

# From Facts to Insights: A Study on the Generation and Evaluation of Analytical Reports for Deciphering Earnings Calls

Tomas Goldsack,<sup>1</sup> Yang Wang,<sup>1</sup> Chenghua Lin,<sup>1,2</sup> Chung-Chi Chen<sup>3</sup>

<sup>1</sup>Department of Computer Science, University of Sheffield, UK

<sup>2</sup>Department of Computer Science, University of Manchester, UK

<sup>3</sup>Artificial Intelligence Research Center, AIST, Japan

{tgold sack1, Y.Wang4}@sheffield.ac.uk chenghua.lin@manchester.ac.uk

c.c.chen@acm.org

## Abstract

This paper explores the use of Large Language Models (LLMs) in the generation and evaluation of analytical reports derived from Earnings Calls (ECs). Addressing a current gap in research, we explore the generation of analytical reports with LLMs in a multi-agent framework, designing specialized agents that introduce diverse viewpoints and desirable topics of analysis into the report generation process. Through multiple analyses, we examine the alignment between generated and human-written reports and the impact of both individual and collective agents. Our findings suggest that the introduction of additional agents results in more insightful reports, although reports generated by human experts remain preferred in the majority of cases. Finally, we address the challenging issue of report evaluation, we examine the limitations and strengths of LLMs in assessing the quality of generated reports in different settings, revealing a significant correlation with human experts across multiple dimensions.

## 1 Introduction

Earnings Calls (ECs), critical quarterly meetings conducted by publicly traded companies to discuss financial performance with professional analysts, have been extensively studied for various prediction tasks. These tasks include volatility prediction (Sawhney et al., 2021; Niu et al., 2023), analyst decision prediction (Keith and Stent, 2019), financial risk prediction (Qin and Yang, 2019), and earnings surprise prediction (Koval et al., 2023), highlighting ECs’ significance in investment decision-making. Because ECs’ typical duration of about one hour, another prominent research area in this domain is summarizing lengthy EC transcripts (Mukherjee et al., 2022). Post-EC, two types of summaries emerge: *Journalistic Reports*, in which journalists concisely summarize the key financial takeaways from the meeting, and *Analytical Reports*, in which professional analysts offer a con-

siderably more extensive and multifaceted analysis of meeting events, financial performance, and implications on investment strategies. Whilst the automatic generation of journalistic reports has been addressed in previous studies (Mukherjee et al., 2022), no work to our knowledge has explored the task of generating analytical reports, despite numerous potential benefits. For example, the automatic generation of analytical reports could reduce the burden placed on analysts, enable the immediate distribution of key information to a broad range of stakeholders, and introduce novel insights through scalable interpretation of complex data. However, given the inherent complexity of generating analytical reports, success in this challenging task requires a methodology that can enable an in-depth analysis across multiple varied and important technical aspects, such as expectations on future operations and managers’ attitudes during ECs.

One promising candidate for such an approach comes in the form of multi-agent frameworks: an exciting avenue of recent research that explores how multiple role-playing LLM “Agents” can be deployed to cooperatively solve a task. When deployed for generation tasks such as Software development (Qian et al., 2023) and Trivia-based Creative Writing (Wang et al., 2024), the introduction of role-based division has enabled complicated requirements to be broken down into simple subtasks and processes, reducing the cognitive and contextual burden placed the underlying models. Furthermore, the utilization of role-playing LLM agents has provided expanded opportunities for domain specialization, the leveraging of external data/tools, and the incorporation of diverse viewpoints, all of which could add significant value to the generation of analytical reports. Notably, such an approach also bears a closer resemblance to a human writing process which, from a cognitive science perspective, is a complex, cyclic, and multi-step procedure, often requiring strategic discourse planning and

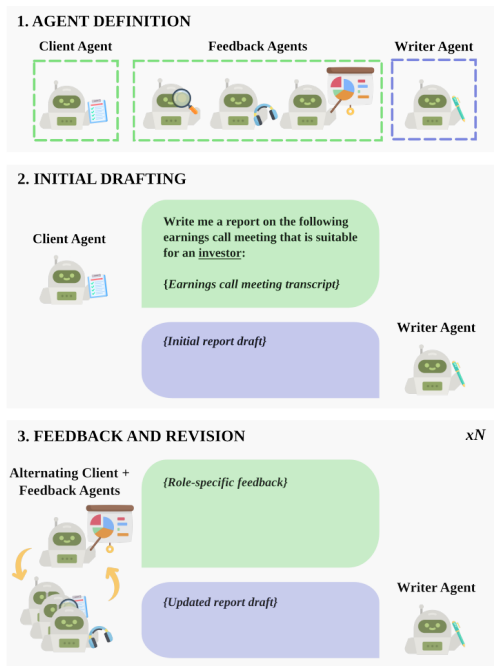


Figure 1: An overview of our multi-agent framework.

multiple iterations to effectively achieve a communicative goal (Flower and Hayes, 1981).

In this work, we explore the utility of a multi-agent framework for generating analytical reports using LLMs. We establish the details of our novel multi-agent framework and specially designed agents (§2), before performing a thorough characterization of the generated reports under different settings (§3), highlighting key differences to human-authored reports and the additional insights offered by feedback agents. Finally, we assess the capabilities of LLMs in evaluating the quality of generated reports (§4), establishing both promising directions and limitations, and laying the groundwork for future research on this task.

Overall, the following three research questions (RQ) are addressed:

**(RQ1):** How do generated analytical reports differ from human-authored analytical reports?

**(RQ2):** Can a multi-agent approach be used to generate more insightful analytical reports?

**(RQ3):** How effective are LLM-based evaluation methods in assessing the quality of analytical reports?

## 2 Multi-Agent Report Generation

We explore a novel framework for generating analytical reports through collaborative multi-agent conversation, leveraging the capacity of LLMs to

refine their output based on feedback. This framework, developed using Microsoft’s AutoGen (Wu et al., 2023), assigns each LLM-powered agent a distinct role through an initialization prompt, dictating their contribution to the conversation. For all experiments and agents, we employ the gpt-4-1106-preview model via the ChatGPT API as the underlying LLM. Figure 1 illustrates the framework, whereby a single agent with the role of **Writer** (✍️) is tasked with drafting and revising the report, guided by feedback from other agents. The process encompasses three stages:

**Stage 1. Agent Definition** Before generation commences, it is imperative to define the agents involved: 1) a **Client** (📝) agent, providing the initial task brief (as shown in Figure 1) and subsequent feedback representing the audience’s perspective; 2) “Feedback” agents, offering insights based on their specific roles. In this study, we explore the integration of three feedback agents alongside the Writer and Client agents. Details of the prompts used to initialize and generate responses for each agent are provided in Appendix B.

**Analyst** (📊) agent tasked with extracting and analyzing historical financial data, the Analyst agent leverages the AlphaVantage API to gather earnings performance data for the previous quarter. This information, presented alongside the preceding conversation context to the LLM, allows the agent to draw additional insights about the company’s current financial performance through comparison to previous performance, enabling the formulation of pertinent feedback.<sup>1</sup>

**Psychologist** (🧠) agent analyzes external data, specifically phonetic statistics from earnings call (EC) audio recordings, to offer additional insights on the level of confidence vocally expressed by management (e.g., CEO, CFO, etc.). Following Qin and Yang (2019) who show that such statistics are useful in the prediction of financial risk, PRAAT features are derived from the utterances of the management team during the EC, enriching the feedback provided to the LLM and the discourse on management’s attitudes towards present or future financial performance.<sup>2</sup>

<sup>1</sup><https://www.alphavantage.co/>

<sup>2</sup>Audio recordings are collected from Seeking Alpha (<https://seekingalpha.com/>) and force-aligned with transcripts using the Aeneas library (<https://github.com/readbeyond/aeneas>).

**Editor** (🗨️) agent ensures the generated report is suitable for the intended audience (in terms of content, style, and structure) and for checking that important information is maintained through revisions.

**Stage 2. Initial Drafting** Upon receiving the task brief from a Client agent, the Writer agent generates the initial draft of the report.

**Stage 3. Feedback and Revision** In an iterative process, each agent furnishes feedback aimed at enhancing the report, concentrating on elements relevant to their roles. Following each feedback round, the Writer updates the report. This cycle concludes when the preset maximum of  $N$  iterations is reached or upon the Client agent’s determination that the report is complete.<sup>3</sup>

### 3 Generated Report Characterization

To gain a full understanding of the style, content, and utility of generated reports, we conduct several in-depth experiments and analyses. Adopting multiple configurations of our multi-agent framework, we generate reports for a sample of 60 EC transcripts used in previous work (Mukherjee et al., 2022), basing our sample size on previous work for multi-agent text generation tasks (Chan et al., 2023; Wang et al., 2024).

#### 3.1 Generated vs. Human-Authored Reports

Given that our analytical reports are generated in a zero-shot setting, there is no guarantee that they will closely resemble those written by human experts. Therefore, to answer **RQ1**, we start by analyzing the similarities and differences between generated reports and those produced by human experts. For human-authored reports, we use a sample of 26 equality research reports from the Bloomberg Terminal that are authored by professional analysts at J.P. Morgan, to which we were granted restricted access.

**Content** To identify the key topics of discussion within each report type, we employ the aspect-extraction method outlined by Tulkens and van Cranenburgh (2020), we use SpaCy (Honnibal and Montani, 2017) to extract the 250 most frequent nouns in the earnings call transcripts of ECTSumm and analyze their presence in both generated and

<sup>3</sup>We set the value of  $N$  to 10 for all experiments, but find that it rarely reaches this threshold without being stopped by the Client.

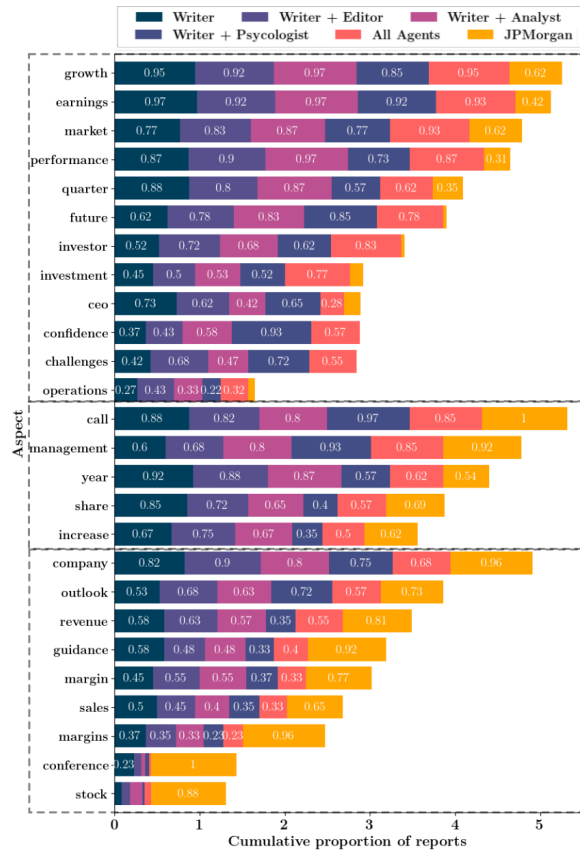


Figure 2: A visualization of the proportion of aspect occurrences for different report types, including at least the 10 most common aspects for each.

reference (JP Morgan) reports. Figure 2 presents the results of the content analysis. Using the Figure we can see that, for example, the topic of “growth” appears in 95% of analytical reports generated with all agents (All Agents), but only in 65% of the human-authored reports of JP Morgan. To enhance clarity, we’ve divided the figure into three segments based on the similarity of aspect occurrences between the generated reports (with all agents) and the reference reports. The top segment represents aspects that occur more frequently in the generated reports, the middle segment includes aspects with similar frequencies, and the bottom segment comprises aspects that occur more frequently in the reference reports. The Figure shows that, although generated and human-authored reports can be seen to discuss aspects like “share”, “management”, and “increase” at a similar rate, there exists a significant divergence in content emphasis. For instance, human-authored reports place a greater focus on financial statistics, with aspects like “margin(s)”, “revenue”, “sales”, and “stock” occurring more frequently. In contrast, generated








Agents	# Sents	FKGL	CLI	ARI	Abst
	24.35	12.88	16.42	16.87	41.74
	22.90	13.67	17.55	17.83	48.03
	21.43	13.44	17.32	17.24	49.46
	20.03	15.71	19.03	20.26	57.95
	19.65	14.76	18.33	19.10	53.40
	19.68	15.69	19.18	20.11	56.87
	18.58	15.11	18.98	19.46	56.72
References	19.25	7.26	8.54	8.85	47.14

Table 1: Readability with different feedback agent configurations.

reports can be seen to emphasize aspects such as “performance”, “future”, “earnings”, and “market” which, whilst different from references, remain indicative of analytical discussion. Furthermore, generated reports introduce several aspects that very rarely occur in reference reports (or do not occur at all), including “investor”, “investment”, and “confidence” implying they address their intended audience more explicitly.

**Style** Table 1 presents the statistical results of several metrics selected to measure the stylistic properties of reports. Here, we employ readability metrics Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975), Coleman-Liau index (CLI) (Coleman and Liau, 1975), and Automated Readability Index (ARI) (Senter and Smith, 1967), to assess text complexity (Goldsack et al., 2023). These widely used metrics are calculated based on such factors as the number of characters (ARI and CLI), syllables (FKGL), words (all), and sentences (all) present in a given text. Higher scores indicate greater document complexity. To provide insights on document length and content novelty, we also calculate the number of sentences using NLTK (Bird et al., 2009) and the abstractiveness (% of unigrams used that do not occur in the source transcript).



Here, we can see that, although both generated and human-authored reports are generally of a similar length and level of abstractiveness, there is a significant divergence in the scores of readability metrics. Expert-written reports obtain readability scores ranging from 7-10, deemed suitable for a majority of marketing materials, and indicative of shorter statement-like sentences (i.e., containing fewer syllables or characters). Contrastingly, readability scores of generated reports span a range of 12-20, indicative of longer more complex sentences and aligning with materials intended for a highly skilled readership, such as academic publications.

### 3.2 Impact of Feedback Agents

Again utilizing Figure 2 and Table 1, we can begin to assess the impact each agent has on the output report. Firstly, the incorporation of both the Editor and the Analyst agents is shown to have a similar effect on content, increasing the rate at which aspects like “outlook”, “market”, “management”, and “future” are discussed when compared to the Writer agent alone. This implies that both the additional financial statistics introduced by the Analyst and the critical feedback of the Editor induce a broader and more speculative analysis of company performance, causing the content to transition from focusing primarily on reporting the facts and figures from the transcript to a more speculative and potentially more insightful discussion. Table 1 supports this, showing an increase in abstractiveness (Abst) upon the introduction of each agent, suggesting that report content becomes less based on the source transcript and more based on external data and agent discussion. Additionally, readability metrics (FKGL, CLI, and ARI) can also be seen to increase, denoting the use of longer and more complex words/sentences.

For the Psychologist agent, Figure 2, shows a significant decrease in the reporting of aspects relating to financial performance figures (“year”, “share”, “quarter”, “increase”, “revenue”, “margins”) in favor of aspects relating to the attitude and confidence of management (“management”, “confidence”, “future”, “outlook”), areas that this agent was designed to focus on. In addition to the change of focus, Table 1 suggests a significant change in the style of reporting. More than any other agent, the Psychologist causes readability and abstractiveness metrics to rise, demonstrating its ability to influence the generated text through the introduction of novel content.

### 3.3 Insightfulness of Generated Reports

To answer RQ2, and determine how effective a multi-agent approach is at providing insights that are potentially useful to an investor, we conduct an in-depth human evaluation utilizing domain experts. We employ three evaluators, all of whom are pursuing postgraduate studies in Finance, and ask them to assess reports generated by the Writer () alone with those produced using all agents () for 32 randomly-selected earnings call instances, allowing us to see the impact of our designed feedback agents. Specifically, the evalua-

Report characteristic	Description
Financial takeaways	The key financial details from the meeting (i.e., numerical statistics relating to company performance for the quarter).
Financial context	Any additional information (e.g., financial details from previous quarters) that helps to contextualize the current financial performance.
Management attitudes	Information on how management (e.g., CEO, CFO, etc..) feels about the company’s financial performance.
Management expectation	Details about how the company is expected to perform in the future/next quarter.
Possible future events	Details surrounding any noteworthy events/scenarios that are likely to occur in the future.

Table 2: Human evaluation assessment criteria descriptions.

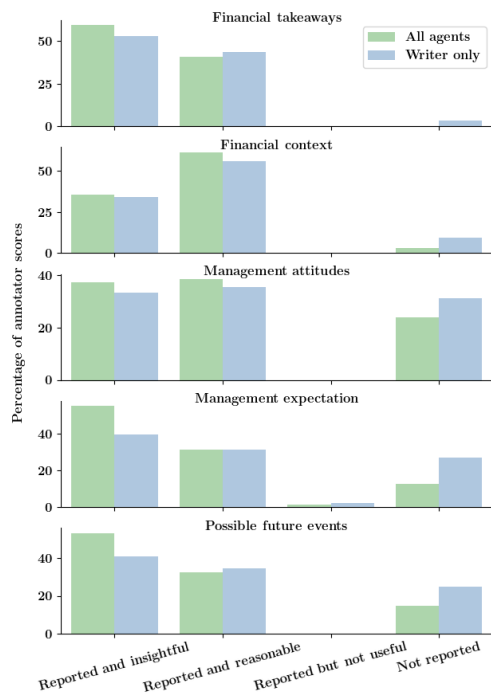


Figure 3: Characteristic-based human evaluation results.

tors’ task is broken down into the assessment of the following key report characteristics: 1) Financial takeaways, 2) Financial context, 3) Management attitudes, 4) Management expectations, 5) Possible future events. For each characteristic, evaluators assign one of the following labels, identifying both the presence and insightfulness of relevant information: 1) Reported and insightful, 2) Reported and reasonable, 3) Reported but not useful, and 4) Not reported. Our evaluation characteristics and labels were developed in collaboration with professional analysts, and are crucial in capturing report value to potential investors. Full descriptions of our characteristics and labels are provided in Tables 2 and 11 (Appendix C), respectively.<sup>4</sup>

<sup>4</sup>A discussion of inter-annotator agreement is also provided in Appendix C.

**Evaluation results** The distribution of evaluator ratings for each system is depicted in Figure 3. Firstly, it can be observed that, for both systems, the percentage of positive labels (i.e., “reported and insightful” and “reported and useful”) significantly outweighs the more negative labels for all characteristics. Additionally, there are almost no cases of the “reported but not useful” label for either system, attesting to the quality of the information that is included in the report. Of all the characteristics that are discussed in the reports generated by both systems, Financial takeaways are the most frequently included, whereas Management attitudes are the most frequently absent.

Results also indicate that the inclusion of all agents leads to a decrease in the percentage of characteristics labeled as “not reported” and an increase in the percentage of characteristics deemed “reported and insightful” in all instances. Particularly notable improvements are observed for the management expectations and possible future events characteristics, illustrating how the additional feedback agents significantly improve the speculation and anticipation of future performance in generated reports, factors that are particularly relevant to investment decision-making.

**Case study** To gain a better understanding of insights generated by the multi-agent framework, we perform a closer examination of reports and annotator classifications, presenting examples of our findings in the form of a case study in Figure 4. The given examples demonstrate how additional agent feedback can make the reporting improve the reporting of important information. Looking at each of the provided instances, example A shows how the same financial takeaways are presented alongside additional insights for investors (a likely influence of the Editor); example B illustrates how novel content relating to the confidence of man-

<p><b>A. Financial takeaways</b></p> <p><b>Writer only</b> - Motorola Solutions showcased a robust financial performance in the third quarter of 2021. The company reported a revenue increase of 13% and a surge in earnings per share (EPS) by 21%, signaling strong demand across businesses. Operating margins also expanded by 150 basis points. A record backlog of \$11.4 billion was achieved, marking a 7% year-over-year increase.<sup>1/3</sup></p> <p><b>All agents</b> - Motorola Solutions persists in its upward financial trajectory, as illustrated by pivotal metrics:</p> <p>- <i>Revenues</i>: A robust 13% increase substantiates the company's firm market footing and demand for its solutions. - <i>Earnings Per Share (EPS)</i>: The 21% rise in EPS reflects burgeoning profitability, a reassuring signal for shareholders. - <i>Operating Margins</i>: The expansion by 150 basis points testifies to enhanced operational rigor and disciplined cost governance. - <i>Record Backlog</i>: The \$11.4 billion figure – up 7% year-over-year – fortifies visibility into future revenue streams.<sup>3/3</sup></p>
<p><b>b. Management attitudes</b></p> <p><b>Writer only</b> - As the economy continues to transition, Dynex remains cautiously optimistic about capitalizing on emerging opportunities while maintaining its established principles of responsible money management and transparency.<sup>1/3</sup></p> <p><b>All agents</b> - The managers' communication styles, marked by confidence and clarity, supplemented the strong financial results by providing an indication of management's collective poise and tactful approach to stewardship, which might positively influence investor sentiment.<sup>3/3</sup></p>
<p><b>c. Management expectation</b></p> <p><b>Writer only</b> - In regards to the outlook for Q4 of fiscal '22, RPM projected consolidated sales growth in the low teens versus the prior year, despite anticipated supply chain challenges and currency headwinds due to the strengthening U.S. dollar. Notwithstanding these obstacles, it was suggested that RPM could achieve record EBIT on a consolidated basis in Q4.<sup>1/3</sup></p> <p><b>All agents</b> - <i>Growth Prospects</i>: RPM foresees low teens growth in consolidated sales and adjusted EBIT for Q4. This would continue the trend of sustained growth even when set against a challenging economic backdrop, including impacts from the Ukrainian conflict.<sup>3/3</sup></p>

Figure 4: A case study comparing how the introduction of feedback agents changes the discussion of key aspects. Colors and superscript are used to denote the number of evaluators who judged the sentence as “reported and insightful (e.g., <sup>3/3</sup> = 3 out of 3 evaluators).

agement is introduced by the Psychologist’s audio analysis; and example C demonstrates how past financial from the Analyst can contextualize relevant facts.

## 4 Evaluating Analytical Reports with LLMs

In this section, we address **RQ3** and attempt to establish how future works may address the challenge of evaluating generated analytical reports. Specifically, we first describe the challenges of a conventional reference-based evaluation (§4.1), before exploring the potential of LLMs for reference-free evaluation (§4.2).

### 4.1 Challenges of Reference-based Evaluation

Significantly, the more conventional method of assessing generated reports against references using automatic metrics faces several inherent limitations in the context of analytical reports. Firstly, the scarcity of available data samples is compounded by the fact that these reports, being internally generated by corporate entities, pose challenges in establishing public benchmarks. The small dataset of 26 reports we’ve gathered is subject to strict redistribution restrictions, precluding us from making them publicly accessible for benchmark creation. Even if such access were feasible, these reports typically adhere to in-house guidelines and practices,

contributing to disparities in content and style between human and machine-generated reports such as those outlined in §3.1. Accordingly, any novel insights provided during generation are unlikely to be adequately captured or rewarded through reference-based comparison. Furthermore, these instances are based on Earnings Calls (ECs) considerably older (2012-2016) than others utilized in this study (2019-2022), which played a role in our granted access. This raises questions about their usefulness as a potential point of comparison, considering a possible evolution of financial reporting practices.

### 4.2 Reference-free Evaluation with LLMs

Given the limitations described in §4.1, we explore the use of LLMs for the reference-free evaluation of generated outputs, a direction that has proved promising in previous studies (Liu et al., 2023; Luo et al., 2023; Chan et al., 2023). Utilizing their respective APIs, we experiment with GPT-4 (OpenAI, 2023), Gemini-pro (Gemini, 2023), and Mistral-medium (Jiang et al., 2024), instructing each to embody a financial expert and reenact evaluations performed by experts. We assess the performance of LLM evaluators in two popular human evaluation settings: 1) a characteristic-based setting and 2) a preference-based (ranking) setting. The prompts used for each setting are provided in Appendix B, Table 7.

Characteristic	GPT-4			Gemini-pro			Mistral-medium		
	$\gamma$	$\rho$	$\tau$	$\gamma$	$\rho$	$\tau$	$\gamma$	$\rho$	$\tau$
Financial Takeaways	0.375	0.160	0.412	0.156	0.018	0.014	0.139	0.205	0.192
Financial Context	0.597	0.455	0.397	0.341	0.330	0.292	0.758	0.437	0.397
Management Attitudes	0.570	0.524	0.463	0.248	0.301	0.266	0.463	0.558	0.492
Management Expectation	0.529	0.511	0.441	0.643	0.598	0.521	0.670	0.661	0.581
Future Events	0.472	0.379	0.327	0.179	0.194	0.167	0.422	0.382	0.330
Average	0.509	0.405	0.408	0.313	0.288	0.252	0.490	0.449	0.398

Table 3: Correlation statistics of LLMs vs. human evaluators (averaged) for each report characteristic.

**Characteristic-based Evaluation** For a characteristic-based evaluation, LLM evaluators are tasked with performing the evaluation described in §3.3. Here, we adhere to established evaluation methodologies from previous studies (Zhong et al., 2022; Chan et al., 2023) and employ Pearson ( $\gamma$ ), Spearman ( $\rho$ ), and Kendall ( $\tau$ ) correlation coefficients between LLM and human evaluators.<sup>5</sup> Table 3 presents our findings. Here, we can see that GPT-4 and Mistral obtain a good level of correlation with human experts, whereas Gemini-pro performs slightly works in terms of average correlation scores.

Looking closer at individual characteristics, all models can be seen to achieve a strong level of correlation ( $> 0.5$ ) for at least one of the listed characteristics and GPT-4 maintains at least a moderate level ( $> 0.3$ ) of correlation across all characteristics. Contrastingly, Gemini and Mistral achieve a broader range of correlation scores, with particularly strong scores for some aspects (e.g., Management Expectations), but weaker scores for others (e.g., Financial Takeaways).

Overall, these results indicate that LLMs have significant potential in the evaluation of analytical reports when assessing fine-grained characteristics, but that performance is likely to differ depending on the LLM. Although GPT-4 can be considered the best all-round evaluator, the fact that different LLMs achieve stronger correlations for specific characteristics is something that future works should consider when designing their evaluation.

**Preference-based Evaluation** In addition to the characteristic-based evaluation, we perform a preference-based evaluation utilizing the professional analytics reports we collect from JP Morgan (described in §3.1). Specifically, evaluators are required to compare the reference report and the

<sup>5</sup>To calculate correlation, we convert our labels into numeric scores ranging from 1-4, with 4 being the most positive classification (reported and insightful) and 1 being the most negative (not reported).

Report	GPT-4		Gemini-pro		Mistral-med	
	#1	#2	#1	#2	#1	#2
Generated	100.0	70.83	87.5	100.0	91.67	16.67
Reference	0	29.17	12.5	0.0	8.33	83.33

Table 4: The % of preference annotations given by each LLM evaluator for generated (📊🔍) and reference reports. #1 = generated report given first in prompt, #2 = reference report given first in prompt.

report generated by our system with all possible agents (📊🔍)<sup>6</sup> Here, we integrate argument quality evaluation principles (Gretz et al., 2020), which hinge on *whether evaluators would recommend a friend to use that argument as is in a speech supporting/contesting the topic, regardless of personal opinion*. In our case, evaluators indicate which report they would recommend to someone who would be making an investment decision based on the information released in the EC.

After conducting this evaluation using the same three experts as in our characteristic-based evaluation, we find that there remains a large preference for human-authored reports over generated reports, with human annotators preferring the reference reports of JP Morgan 83.33% of the time (on average). To gain further insights into human preferences, we conducted in-depth interviews with human annotators, which revealed the general preference for reference reports was attributed to their detailed, forward-looking evaluations, comprehensive risk assessments, and specific financial performance forecasts. These reports adeptly juxtapose company guidance against market expectations and consider the implications of company policies and regional market dynamics. This feedback serves as a cornerstone for future research on generating analytical reports for ECs.

Table 4 presents the results of LLM evaluators for this preference-based evaluation. Given that

<sup>6</sup>Note that, due to these ECs being from earlier dates, their audio recordings are unavailable and we were unable to include our Psychologist agent in report generation process for these instances.

previous work (Wang et al., 2023) has identified a tendency of LLMs to favor the first displayed candidate in a ranking scenario, we perform this experiment with both possible candidate orderings. Interestingly, we find that only Mistral exhibits the strong positional bias described by Wang et al. (2023). For GPT-4 and Gemini-pro, the stronger bias is shown toward generated outputs, with both models overwhelmingly favoring them regardless of candidate orderings, starkly contrasting with human experts. Furthermore, in all cases, the models are largely inconsistent across both runs. These factors highlight serious limitations in using LLMs to assess the overarching quality of analytical reports for ECs, particularly in ranking scenarios involving both human- and machine-generated outputs.

## 5 Related Work

### 5.1 LLMs as Task-solving Agents

The deployment of multiple LLMs to collaboratively work on a task has recently emerged as a trend in NLP research. While one branch of this research has typically sought to answer if adopting such an approach can improve collective reasoning (Du et al., 2023; Liang et al., 2023) another has explored how the unique opportunities afforded by this approach might allow us to attempt yet more complex tasks (Qian et al., 2023; Wu et al., 2023; Chan et al., 2023; Wang et al., 2024). Of these, our work is in closer alignment with the latter. Flexible and generic frameworks such as Autogen (Wu et al., 2023), utilized in this work, have recently emerged, enabling the development of agent-based approaches that are easily customizable and utilize external tools. Related studies have taken the initial steps in exploring challenges such as assessing generated text by employing multiple agents (Chan et al., 2023), establishing the importance of diverse agent roles/personas. Similarly, Wang et al. (2024) leverages multiple language model personas to enhance performance in tasks demanding knowledge and reasoning, such as creative writing based on trivia and solving logic puzzles. In contrast to these prior efforts, our focus centers on a more specialized task, necessitating the development of agents with domain expertise and benefiting from the incorporation of external data.

### 5.2 Earnings Call Processing

As mentioned in §1, ECs have proved a popular topic of study for previous work, due largely

to their significance in investor decision-making. Of these works, the most related to ours is that of Mukherjee et al. (2022) which introduces and benchmarks ECTSumm, a dataset for the generation of journalistic EC reports. However, in contrast to this work, we focus on generating analytical reports using a multi-agent framework, a notably more challenging task.<sup>7</sup>

Another more well-explored branch of EC processing focuses on the utilization of transcripts as the source documents for predictive NLP tasks, including volatility prediction (Sawhney et al., 2021; Niu et al., 2023), analyst decision prediction (Keith and Stent, 2019), financial risk prediction (Qin and Yang, 2019), and earnings surprise prediction (Koval et al., 2023). In contrast to these works, we tackle a complex generation task, although we in inspiration from their findings. For instance, the inclusion and design of our Psychologist agent is influenced by the work of Qin and Yang (2019), who demonstrate the efficacy of features based on EC audio in the prediction of financial risk.

## 6 Conclusion

This study explores the novel task of generating analytical reports for ECs. Following an investigation of the key distinctions between analytical reports and previously studied journalistic reports, we address the generation of analytical reports using an LLM-based multi-agent framework. We perform a thorough characterization of generated reports, revealing key divergences with human-authored reports while also highlighting the ability of agents to introduce useful insights. Finally, we address the open challenge of evaluation using LLMs, establishing the utility of different setups, and laying the groundwork for future research. Here, our findings illustrate a detrimental tendency for LLMs to favor generated over human-authored reports, but reveal that LLMs largely achieve good alignment with human experts when it comes to evaluating fine-grained criteria. While our framework aims to generate insightful analytical reports, there remains a significant opportunity to explore generative techniques that can produce novel insights. Future research could greatly benefit from incorporating real-time financial data, news, and market trends for more useful analyses.

---

<sup>7</sup>We provide an analysis of the differences between journalistic and analytical reports in detail in Section A of the Appendix.



## Limitations

Given the flexibility of multi-agent frameworks, there are undoubtedly many alternative options in terms of agent design and interaction that could prove beneficial and are worthy of investigation. However, as this work represents a first exploration of this task, we primarily concentrate our research efforts on establishing knowledge and practices that will benefit future work, for example, in the form of our generated report characterization (§3), and in the investigation of LLM-based evaluation (§4).

As discussed in the paper, another limitation in exploring the task of generating analytical reports in open research is the lack of suitable reference reports, an issue that naturally arises from their typically corporate origins. We attempt to address this issue by exploring the use of Large Language Models for reference-less human-style evaluation and hope that these findings will have a positive impact on future research on this task.

## Acknowledgements

This paper is based on results obtained from a project JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO), and JSPS KAKENHI Grant Number 23K16956. This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

## References

- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc."
- Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. [Chateval: Towards better llm-based evaluators through multi-agent debate](#).
- Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. [Improving factuality and reasoning in language models through multiagent debate](#).
- Linda Flower and John R Hayes. 1981. A cognitive process theory of writing. *College composition and communication*, 32(4):365–387.
- Team Gemini. 2023. [Gemini: A family of highly capable multimodal models](#).
- Tomas Goldsack, Zhihao Zhang, Chen Tang, Carolina Scarton, and Chenghua Lin. 2023. [Enhancing biomedical lay summarisation with external knowledge graphs](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8016–8032.
- Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Asaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7805–7813.
- Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th ophile Gervet, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. 2024. [Mixture of experts](#).
- Katherine Keith and Amanda Stent. 2019. [Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 493–503, Florence, Italy. Association for Computational Linguistics.
- J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.
- Ross Koval, Nicholas Andrews, and Xifeng Yan. 2023. [Forecasting earnings surprises from conference call transcripts](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8197–8209, Toronto, Canada. Association for Computational Linguistics.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. [Encouraging divergent thinking in large language models through multi-agent debate](#).
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruo Chen Xu, and Chenguang Zhu. 2023. [G-eval](#):

NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for text summarization.

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.

Hao Niu, Yun Xiong, Xiaosu Wang, Wenjing Yu, Yao Zhang, and Weizu Yang. 2023. KeFVP: Knowledge-enhanced financial volatility prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11499–11513, Singapore. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Chen Qian, Xin Cong, Wei Liu, Cheng Yang, Weize Chen, Yusheng Su, Yufan Dang, Jiahao Li, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development.

Yu Qin and Yi Yang. 2019. What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 390–401, Florence, Italy. Association for Computational Linguistics.

Ramit Sawhney, Arshiya Aggarwal, and Rajiv Ratn Shah. 2021. An empirical investigation of bias in the multimodal analysis of financial earnings calls. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3751–3757, Online. Association for Computational Linguistics.

RJ Senter and Edgar A Smith. 1967. Automated readability index. Technical report, Technical report, DTIC document.

Stéphan Tulkens and Andreas van Cranenburgh. 2020. Embarrassingly simple unsupervised aspect extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3182–3187, Online. Association for Computational Linguistics.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators.

Report	# Sents	FKGL	CLI	ARI	Abst
Journalistic	4.2	5.73	8.78	7.64	42.06
Analytical	19.25	7.26	8.54	8.85	47.14

Table 5: Average statistics of journalistic and analytical reports.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2023–2038, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

## A Journalistic vs. Analytical Reports

Here we perform a comparative study on the style and content of journalistic and analytical reports. For journalistic reports, we utilize the references of ECTSum (Mukherjee et al., 2022), a dataset consisting of EC transcripts paired with bullet-point summaries derived from online Reuters articles that report on the key financial takeaways from the EC for a general audience.<sup>8</sup> For analytical reports, we utilize the J. P. Morgan samples discussed in the main text.

**Style** The results in Table 5 show that two out of three readability metrics indicate that analytical reports are more structurally complex than journalistic reports, featuring longer and more intricate sentences. Higher abstractiveness also indicates that the content of analytical reports is less directly derived from the source transcript. These factors are indicative of the anticipated complexity in the discussion found in analytical reports, aligning with their purpose of offering in-depth insights to potential investors.

<sup>8</sup>Reuters article example: <https://tinyurl.com/yc3z9sbj>

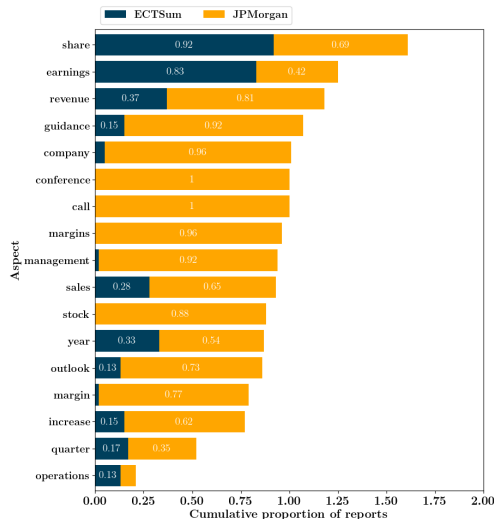


Figure 5: The most commonly discussed aspects of each report journalistic and analytical reports.

**Content** Figure 5 illustrates the predominant topics for each report type. For instance, the Figure shows that the topic of “earnings” appears in 42% of analytical reports compared to 83% of journalistic reports, indicating that while “earnings” is a significant aspect, analysts do not always explicitly focus on it as much as journalists do. Whereas the shorter journalistic reports are shown to concentrate on a few select topics, longer analytical reports are shown to cover a much broader range of topics, likely caused by the divergence in the target audience. Whereas journalistic reports are intended for a more general audience looking for key financial statistics, reports produced by professional analysts at J.P. Morgan are largely aimed at internal investors and are tailored as such. For instance, topics such as management, “outlook”, “guidance”, and “increasing” are all representative of a more in-depth discussion that goes beyond the key statistics, addressing the attitudes of management, analyzing trends in financial performance, and speculating about the future.

Overall, this divergence not only underscores the varied emphases between report types but also reveals that analysts are more inclined towards forward-looking analyses, whereas journalists predominantly concentrate on summarizing key points without providing in-depth analysis. Furthermore, the complexity and diversity of discussed topics are likely to present a significant challenge to a single model working off only the EC transcript, requiring the model to follow complex and multifaceted instructions and pushing the limitations of its limited

context window. For this reason, we believe this task lends itself to a multi-agent approach, which allows us to employ specialist agents that utilize expert-based role-play and external data to address specific aspects of report analysis.

## B Prompts

**System Prompts** Here, we provide details concerning all aspects of the prompts used within our system. Notably, the initialisation prompts that are used to dictate each agent’s role within the report generation process are provided in Table 6. For the Writer, Editor, and Client agents, this is the only controlling attribute for agent behavior with the rest of their interaction being handled by AutoGen.

However, for our more specialized agents, the Analyst and and Psychologist, we implemented additional functionality to allow them to utilize external data. Specifically, these agents are implemented such that the relevant data is collected before each response, examples of which are provided in Table 8. This data is then utilized in a prompt sent to the underlying LLM that we specifically designed to extract the desired feedback. Prompt formats for both of these agents are in Table 9.

**LLM Evaluation Prompts** As stated in §4, we provide LLM evaluators with the same instructions that are given to human evaluators, for both our characteristic-based and preference-based manual evaluations. The specific prompts provided to each LLM model are given in Table 7.

## C Additional Experimental Details and Results

**Metric calculation** All readability metrics were computed using the `textstat` package.

**Analytical Report Earnings Call Samples** Details of the ECs that our reference analytical report ECs are based on are provided in Table 10. It is important to note that, given that these reports are not publicly available and we are granted restricted access to them from JP Morgan, there is very little of LLMs having encountered them in training (i.e., data contamination).

**Characteristic-based Human Evaluation Details and Annotator Agreement** We provide a full description of each characteristic for our characteristic-based human evaluation in Table 2.

Agent	Initialisation Prompt
Writer 🖋️	You are a Writer who is responsible for drafting the requested output text and making adjustments based on other agents’ suggestions. Note that, unless otherwise specified, you should avoid completely rewriting the report and focus on making smaller targeted changes or additions based on other agent’s feedback. You should only respond with updated versions of the report.
Client (Investor) 📄	You are an Investor who requires accurate investment and market analysis data to build investment strategies. You are responsible for ensuring the report contains the information that is relevant to you by providing feedback to the Writer. If you are happy with the report, respond with “TERMINATE”.
Analyst 📈	You are an Analyst, a financial expert who is responsible for determining what past financial data might be relevant to the report and explaining this data to the Writer.
Psychologist 🎧	You are a Psychologist who is responsible for using data derived from the audio recording to identify notable features (e.g., that may express confidence, doubt, or other emotional giveaways) in audio-derived statistics of management’s answers in the Q&A session that might be relevant to the report and explaining these features to the Writer.
Editor 🔍	You are an Editor who is responsible for ensuring that the output text is suitable for the intended audience (in terms of content, style, and structure) and that important information from previous revisions of the report is not lost by providing feedback to the Writer.

Table 6: Agent initialization prompts.

We also measure the annotator agreement for this evaluation, calculating pairwise Cohen’s  $\kappa$ , getting an average score of 0.171, indicative of weak agreement between annotators. We perform a close inspection of our annotator labels to identify the source of this, finding that the majority (65.25%) of all pairwise disagreements occur between labels “Reported and insightful” and “Reported and reasonable”, the difference between which represents the most subjective decision within our evaluation whereby, after deciding if relevant information is included and at least somewhat useful, annotators must judge *how* useful they personally find the reported information. If we were to treat these labels as one, we find that the pairwise Cohen’s  $\kappa$  increases significantly to 0.476, indicative of a much stronger agreement. Therefore, we do not judge this to be an issue with our evaluation. Furthermore, in presenting the results of the evaluation, we respect any differences in the opinions of expert evaluators by calculating the statistics in Table 3 based on all evaluator votes rather than performing vote aggregation (e.g., majority vote).

**LLM Correlation Results Breakdown** To provide further insight into the correlation of LLMs with expert evaluators, Table 12 provides a breakdown of the results presented in Table 3 which shows the correlation statistics between LLMs and

individual annotators.

## D Examples

**Feedback** Figure 6 visualises an example of a typical run within our system, displaying the feedback of different agents.

**Generated reports** Figures 7 and 8 contain examples of full reports generated using only the Writer agent and all agents, respectively.

Evaluation	LLM Prompt
Characteristic-based	<p># INSTRUCTIONS</p> <p>You are a financial expert tasked with evaluating a summary of an earnings call meeting intended to provide useful information to a potential investor.</p> <p># CRITERION</p> <p>You must identify whether or not the summary contains the information relating to the aspect described below and, if it does so, assess how well the information is reported.</p> <p><i>{criterion}: {description}</i></p> <p># LABELS</p> <p>Below are the possible labels you can assign to the summary based on the described criterion. Respond using only the number of the label.</p> <ol style="list-style-type: none"> <li>1. Reported and insightful: the relevant information is included in the report and is very well explained, offering additional insights/interpretations that would likely be useful to a potential investor.,</li> <li>2. Reported and reasonable: the relevant information is included in the report and is reported reasonably well, including either no insights at all (e.g., as a statement of facts) or suggesting interpretations that are unlikely to be particularly useful to a potential investor.,</li> <li>3. Reported but not useful: the relevant information is included in the report, but it is either incorrect (i.e., there is contradictory evidence in the references) or explained in a way that is likely to mislead or misinform a potential investor.,</li> <li>4. Not reported: no information relevant to this aspect is included in the report.</li> </ol> <p># SUMMARY</p> <p><i>{generated_report}</i></p> <p># ASSIGNED LABEL</p>
	Preference-based

Table 7: LLM evaluator prompt.

Agent	Data example
Analyst 	<pre>{   "fiscalDateEnding": "2021-07-31",   "reportedDate": "2021-08-20",   "reportedEPS": "5.25",   "estimatedEPS": "4.58",   "surprise": "0.67",   "surprisePercentage": "14.6288" }</pre>
Psychologist 	<pre>{   "minimum_intensity": -14.925902805117943,   "maximum_intensity": 82.11127894879778,   "mean_intensity": 51.97292655136569,   "minimum_pitch": 75.04017645717074,   "maximum_pitch": 599.378734309719,   "mean_pitch": 143.7376593336546,   "num_pulses": 51218,   "num_periods": 51217,   "mean_periods": 0.013944367276250947,   "stddev_periods": 0.05998498783743047,   "fraction_unvoiced": 0.5148897444872244,   "degree_of_voice_breaks": 0.5146557157170372,   "jitter_local": 0.02535925970608026,   "jitter_local_absolute": 0.00018326521186048697,   "jitter_rap": 0.010346084585498544,   "jitter_ppq5": 0.012399774771964451,   "jitter_ddp": 0.031038253756495632,   "shimmer_local": 0.13931305032875707,   "shimmer_localdb": 1.2625161389585877,   "shimmer_apq3": 0.05669263538384766,   "shimmer_aqpq5": 0.08401808667057334,   "shimmer_dda": 0.17007790615154297,   "hnr": 10.84568288161106 }</pre>

Table 8: Examples of external data provided to specialized agents. For the Analyst, the exemplified data is provided from the quarter before that which is reported in the EC. For the Psychologist, the exemplified data is provided for each management utterance in the Q&A Session of the EC.



Agent	Response prompts
Analyst 	Based on your expert analysis of the Earnings Call meeting and the above conversation, identify any notable features in the following statistics, derived from the audio of the meeting for each management response in the QA session and explain how and why they should be included in the report:
Psychologist 	Based on your expert analysis of the Earnings Call meeting and the above conversation, explain why and how the following earnings information from the the companys' previous quarter should be included in the report:

Table 9: Prompt formats used by specialized agents to introduce external data. Note that, after each prompt, the relevant data is printed in JSON format.

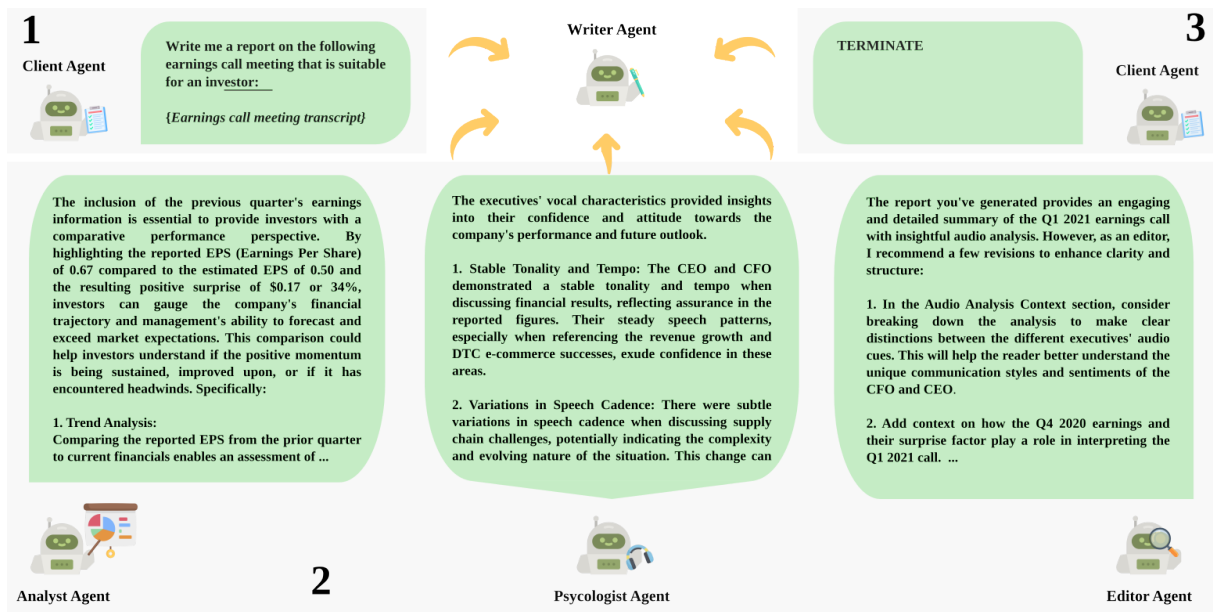


Figure 6: An example of commonly seen run within our framework where: 1) the task is given by the Client agent; 2) each feedback agent provides feedback relative to their role; and 3) the Client deems the report acceptable and terminates composition process.

Company code	Year	Quarter
CMI	2013	q4
	2014	q1
	2014	q3
	2014	q3
	2015	q1
	2015	q4
DE	2012	q4
	2013	q3
	2014	q1
	2014	q2
	2014	q3
ETN	2014	q4
	2014	q1
PCAR	2014	q1
	2014	q2
	2014	q3
	2014	q4
	2015	q1
	2015	q2
	2015	q3
	2015	q4
	2016	q1
UNH	2014	q2
WYNN	2014	q2

Table 10: The earnings call meetings from which professional J.P. Morgan reports are derived.

Label	Description
Reported and insightful	The relevant information is included in the report and is very well explained, offering additional insights/interpretations that would likely be useful to a potential investor.
Reported and reasonable	The relevant information is included in the report and is reported reasonably well, including either no insights at all (e.g., as a statement of facts) or suggesting interpretations that are unlikely to be particularly useful to a potential investor.
Reported but not useful	The relevant information is included in the report, but it is either incorrect (i.e., there is contradictory evidence on the transcript) or explained in a way that is likely to mislead or misinform a potential investor.
Not reported	No information relevant to this aspect is included in the report.

Table 11: Human evaluation annotation label descriptions.

Characteristic	GPT-4			Gemini			Mistral			
	$\gamma$	$\rho$	$\tau$	$\gamma$	$\rho$	$\tau$	$\gamma$	$\rho$	$\tau$	
Evaluator 1	Financial Takeaways	0.656	0.661	0.631	0.332	0.352	0.330	0.202	0.344	0.339
	Financial Context	0.484	0.426	0.390	0.112	0.011	0.005	0.493	0.422	0.404
	Management Attitudes	0.454	0.391	0.373	0.293	0.363	0.347	0.301	0.331	0.312
	Management Expectation	0.238	0.243	0.229	0.493	0.464	0.429	0.435	0.409	0.374
	Future Events	0.353	0.273	0.266	0.102	0.080	0.078	0.402	0.418	0.401
Evaluator 2	Financial Takeaways	0.201	-0.041	-0.046	0.174	0.007	0.005	0.167	0.127	0.124
	Financial Context	0.395	0.210	0.191	0.424	0.423	0.390	0.672	0.473	0.458
	Management Attitudes	0.436	0.429	0.412	0.157	0.173	0.168	0.506	0.485	0.461
	Management Expectation	0.579	0.587	0.542	0.639	0.554	0.496	0.728	0.719	0.665
	Future Events	0.176	0.001	0.0	0.079	0.046	0.043	0.201	0.048	0.044
Evaluator 3	Financial Takeaways	-0.165	-0.168	-0.163	-0.282	-0.265	-0.253	-0.146	-0.146	-0.146
	Financial Context	0.551	0.468	0.441	0.312	0.264	0.244	0.677	0.451	0.436
	Management Attitudes	0.505	0.512	0.483	0.163	0.125	0.119	0.348	0.360	0.340
	Management Expectation	0.560	0.483	0.462	0.319	0.319	0.301	0.405	0.380	0.353
	Future Events	0.409	0.417	0.392	0.186	0.156	0.147	0.231	0.229	0.211

Table 12: Correlation statistics of LLMs vs. human evaluators (all) for each report characteristic.



**Financial Overview**

- **Company Symbol:** CFR
- **Reported Earnings:** \$109.8 million (\$1.73 per diluted share)
- **Comparable Earnings (Previous Year):** \$115.8 million (\$1.78 per share) in Q3 2018
- **Return on Average Assets:** 1.35% compared to 1.49% in Q3 2018
- **Average Deposits:** \$26.4 billion, up from \$26.2 billion in Q3 2018
- **Average Loans:** \$14.5 billion, a 5.8% increase from Q3 2018
- **Nonperforming Assets:** \$105 million, up from \$86.4 million in Q3 2018, largely due to a single energy credit
- **Net Charge-offs:** \$6.4 million, down from \$15.3 million in Q3 2018
- **Provision for Loan Losses:** \$8 million, increased from \$2.7 million in Q3 2018
- **Problem Loans:** Ending at \$487 million, down from \$504 million in Q3 of the previous year

**Strategic Highlights and Outlook**

- Marked growth in C&I, CRE, and Consumer segments despite increased competition in CRE leading to losses due to structure in some deals.
- The weighted loan pipeline increased by about 30% over Q2, indicating a stronger Q4.
- The bank continues to open new financial centers in Houston, with the series planned expansion enhancing long-term growth prospects.
- Significant uptick in net new consumer customer growth by 48% compared to the previous year.
- Online account openings surged by 56% compared to the previous year, reflecting the bank's technology strength.
- Despite challenges due to interest rate cuts, the bank focuses on maintaining quality standards across business segments.

**Economic and Financial Detailed Insights**

- Texas job growth showed a deceleration but remains healthy overall.
- The net interest margin for Q3 was 3.76%, a decrease from 3.85% in the last quarter.
- Active management of investment portfolios and deposit pricing strategy in response to rate cuts.
- Increased noninterest expenses by about 7.8% year-over-year in part due to expansion and technology investments.
- The bank is well-positioned with robust customer satisfaction accolades and industry recognition, underpinning its service-centric strategy.

**Forward Guidance**

- For full-year 2019, the mean of analyst estimates at \$6.81 is deemed reasonable.
- Continued investment in expansion and technology could dictate a rise in noninterest expenses.
- CECL is expected to impact reserves, estimated at 15%-25% above the current allowance by the year's end.

**Discussion Points from Q&A Session**

- Loan growth optimism, albeit more competitive in pricing, was underscored by better expected performances from C&I opportunities.
- Ongoing pressure is anticipated on the net interest margin, attributable to competition and asset yield pressures.
- The continued focus remains on branch expansion in Houston, with a planned acceleration of openings in 2020, amounting to an estimated 10 by the end of 2019.
- Resolution on the new energy nonperforming loan is not imminent; management is cautious about the challenging environment for asset sales.

**Concluding Remarks**

Cullen/Frost's commitment to growth, despite short-term expense pressures, showcases long-term strategic positioning and a strong focus on customer service excellence, which bodes well for future financial performance.

Figure 7: Full example of a report generated with only the writer agent.

**Financial Overview**

- **Company Symbol:** CFR

- **Reported Earnings:** \$109.8 million (\$1.73 per diluted share). Comparison with Q2 2019's \$1.72 EPS shows incremental growth, reinforcing the company's financial stability.

- **Comparative Earnings (Q3 2018):** A slight year-over-year decrease from \$115.8 million (\$1.78 per share), which may reflect broader market trends affecting the banking industry.

- **Return on Average Assets:** 1.35%, experiencing a marginal decrease from the 1.49% reported in Q3 2018.

- **Average Deposits:** Demonstrated growth to \$26.4 billion compared with \$25.6 billion in Q2 and \$26.2 billion in Q3 2018.

- **Average Loans:** Increased to \$14.5 billion, up from \$13.7 billion in Q2. Year-over-year comparison with Q3 2018's \$13.7 billion underlines sustained growth.

- **Net Charge-offs:** \$6.4 million, escalating from Q2's \$4.5 million, yet significantly lower than Q3 2018's \$15.3 million.

- **Provision for Loan Losses:** Rose to \$8 million from \$5 million in Q2 amidst a challenging credit environment.

**Strategic Outlook and Investments**

- Management's investment in long-term strategic expansion, particularly in the Houston market, displays a commitment to future profitability and market leadership, despite impacting short-term operating leverage.

- The bank's strategic initiatives, such as opening new branches and technological investments, are expected to improve the bank's efficiency metrics in future reporting periods. However, this may temporarily reflect a higher efficiency ratio.

**Interest Rate Environment and Performance**

- The net interest margin faced pressures, partially attributed to the Federal Reserve's recent rate cuts, underscoring the sensitivity of bank operations to external rate fluctuations.

**Insights from Q&A Session**

- Management expressed cautious optimism regarding Q4 loan growth, particularly from the stronger C&I pipeline. This aligns with the pragmatic but forward-looking sentiment conveyed throughout the call.

- Concerns about single energy credit impact on nonperforming assets were tackled with measured realism, noting the challenge but expressing confidence in risk management strategies.

**Projections and External Perspectives**

- The reiteration of \$6.81 EPS guidance for the full year, despite the fluctuations observed, signals management's trust in the robustness of the bank's core operations and the expected realization of long-term strategic benefits.

- The analysis benefits from considering a range of financial analyst commentaries, providing insight into how Cullen/Frost Bankers' performance is interpreted across the investment community.

**Conclusion and Investor Takeaways**

- **Consistent Performance:** The Q3 earnings incrementally build upon Q2 2019's results, exhibiting consistency in meeting analyst estimates.

- **Strategic Growth Focus:** Long-term investments in markets and technology signal a pursuit of sustainable growth and service excellence.

- **Adaptation to Economic Conditions:** The bank continues to navigate the complexities of the current economic climate, especially in response to interest rate changes.

- **Ongoing Risk Management:** Attention to credit quality and problem loans underscores a vigilant approach to risk during economic uncertainties. Investors should be reassured by the bank's strategic consistency and responsiveness to economic indicators, reflecting a blend of ambitious expansionism tempered by conservative risk management—factors that contribute to the bank's reputation for reliability and steady growth prospects.

Figure 8: Full example of a report generated with all agents.