy or sell decisions for the next day, week,
     and month
Here are some things you can consider when analyzing the summary:
    1. Would you recommend this to an investor?
    3. What type of investors would this be best suited for and why?
    4. What is common in personality of the investors you select?

Your report should provide CLEAR investment guidance that will help investors make PROFITABLE,
     CORRECT and INFORMED decisions.

## [INPUT]

Financial Earnings Call Transcript Summary: Summary of the original call transcript along with each
     Question-Answer pair, an assigned sentiment to the pair, and a signal for what it means for
     stock performance.

$financial_earnings_summary

Examples of Investor Types: Examples of personality types and what investments they usually prefer.

$investor_df

## [ANALYSIS FRAMEWORK]
This is to help you create a framework structure for your report.
1. FINANCIAL PERFORMANCE
    - Revenue trends, growth rates, profitability margins
    - Beat or miss vs guidance and expectations
    - Key performance indicators and metrics
    - Leading, lagging or matching

2. BUSINESS FUNDAMENTALS
    - Market position, competitive advantages
    - Management quality and strategic vision
    - Growth catalyst and expansion opportunities

3. RISK ASSESSMENT
    - Industry and company-specific risks
    - Operational challenges
    - Market volatility factors

4. FORWARD OUTLOOK
    - Management guidance and expectations
    - Industry trends affecting future performance
    - Short and medium-term price drivers

5. PERSONAL FACTORS
    - Risk tolerance
    - Financial situation
    - Investment goals
    - Length of time to hold investment before needing funds

6. CONVICTION SCORING
    - 90-100%: Multiple bullish signals, minimal risks, clear catalysts
    - 70-89%: Strong thesis, but some uncertainty/risks present
    - 50-69%: Mixed signals, requires smaller position sizing
    - 30-49%: Weak thesis, high uncertainty, avoid or minimal exposure
    - 0-30%: Strong negative signals, consider short position
```

Cognitive Reasoner Prompt

```
## [ANALYZING INSTRUCTIONS]
1. Read through the given financial earnings transcript summary carefully.
2. Read through each Question-Answer pair and the sentiment assigned to it.
3. Provide a recommendation based on your analysis for the next day, week, and month. It can be LONG
     position or SHORT position.
    - LONG: buy, assumption is that the price of a security will increase over time
    - SHORT: sell, assumption is that the price of a security will decrease over time.
4. Identify what types of investors are best suited to invest in the product.
    - The given input for investors list is only to help you.
    - You may list some other investor types.
    - Be careful in identifying if the product should be recommended.
5. Think through the different personalities that you are selecting as suitors and identify what
     ties them together.
    - Use this for your reasoning.
    - Specifically, ask yourself
        - "What do these investors have in common and why should this investment be the best for
            them?"
        - "If I were an advisor, what kind of investor personality would I recommend this product
            to?"
6. AVOID THESE COMMON ERRORS:
    - Don't be overly optimistic just because the tone of the management sounds confident.
    - Watch for guidance raises that are actually just catching up to reality.
    - Be skeptical or margin expansion claims without clear drivers.
    - Flag when management avoids giving specific numbers.
7. Only use information from the given input call transcript. Do NOT use any external information or
     data.

## [REPORT INSTRUCTIONS]
1. Your report MUST be persuasive enough that when human evaluators read it, they can make the
     CORRECT investment decisions.
2. Use professional, authoritative tone for investors.
3. Support all your conclusions with evidence from the given input call transcript.
4. Balance all bullish and bearish factors.
5. Focus on driving stock performance over the next day, week, and month.
6. Use the ANALYSIS FRAMEWORK while writing your report.
7. Strictly adhere to the report structure given to you.
8. Ensure the report is concise, short and easy to read.

## [REPORT STRUCTURE]
1. Executive summary: 2-3 clear sentences with a clear "LONG" or "SHORT" recommendation for each
     timeline in natural language.
2. Driving stock performance: provide recommendation of "LONG" or "SHORT" for each of the next day,
     week, and month.
    - The recommendation can be different for each timeline. Use only "LONG" or "SHORT" and no
        synonyms of these words.
    - In ONE line, explain what would change your recommendation.
    - Be careful during your analysis and recommendation for each timeline.
    - You MUST provide the recommendation for all 3 timelines.
    - Explain why you are recommending the position.
    - Do not use "We" or "I" while recommending.
    - Do NOT use words like "downgrade" during your recommendation when describing why you would
        change from "LONG" to "SHORT".
3. Additional Information:
    3a. Financial highlights: key numbers and performance vs expectations with evidences from the
        call transcript.
    3b. Strategic updates: business initiatives, market developments from the call transcript.
    3c. Risk factors: highlight any concerns, potential headwinds.
    3d. Investment thesis: why buy or sell, what timings to consider.
    3e. Environmental, Social, Governance (ESG) and Other Qualitative Factors: identify management
        comments on sustainability, employee issues, ethics and other such factors and assess the
        credibility. Also evaluate any ESG risk possible that could impact the investment.
    3f. Insights: provide detailed, concise insights into the CONVICTION SCORING, strategy, risks,
        and market positioning mentioned in the call transcript. Include anything from the ANALYSIS
        FRAMEWORK here.
    3g. Investor Suitability: list the best suited investor profiles.
        - You may use some examples from the list.
        - Include the difference in the strategy of investment for different types of investors.

Your report will be evaluated on whether it convinces investors to make correct investment decisions.
```