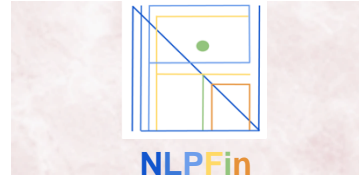


From Facts to Insights: A Study on the Generation and Evaluation of Analytical Reports for Deciphering Earnings Calls



Tomas Goldsack, Yang Wang, Chenghua Lin, **Chung-Chi Chen**

Department of Computer Science, University of Sheffield, UK

Department of Computer Science, University of Manchester, UK

Artificial Intelligence Research Center, AIST, Japan



Earnings Conference Calls

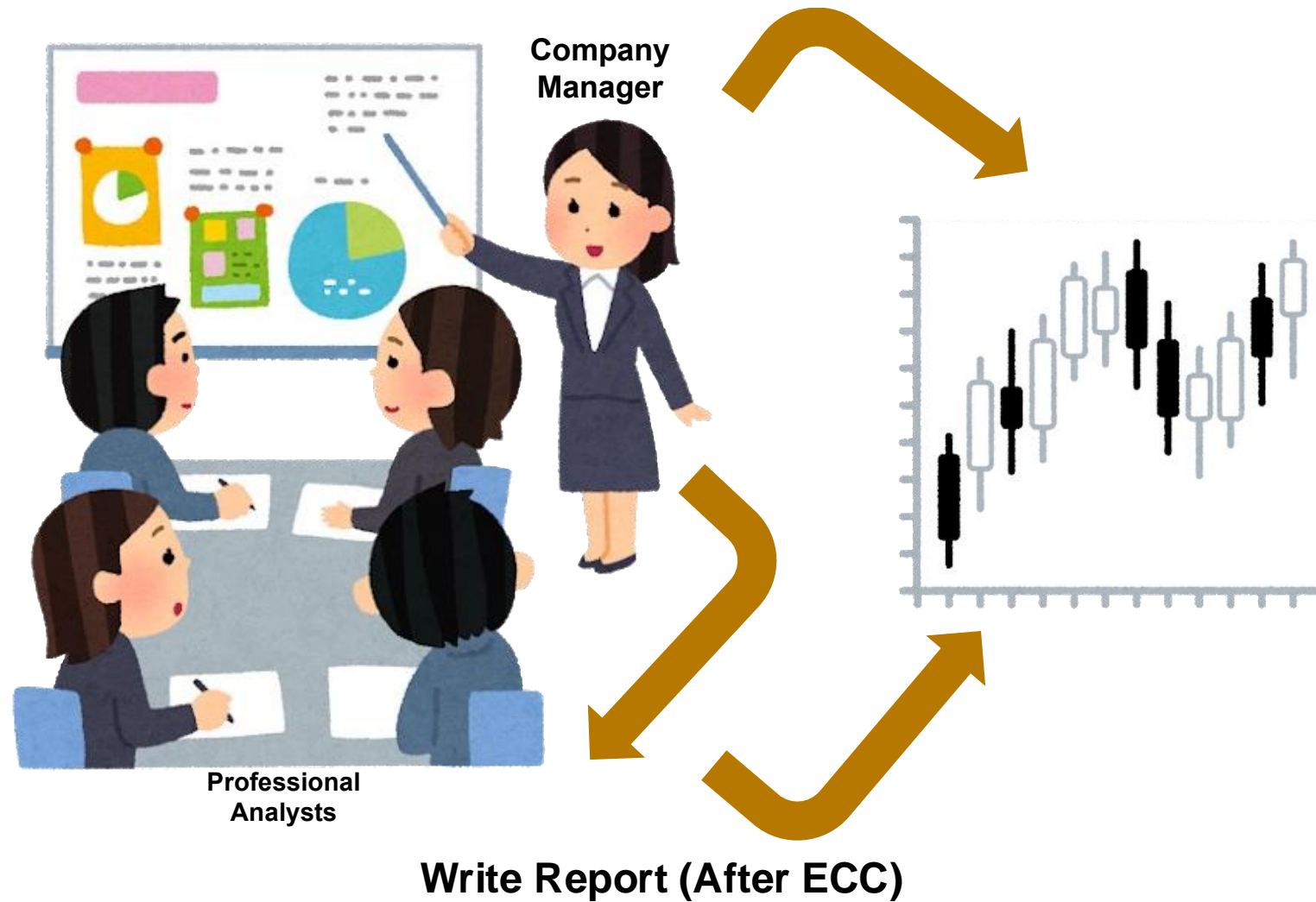


Outline:

- **Prepared Remarks**
 1. Operator
 2. Director, Investor Relations and Corporate Finance
 3. Chief Executive Officer
 4. Chief Financial Officer
- **Questions and Answers**
 1. Operator
 2. **Q: UBS – Analyst**
A: CEO
 3. **Q: Credit Suisse – Analyst**
A: CFO
 4. **Q: Credit Suisse – Analyst**
A: CEO

...

Generate Analytical Reports

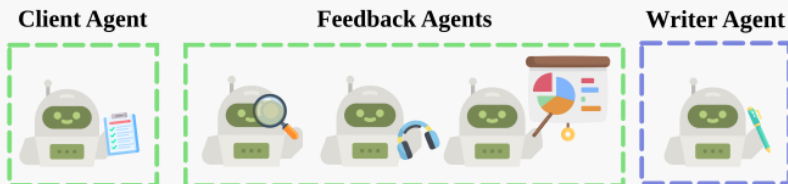


Research Questions

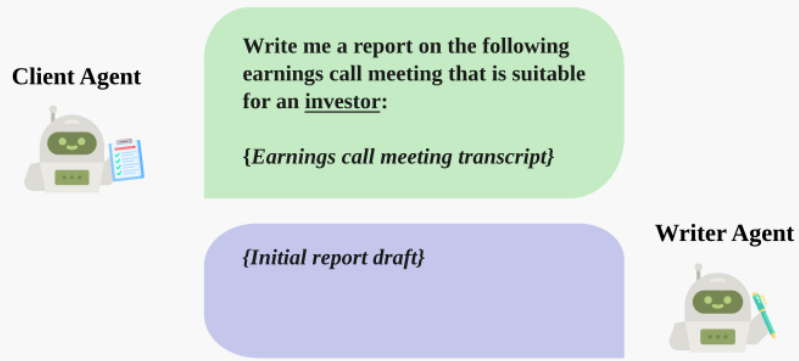
1. How do generated analytical reports differ from **human-authored** analytical reports?
 - Readability
 - Topics
2. Can a **multi-agent** approach be used to generate more insightful analytical reports?
 - Financial takeaways
 - Financial context
 - Management attitudes
 - Management expectation
 - Possible future events
3. How effective are LLM-based **evaluation** methods in assessing the quality of analytical reports?
 - Correlation
 - Preference

Approach

1. AGENT DEFINITION



2. INITIAL DRAFTING



3. FEEDBACK AND REVISION

xN



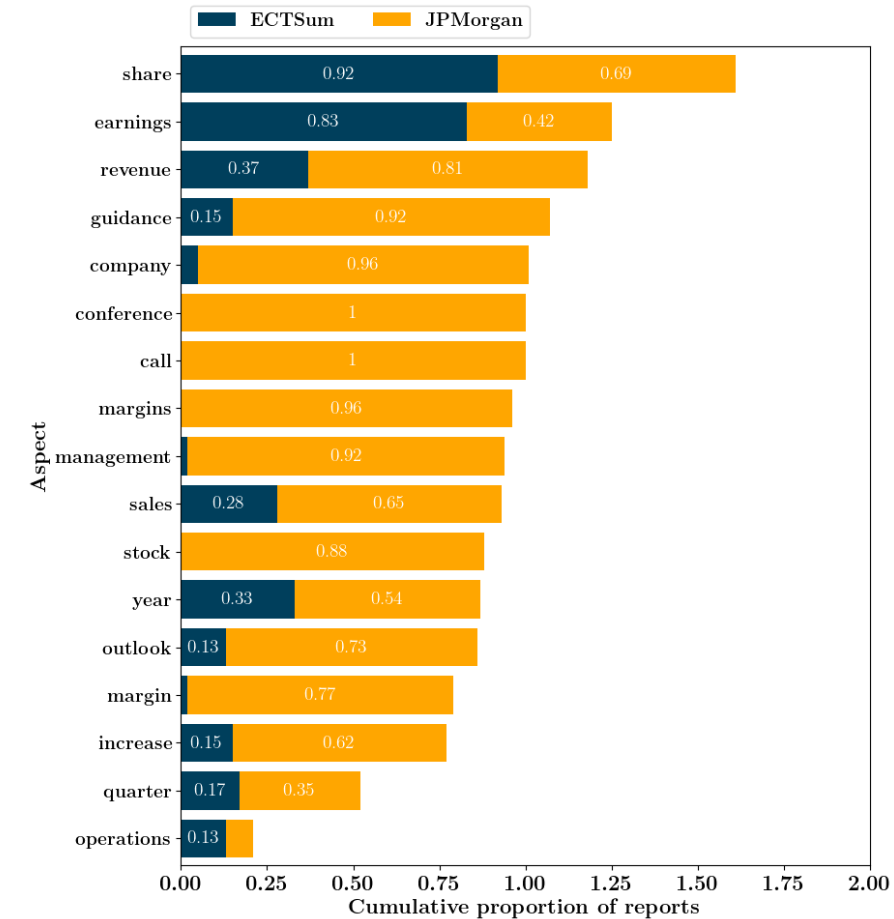
Agent	Initialisation Prompt
Writer 🖋️	You are a Writer who is responsible for drafting the requested output text and making adjustments based on other agents' suggestions. Note that, unless otherwise specified, you should avoid completely rewriting the report and focus on making smaller targeted changes 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.

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



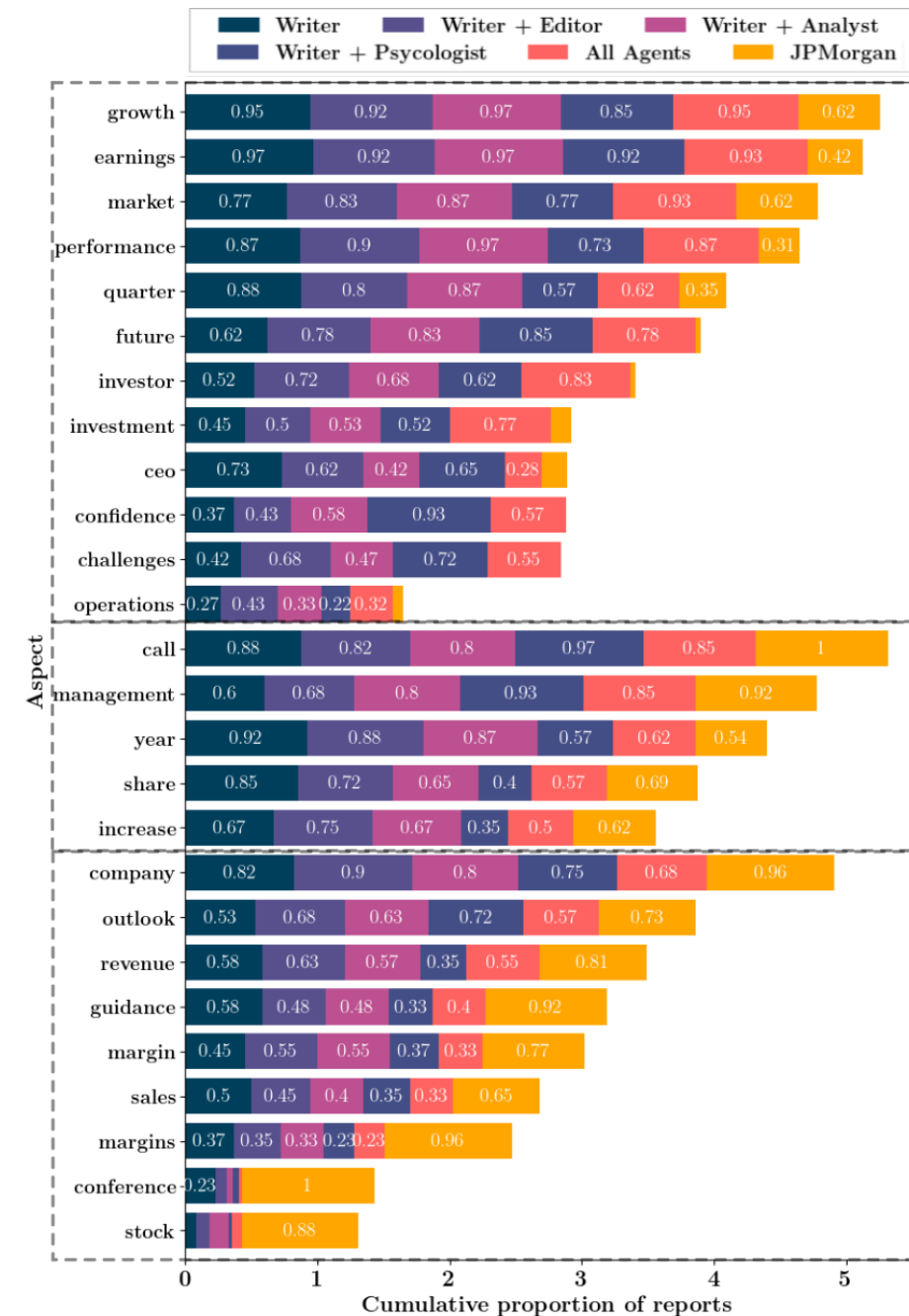
Human (Journalist) vs. Human (Analyst)

- **Topic Distribution:** Figure highlights distinct topic distributions between report types; "earnings" appear in 42% of analytical reports vs. 83% in journalistic reports, indicating analysts' less frequent but broader thematic coverage.
- **Depth of Analysis:** Journalistic reports focus on summarizing key financial statistics for a general audience. Analytical reports, aimed at internal investors, delve into deeper topics like management attitudes, future outlooks, and detailed financial trends.
- **Exclusive Topics in Analytical Reports:** Certain topics like "management," "outlook," "guidance," and "increasing" are primarily found in analytical reports, reflecting their tailored approach to meet the needs of internal stakeholders.
- **Forward-looking vs. Summary Focus:** Analysts engage in forward-looking analyses, speculating on future trends and performance. Journalists concentrate on summarizing current financial statistics and events, providing less speculative content.
- **Challenges for Automated Reporting:** The complexity and diversity of topics in analytical reports pose challenges for automation using a single model approach.










Topic – LLM vs. Analyst

- **Common Aspects:**
 - Both report types discuss aspects like "share", "management", and "increase" at comparable rates.
- **Divergence in Emphasis:**
 - Human-authored reports focus more on "margin(s)", "revenue", "sales", and "stock", indicating a **preference for detailed financial statistics**.
 - Generated reports emphasize "performance", "future", "earnings", and "market", suggesting a broader, **forward-looking analytical** approach.
 - **Unique Aspects in Generated Reports:** Aspects like "investor", "investment", and "confidence" are more prevalent, indicating explicit addressing of audience concerns and expectations.



Readability – LLM vs. Analyst

- Both report types show **similar lengths** and levels of abstractiveness, suggesting comparable content coverage.
- Significant divergence in readability metrics, with expert-written reports scoring between 7-10 and generated reports scoring between 12-20.
- Human-authored reports have lower readability scores, indicating they are more accessible and suitable for broader marketing materials with simpler sentence structures.
- **Generated reports, with their higher scores, contain longer and more complex sentences**, akin to academic texts, suggesting a style more suited to a specialized or highly skilled audience.
- **More Agent More Complex**
- The readability disparities highlight a stylistic mismatch between the generated content and potentially preferred simpler styles found in human-authored reports.
- Indicates a need for tuning generated report styles to better align with industry standards for readability and audience accessibility.

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

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

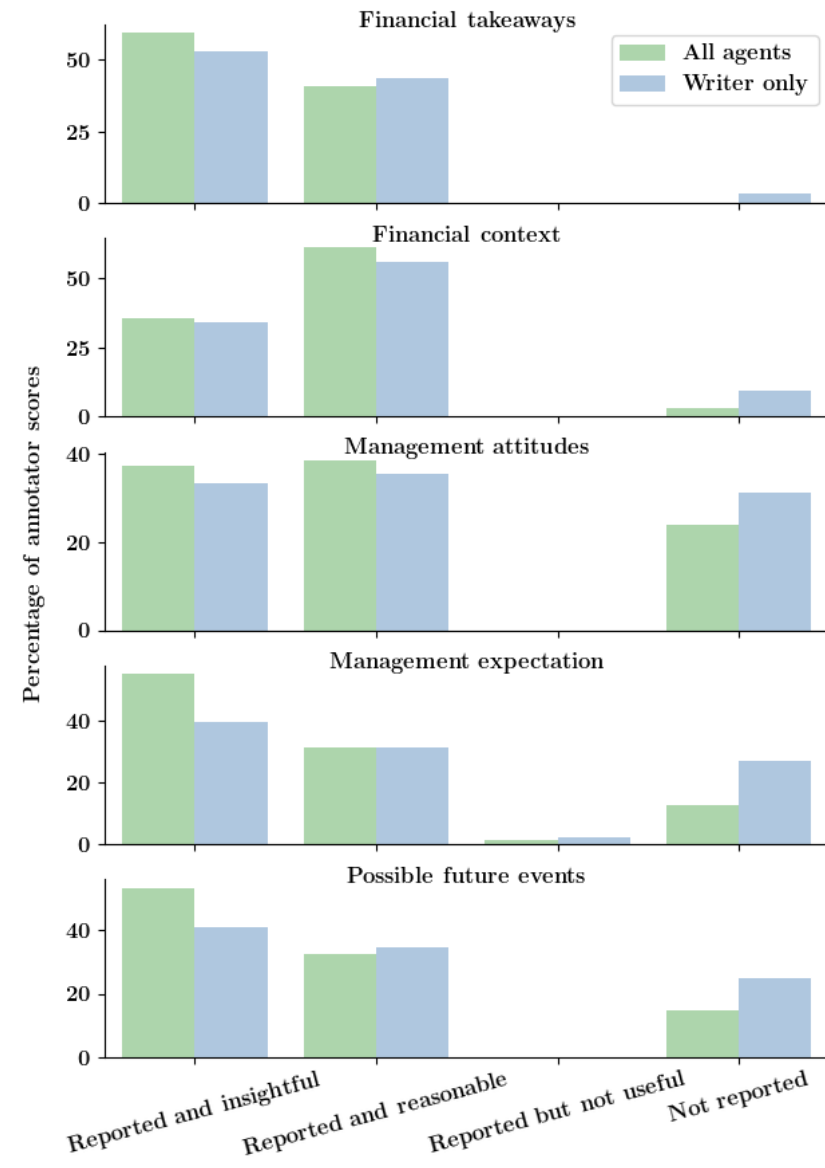


Human Evaluation Aspects

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.

Human Evaluation Results

- Predominantly **positive ratings** for both systems, with "reported and insightful" and "reported and useful" labels significantly outnumbering negative assessments.
- Financial takeaways are the most frequently included topics, indicating a strong focus on essential financial metrics.
- Management attitudes are the least reported, suggesting an area for **potential enhancement in future reports**.
- Introduction of all agents leads to **reduced occurrences of characteristics labeled as "not reported."**
- Significant increase in characteristics deemed "reported and insightful," particularly for management expectations and future events.



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



Correlation Statistics of LLMs vs. Human Evaluators

- **GPT-4 and Mistral** show good correlation levels with human experts, indicating their effectiveness in evaluating report characteristics.
- Gemini-pro displays slightly lower average correlation scores, highlighting variability in performance among different LLMs.
- All models achieve strong correlations (>0.5) for at least one characteristic, demonstrating their potential in specific areas of report evaluation.
- The variance in correlation scores among different LLMs underscores the need to **choose the right model based on specific evaluation needs**.
- GPT-4's consistent performance across different characteristics positions it as a reliable all-around evaluator for analytical reports.

Characteristic	GPT-4			Gemini-pro			Mistral-medium		
	γ	ρ	τ	γ	ρ	τ	γ	ρ	τ
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

LLM Preference



- Mistral shows a strong positional bias, confirming prior research about LLM tendencies in ranking scenarios.
- Both GPT-4 and Gemini-pro demonstrate a pronounced preference for generated outputs over expert-authored ones, regardless of the order in which they are presented.
- Human evaluators show a clear preference for expert-composed reports, in sharp contrast to the LLMs' favoring of generated content. This discrepancy highlights a potential misalignment between human judgment and LLM evaluations in assessing report quality.

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

Conclusion

- Study Overview
 - Explores the generation of analytical reports for earnings calls (ECs) using an LLM-based multi-agent framework.
 - Investigates the differences between analytical and journalistic reports.
- Key Findings:
 - Generated reports show significant divergences from human-authored reports in both style and substance.
 - Agents in the framework are capable of introducing useful insights, enhancing the depth and relevance of generated reports.
- Evaluation Challenges and Results:
 - LLM evaluations exhibit a bias favoring generated reports over human-authored ones.
 - Despite this bias, LLMs generally align well with human evaluators on fine-grained evaluation criteria, indicating effective performance in detailed analytical tasks.
- Implications for Future Research:
 - Highlights the need for further development in generative techniques to produce novel insights.
 - Suggests incorporating real-time financial data, news, and market trends to enhance the utility and relevance of generated analytical reports.
- Future Research Directions:
 - Encourages exploration of advanced generative models and data integration techniques to improve the quality and insightfulness of analytical reports.
 - Proposes further studies on reducing LLM biases and aligning automated evaluations more closely with human judgments.

Related Events

- ACL Special Interest Group on Economic and Financial Natural Language Processing (SIG-FinTech)
 - <https://sigfintech.github.io/index.html>
- The 10th Workshop on Financial Technology and Natural Language Processing
 - EMNLP-2025, Nov. 5th-9th, 2025, Suzhou, China
 - Call for Paper
 - <https://sigfintech.github.io/finnlp.html>
- Financial Information Access and Evaluation (FinEval)
 - Call for Task Proposal
 - Task proposals Due: 31 January 2025
 - <https://sigfintech.github.io/fineval.html>

