Capturing Analysts' Questioning Strategies in Earnings Calls via a Question Cornering Score (QCS)

Giulia D'Agostino, Andrea Rocci
IALS
Università della Svizzera italiana
Switzerland
{name.surname}@usi.ch

Chris Reed
Centre for Argument Technology
University of Dundee
Scotland, UK
c.a.reed@dundee.ac.uk

Abstract

The study of questions in the setting of dialogical interactions in corporate communication has the purpose of understanding and capitalizing on the opinions that the questioner has with respect to the questioned company. Particularly, financial analysts have the maximal incentive to be right in their forecasts about the company's performance, but they are also incentivized and expected to maintain a good relationship with the management – and therefore, not to be too challenging in their questions. While avoiding overt adversarialness, analysts adopt alternative strategies to seek the desired information; among which modulating the *cornering* quality of questions. This paper presents a way of measuring such cornering property, automatically extracting feature scores, and comparing the results with a manually annotated gold standard. Results encourage further research along this stream, particularly towards the study of replies and their degree of *answerhood* with respect to the cornering quality of the prompting question.

1 Introduction

Multiple strategies can be put in place to make a question more effective, that is, to enhance its likelihood of eliciting a reply maximally compliant with the type of answer that the questioner wished to obtain.

In political press conferences (henceforth: PPCs) the questioner role is held by journalists. In this context, questioners have the incentive of being on the side *opposite* to the politician they are interviewing, if need be; this because the institutional role of journalists is to gain information *to the benefit of the public*. An effective question, in this environment, may feature an openly confrontational stance towards the interlocutor in order to obtain the sought after information and/or hold the interlocutor accountable. Therefore, adversarialness (Clayman and Heritage, 2002; Clayman et al., 2007) is a measure that captures the confrontational attitude with which journalists load their questions.

In earnings conference calls (henceforth: ECCs), on the other hand, questioners are financial analysts. Whilst representing the interests of investors they also need to maintain a good relationship with the interlocutor (Palmieri et al., 2015; Koller and Wu, 2023). Therefore, to model effective questions analysts need to concoct tactics that are functionally equivalent to adversarialness but rooted in distinct mechanisms, since their role discourages them from showing any confrontational attitude; such tactics modulate what we call the *cornering* quality of questions. The idea is that "to corner" means to constrain the set of possible moves and to limit the ability to perform evasive maneuvers; thus, cornering questions constrain what counts as a valid answer and are difficult to evade.

The present contribution proposes a metric for the evaluation of the degree of such a cornering attitude in ECCs; we call it Question Cornering Score (QCS). A baseline QCS is calculated on the basis of manually annotated features, and later GPT-4 performance in the scoring of the same features is tested against the baseline. We conduct a case-study, comparing the variation of call-QCS in the annotated sample with financial data describing the evolution of the corresponding stock price, as well as forecasted and actual earnings per share (EPS), their delta and their surprise component for each financial quarter. Results show that the score has a correspondence in the likely stance that the questioner developed towards the issuing corporation due to event-external relevant factors. Encouraged from these results, we finally present some future goals that extend from the (automatic) assessment of the QCS towards the estimation of the answerhood degree of the corresponding replies.

Extracting the cornering degree of questions, therefore, represents a fundamental advancement towards a systematic study of Q&A pairs, captur-

Proceedings of the Joint Workshop of the 8th Financial Technology and Natural Language Processing (FinNLP) and the 1st Agent AI for Scenario Planning (AgentScen), Jeju, South Korea, August 3, 2024

ing answerhood and cooperation vs. evasion in corporate-side answers – which will arguably produce insights into corporate performance and market reaction.

1.1 Earnings Conference Calls

ECCs are quarterly public events with a formulaic and predefined structure, during which companies present financial results of the previous quarter, explain the current situation and share an outlook for the following quarters (Crawford Camiciottoli, 2010). They are voluntary events which are however held by the vast majority of listed companies, since they represent one of the few public interactive moments between corporate representatives and financial analysts, among the investor relations activities (Rocci and Raimondo, 2017). The significance of such an activity is widely recognized from both sides, and this is testified by the invariable attendance of high-level managers, typically even the CEO and/or the CFO.

ECCs comprise a Question-and-Answer (Q&A) session, during which analysts can seek to elicit as much contextual information around the disclosed results as it is possible to obtain without creating the obligation for the company to provide a supplementary disclosure of *material information* (Clark, 2021). To do so, questioners apply multiple rhetorical strategies to get the right understanding about whether the company is really worth investing in, and later communicate their valuation and recommendation in reports made available to investors (Palmieri et al., 2015). The tone, content and interaction of Q&A sessions have already been shown to have an impact on stock price (Chen et al., 2018), particularly due to analyst intervention.

Analysts' questions in ECCs, which make the object of this contribution, are therefore powerful carriers of information. Additionally, the measure here developed and its analysis would be applicable to comparable Q&A interaction schemes, both in the financial domain (e.g., interviews with the top management) and outside it (e.g., press conferences).

2 Theoretical background

2.1 NLP for finance

The use of Natural Language Processing (NLP) techniques in finance is mostly of the text-mining kind (Kumar and Ravi, 2016) and caters for the needs of both investors and traders, and of the firm's Investor Relations. NLP is employed to extract information about what is explicitly stated in documents, disclosures or exchanges, or to explore the implicit content that lies behind the statements – whether sentiment, opinions, or argumentation.

FinTech applications mostly deal with explicit content and are typically developed for (or from) the corporate side, to enhance the effectiveness of the communication with clients or investors (see Chen et al., 2020).

The mining of implicit content, on the other hand, is more commonly a domain that potentially helps clients or investors making informed decisions. The goal in this case is to acquire insights about the past performance and inferentially predict the future course of a company. Results drawn from such studies could arguably be exploited by companies as well, to check the soundness of their current approach to investors communication and possibly improve it. Techniques of this kind traditionally involve the assessment of the sentiment (see Kearney and Liu, 2014 for a review), but notably also include opinion mining and argument mining (Garcia Villalba and Saint-Dizier, 2012; Liu, 2012; Chen et al., 2021), which deal with inferential connections between (often material) premises and (often evaluative) conclusions.

NLP applied to ECCs has either a descriptive or a predictive approach. Description is aimed at the retrieval of certain trends or patterns (Davis et al., 2015; Rocci et al., 2019), possibly correlating them with financial data; not necessarily confined to *text* analysis (Chen et al., 2023). Prediction is forward-oriented, prognosticating for instance post-event analysts' recommendations on the basis of questions formulation and answers tone (Keith and Stent, 2019; Pazienza et al., 2020).

The current contribution is text-based and primarily descriptive, but as part of a planned pipeline including answerhood evaluation and argument mining it has the potential to feature in a range of NLP application for companies and investors, including those aimed at forecasting.

2.2 Adversarialness in political press conferences

Structurally similar to ECCs, PPCs are a field in which descriptive research on question design and questioning strategies flourished for years (Heritage, 2003; Clayman et al., 2006, 2012; Heritage and Clayman, 2013). One development from which this study draws inspiration is the study and measure of the *adversarialness* of journalists' questions in PPCs (Clayman and Heritage, 2002; Clayman et al., 2007). For ECCs, question reformulations and their relative adversarial strength has been investigated only qualitatively by de Oliveira and Pereira (2017).

The original proponents of the adversarialness measure were primarily data-driven in the decomposition of the concept into relevant features and the scores attributed to each of them. The way of computing the total, however, was convoluted and opaque to the reader.

The property and score that are proposed in the current contribution differ from the concept of adversarialness in the following regards:

a. The property is not a characteristic of a single question but of a wider textual unit called MIU, presented below in §2.3

b. The property does not describe how hostile a question is, but how much difficult (i.e., reputationally costly) it is to evade a proper answer to the question

c. The score computed to evaluate the property derives from the plain sum of the scores attributed to its constituent features

2.3 Text segmentation: Maximal Interrogative Units

In ECCs there is a conventional limit on the number of turns an analyst is granted before ceding the floor to the next questioner. Analysts typically have a number of issues they aim to resolve and a number of questions to ask, and so construct individual turns such that they introduce multiple questions which in another activity type, such as spontaneous informal conversation, could be spread out over a series of shorter turns (D'Agostino et al., 2024b).

These multi-issue question turns are segmented by speakers into topically homogeneous sequences of utterances, called Maximal Interrogative Units (MIUs): questioning units typically below the level of the turn, but above the level of the clause or individual speech act. ECC speakers at times make explicit reference to MIU segmentation; a case of this is illustrated by Example (1), further discussed below, in which the speaker, analyst Jeremy Sigee, explicitly and repeatedly marks two sections of their turn as forming a first and second 'question'.

(1) Morning. Thank you very much. Apologies for taking on the painful bits, but I still think there's more clarification that we need. **I wanted to just ask two things.**

{**One** is on Greensill. [You've got about CHF5 billion cash, but also about CHF5 billion remaining exposure in those funds.]$_{preface1}$ [And I just wondered if you could put a number on how much of that CHF5 billion remaining exposure is to doubtful borrowers, including, obviously, Gupta, but also some of the other doubtful borrowers who seem reluctant to pay.]$_{question\ 1}$ **So, that's my first question.**}$_{MIU\ 1}$

{**And my second question** is on the other painful, like I said, I'm afraid, on the Archegos situation. [Could you walk us through the mechanics of how that loss came about in terms of what the outstanding gross exposure was at the moment of problem?]$_{question\ 2}$ [How much margin you had and the sequence of events in terms of, were you slow to sell down or how do you assess what happened?]$_{question\ 3}$}$_{MIU\ 2}$

Those are my two questions please.

In Example (1), the speaker emphasises the fact that they are asking two 'questions' both at the beginning of the turn (before the first MIU, "I wanted to just ask two things") and at the end of the turn (after the second MIU, "Those are my two questions please"). The closing remark, moreover, also plays the role of concluding the turn, leaving the stage to the management for a reply.

Also the "So, that's my first question" remark that concludes the first MIU engages in a similar double purpose: it both reiterates the enumeration of 'questions' and declares the conclusion of the first questioning act.

Finally, both MIUs are introduced by a discourse marker ("One", "And my second question") that serves the purpose of counting the progression of 'questions'.

3 Question cornering score

The core contribution of the current work is of theory development, paired with an exploratory study on the application of such a theory to the context of ECCs and the automatic replication of the measure that the theory proposes. The theoretical construct presented here is the *cornering* property of an MIU and the score (QCS) that is assigned to the MIU on the basis of six discrete, topic independent features that are selected as realistic means of estimating such a property.

An MIU is evaluated to be cornering the more it raises the cost to which the respondent is exposed for not answering properly and fully to it. This means that the higher the cornering score of an MIU performed by an analyst is, the heavier the burden of compliance with cooperation that is cast upon the management is. The cost associated with uncooperativeness can typically be a decreased perception of accountability, reliability, and ultimately value associated with the management and, subsequently, the company overall.

Cooperation is here to be intended in terms of a high degree of answerhood: how much the reply approximates the *principal possible answer*, i.e., can be regarded as logically sufficient and immediate with respect to the prompting question (Wiśniewski, 2015).

The QCS sums up the scoring of six independent features that describe the relevant structural properties of MIUs with respect to their ability to shape such a constriction:

$$QCS\Big|_{MIU} = \sum_{i \in features} (score)_i$$

where $i = 1, ..., 6$ are the features under consideration.

The following cut-off criterion was adopted for the scoring: for each feature, the MIU gets assigned the score of the highest-ranking type that is contains, independently from how many tokens it contains.

Following, an overview of the six features and their individual scores.

(1) Framing preface $\rightarrow$ **score {0, 2}**

This feature tracks the presence of a prefatory statement (Lucchini et al., 2022). If not present, the score is 0. In case there is at least one preface introducing the question(s): score 1 if the statement is neutral or positive; score 2 if it is negative.

(2) Complexity $\rightarrow$ **score {0, 2}**

Complexity counts the number of questions in the MIU. It assigns 0 if one question is present; 1 if questions are 2 or 3; 2 if there are 4 questions or more.

(3) Directness $\rightarrow$ **score {0, 1}**

This feature recognizes whether the MIU contains elements of indirectness (score 0) or whether questions are formulated in a direct way (score 1). Indirectness is both related to hedging and the modality of a question, i.e., formulations such as "*I would be glad to hear* something", but also "*Could you say* something" (see Crawford Camiciottoli, 2009) would be both assigned score 0.

(4) Assertiveness $\rightarrow$ **score {0, 1}**

The assertiveness of questions relates to their formulation. Assuming that each question can be formulated as open or closed, 0 is given if all questions in the MIU are open; 1 if at least one question of the MIU is closed.

(5) Request type $\rightarrow$ **score {0, 2}**

This feature depends on a two-step annotation of each question. First, each question is attributed a certain request type – according to a speech-act typology described and operationalized in Lucchini and D'Agostino (2023). Based on that, the score is assigned to the highest-ranking request in the MIU, according to the following scheme:

- score 0 to requests for elaboration or data
- score 1 to requests for opinion, explanation, clarification or of confirmation of some material data
- score 2 to requests for justification, commitment, or the confirmation of an inference

(6) Time orientation $\rightarrow$ **score {0, 2}**

The time orientation score is 0 if the topic of the question(s) is placed in the present; 1 if in the future; 2 if in the past.

The final cornering score is the sum of the individual scores assigned to an MIU and lies in the range {0, 10}. An MIU is considered to be (increasingly) 'cornering' if its QCS is equal to or higher than 5. Concrete examples showing the application of this scoring are presented in Appendix A.

Beside plain QCS, attributed to MIUs, we name call-QCS the sum of all QCS values of a call. Call-QCS is defined as follows over the $j = 1, ..., n$ MIUs of a call:

$$_{call}QCS = \sum_{j \in MIUs} (QCS)_j$$

Call-QCS is not weighted with the length of the call; this means that it is not calculated considering the number of MIUs per call as a biasing factor. On the contrary, the number of MIUs is acknowledged to be an underlying additional factor that determines the cornering nature of the call.

4 Research questions

Two research questions are addressed in this study:

RQ1. Does the QCS reflect an inquisitive attitude of the speaker, motivated by noteworthy external factors?

RQ2. Is the QCS a measure that can reliably be reproduced by AI tools?

RQ1 will be answered by comparing call-QCS with financial data such as the estimated performance of the company and its actual results over time. RQ2 will be answered by measuring the agreement rate between manual and automatic scoring.

The hypotheses against which the results will be tested are:

H1. The QCS correlates with financial results and news that have a clear impact on such results; particularly, the call-QCS is expected to be higher, the more challenging and potentially disrupting the situation is for the company (and vice-versa).

H2. The measure can be assessed reliably by LLMs insofar as it is decomposed into constitutive features. Some features are harder to score than others.

5 Data and method

The dataset for the current study are the four ECCs held in 2021 by the Swiss bank *Credit Suisse* (CS), for a total of 111 MIUs (483 sentences; 9,853 words; language: English). CS was chosen as a case study because of the poor performance and the sequence of critical issues that the company faced in 2021. The most remarkable features taken into consideration are:

- CS steadily reported losses along the whole financial year
- although a certain variability in CS stock prices can be traced, their value drops around each ECC, with an overall decline over the fiscal year
- CS incurred in at least two major scandals during the period considered (financial and reputational crises); namely the Archegos-Greensill double bankruptcy and the “tuna bonds” fraudulent issuance

In a precarious environment, financial analysts are expected to ask questions that are straight to the point. As a consequence, a study about the cornering degree of questions to CS representatives in 2021 seemed an ideal environment to start testing the soundness of the score. Following, a sketch of the methodological approach.

Setup The first step is the segmentation of question turns into MIUs. This is currently performed manually to ensure precision. The measured inter-annotator reliability for this task (Krippendorff’s alpha (1995) for the unitizing of textual continua among three annotators) is ${}_U\alpha = 0.933$ (see D’Agostino et al., 2024a for further details).

Manual assessment To collect manually annotated data, four trained annotators [1] are instructed to manually score all the 111 pre-segmented MIUs for each feature. Krippendorff’s alpha coefficient for nominal data is the measure employed to evaluate their annotation agreement. Over the single features, the agreement rate ranges from $\alpha = 0.38$ (Request type) to $\alpha = 0.84$ (Assertiveness), as shown in Table 1. Better scores are consistently measured considering only annotators A and B; the remaining two (C and D in Table 1) introduce excessive outliers. Therefore, only the scores by annotators A and B will be used hereafter. The agreement rate among the two best annotators for the QCS (the sum of the single scores) is $\alpha = 0.57$.

Baseline values are calculated as the statistical mode of the manual assessment by the two best annotators. They are determined both at feature (*baseline feature score*) and at the QCS level (*baseline QCS*).

RQ1 The call-QCS is calculated as the sum of the baseline QCS values of a call.

[1] Annotators are student assistants, employed with a part-time contract by the project that funds the current contribution. They are second-year Master’s students in investor relations with a background in languages/linguistics. Their tasks include, but are not limited to, data annotation and the assessment of the current score. For any task, their training is carried jointly by the two PhD students who work on the project. The annotation guidelines for this task that were provided to the annotators are those presented in §3.

Feature	α (4 annotators)	α (A + B)	α (A + B + C)	α (A + B + D)	α (C + D)
Framing preface	0.39	0.58	0.43	0.40	0.46
Complexity	0.75	0.99	0.86	0.75	0.64
Directness	0.49	0.93	0.68	0.47	0.38
Assertiveness	0.84	1.00	0.88	0.84	0.90
Request type	0.38	0.89	0.46	0.53	0.29
Time orientation	0.61	0.89	0.86	0.55	0.37
QCS	0.27	0.57	0.30	0.36	0.19

Table 1: Inter-rater reliability for the manual annotation of the QCS and its constitutive features, measured as Krippendorff's alpha (nominal)

The financial data and relevant news are retrieved from the Bloomberg terminal. These are qualitatively compared with call-QCS measures.

RQ2 The MIUs are passed to GPT-4 via API. The model is instructed with zero-shot prompting to assign a score to each feature of every MIU. The best performing prompt is found through four cycles of instruction-tuning, and holding the best-performing prompt (evaluated in terms of F1 scoring with respect to the baseline) for each feature independently. The set of final prompts can be found in Appendix B. The LLM was tested beforehand and determined to be capable of discerning each value under observation without further instructions or the need for context-specific examples; therefore few-shot prompting was not considered necessary for this exploratory study.

The automatic scoring of both the features and the resulting QCS is tested against the corresponding manual baseline.

6 Results

RQ1 The manual assessment of the call-QCS across the four financial quarters of 2021 is reported in Table 2.

Quarter	call-QCS
Q1	174
Q2	149
Q3	50
Q4	116

Table 2: Manual assessment of call-QCS across the four financial quarters of 2021

Stock prices (closing price every day at 4 p.m.) and main events for the year 2021 are summarized in the line chart of Figure 1, where the values of Table 2 are also displayed as a bar chart. Table 3 reports the analysts' earnings per share (EPS) estimate consensus, the actual EPS results at the end of the period considered, and the resulting surprise; for each financial quarter and annually.

Time period	EPS estimate	EPS result	EPS surprise (%)
Q1	0.86	-0.07	n.a.
Q2	0.37	0.18	-28.8
Q3	0.11	0.15	56.1
Q4	0.03	-0.76	8.8
year	-0.64	-0.61	4.75

Table 3: Earnings per share (estimate consensus, result, percentage surprise) across the four financial quarters of 2021 and for the full year

RQ2 GPT-4 API was called for each QCS feature independently. The feature results were first tested against the feature-baseline and then summed up to form the QCS; the latter was tested against the baseline QCS. GPT-4 predictive performance is measured in terms of F-score; particularly, balanced accuracy F1. Results are presented in Table 4.

Feature	F1
Framing preface	0.59
Complexity	0.62
Directness	0.37
Assertiveness	0.80
Request type	0.53
Time orientation	0.64
QCS	0.20

Table 4: F1 assessment over the entire dataset, testing GPT-4 (best zero-shot prompt) vs. manual QCS-baseline

Manually and automatically assessed QCS was

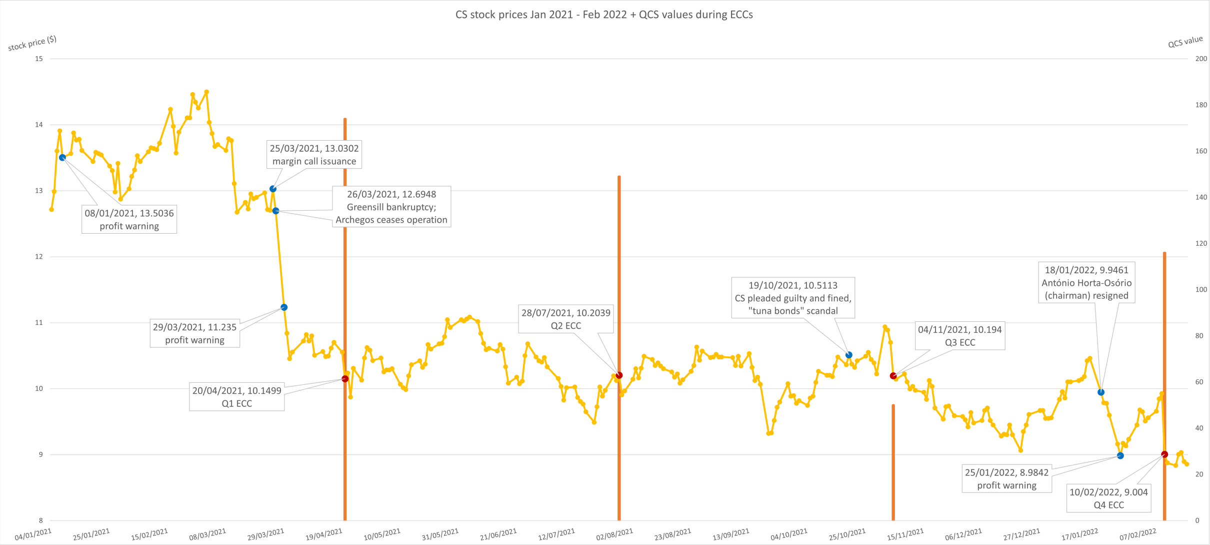

Figure 1: *Line chart*: CS stock prices chart and call-out tags on main events (date, price, event type) occurred to the company between January 2021 and February 2022. Red dots: ECC day; blue dots: disclosure of a negative episode. *Bar chart*: call-QCS values of the four ECCs (as reported in Table 2).

again compared for accuracy, resulting in F1 = 0.20 – as shown in the last line of Table 4.

7 Discussion

RQ1 The call-QCS value for Q1 is equal to 174; the highest of the year. Stock prices, on the other hand, were the lowest of the semester on the day of the call and the ones immediately preceding. The context is presumably critical due to the Greensill Capital and Archegos Capital Management scandals, both of which took place about a month before the issue of the Q1 report. Credit Suisse's misconduct with respect to such events was ascertained by several inquires and the company was pleaded guilty of breaches on risk management and governance; this hugely affected its reliability. Moreover, despite the judiciary issues and two profit warnings since the beginning of the year, the analyst consensus towards Q1 earnings was decidedly positive; results, however, were negative and, most importantly, the difference between estimate and result was almost a point apart ($\Delta = 0.93$): a strong negative surprise. This means that analysts were optimistic and their trust was not rewarded; an optimal ground for a highly cornering tone, aimed at understanding what went wrong.

During Q2, stock prices settled only a little higher than the slump that preceded Q1 announcements: the market was starting to realise that the crisis was not temporary as one might have originally expected. Investigations, especially for the Archegos case, were still ongoing and many points were still unclear. The difference between expected and actual earnings is again negative, but results are positive and the delta is lower ($\Delta = 0.19$). This semblance of recovery may have been the reason for lowering the Q2 call-QCS slightly, but not a reason strong enough to let it drop. In fact, it is the second highest score of the year.

Q3 was marred by the "tuna bonds" scandal which, however, did not seemingly have a huge impact on either the reliability or the profitability of the company: stock prices do not appear affected in the period following the accusations, and earnings were even greater than expected. This correlates with a rather low cornering score for the call; the lowest of the year.

Quarterly results for Q4 are decidedly negative, as it is the difference between estimation and results; their delta is the second highest of the year ($\Delta = 0.79$). With respect to stock market data, on ECC day the stock price reached a historical low among the ECC days of 2021 ($ 9.004); besides, it constitutes the second lowest price of the year up to that point. Such a critical environment would seem to call for a high degree of cornering in the questions of analysts; the call-QCS for Q4, however, is 116, lower than in Q2.

Two additional factors need to be acknowledged:

- As shown in Table 3, the release of quarterly results for Q4 was paired with year-on-year earnings data. Although the quarter was negative, annual results were better than estimated and lead to a earnings-per-share surprise equal to 4.75%; the first positive surprise for this metric in years. This may have softened the stance of some analysts.

- The ECC event was closely followed by an absolute low in value for the company, that reached a cost per share equal to $ 8.836; Credit Suisse's stock price keep decreasing until the acquisition by UBS at the beginning of 2023. Insider knowledge and intuition both may have contributed to a general slacking in the questioning tone of some analysts: if the belief is that the company is not worth investing in anymore, there is no use in asking cornering questions.

Ultimately, the score appears to follow the financial trends, thus confirming hypothesis H1.

RQ2 Among the six parameters that were submitted for classification to GPT, four obtained a satisfactory F1 result (Assertiveness, Time orientation, Complexity, and Framing preface), one a borderline result (Request type), and one did not reach sufficiency (Directness). This confirms part of the hypothesis, although Directness was not the feature that was expected to perform the lowest.

Rather striking is however the assessment of the predictive performance of the overall QCS (i.e., the sum of the individual feature scores), which resulted in F1 = 0.20 – an underwhelming result that evaluates the performance as insufficient. The interpretation of such a measure acknowledges the cumulative nature of F1 scores with respect to each task: whereas feature prediction generally appears to be good enough *on average*, summing the single predictions to evaluate the complex score for each example reveals that they are most often wrong.

In conclusion, the performance of GPT in the classification of constitutive parameters of QCS cannot be assessed as reliable and, subsequently, it cannot be deemed as a valid alternative to the manual assessment of the cornering score. This is in contrast with hypothesis H2.

8 Conclusions and future work

The present study introduces the notions of cornering attitude of a questioner and the Question Cornering Score (QCS) that measures it. It argues for the significance of the QCS in assessing the tone of questions performed by financial analysts over the course of Q&A sessions of Earnings Conference Calls, it shows that the score correlates with the company's financial performance, and it evaluates the reliability of a GPT model in predicting such a score while decomposed into independent constitutive features.

With respect to the purposefully sampled dataset under observation, the QCS appears to be a descriptive measure of the market stance towards a company over time. Besides the extension of the corpus and verification of such results, following steps will include the assessment of whether the QCS can also work as a predictive indicator.

Given the descriptive power of the property and its related score, the automatic measuring of the QCS on text segments constituting a macro questioning act (MIUs) is a critical goal. The LLM GPT-4 is employed to evaluate MIUs with respect to six independent features. Results appear to be generally good by feature, but unsatisfying on the complete score; consequently, the model is not adequate for the assessment of such a score with the proposed methodology. Further research will investigate new ways of using GPT in the assessment of the cornering quality of questions; however, aware of the fact that GPT is not the adequate tool for mimicking sophisticated logical activities such as inference, we might argue that the subtle clues that suggest the cornering tone of a question also fall into this category. More specialized AI engineering will therefore be required to obtain satisfactory results, as for the argument mining domain.

In the perspective of future work featuring the QCS, here are some goals for our research:

- To identify a reliable way to assess the QCS automatically.
- To correct the score by the possible influence of personal style of the questioner and/or casual noise independent from the ECC event.
- To verify whether the score has a predictive value, other than descriptive.
- To identify patterns within the answers provided to cornering MIUs.
- To measure the degree by which an answer to a cornering MIU is cooperative, i.e., provides the desired type of response.

References

Chung Chi Chen, Hen Hsen Huang, and Hsin Hsi Chen. 2020. NLP in FinTech Applications: Past, Present and Future. *arXiv*. ArXiv: 2005.01320.

Chung-Chi Chen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2021. *From Opinion Mining to Financial Argument Mining*. SpringerBriefs in Computer Science. Springer Singapore, Singapore.

Jason V. Chen, Venky Nagar, and Jordan Schoenfeld. 2018. Manager-analyst conversations in earnings conference calls. *Review of Accounting Studies*, 23(4):1315–1354.

Yuan Chen, Dongmei Han, and Xiaofeng Zhou. 2023. Mining the emotional information in the audio of earnings conference calls : A deep learning approach for sentiment analysis of securities analysts' follow-up behavior. *International Review of Financial Analysis*, 88:102704.

Cynthia E. Clark. 2021. How do standard setters define materiality and why does it matter? *Business Ethics, the Environment & Responsibility*, 30(3):378–391.

Steven E. Clayman, Marc N. Elliott, John Heritage, and Megan K. Beckett. 2012. The President's Questioners: Consequential Attributes of the White House Press Corps. *The International Journal of Press/Politics*, 17(1):100–121.

Steven E. Clayman, Marc N. Elliott, John Heritage, and Laurie L. McDonald. 2006. Historical Trends in Questioning Presidents, 1953-2000. *Presidential Studies Quarterly*, 36(4):561–583.

Steven E. Clayman and John Heritage. 2002. Questioning presidents: Journalistic deference and adversarialness in the press conferences of U.S. Presidents Eisenhower and Reagan. *Journal of Communication*, 52(4):749–775. ArXiv: 1011.1669v3 ISBN: 1460-2466.

Steven E. Clayman, John Heritage, Marc N. Elliott, and Laurie L. McDonald. 2007. When Does the Watchdog Bark? Conditions of Aggressive Questioning in Presidential News Conferences. *American Sociological Review*, 72(2005):23–41. ISBN: 0003-1224.

Belinda Crawford Camiciottoli. 2009. "Just wondering if you could comment on that": Indirect requests for information in corporate earnings calls. *Text and Talk*, 29(6):661–681.

Belinda Crawford Camiciottoli. 2010. Earnings calls: Exploring an emerging financial reporting genre. *Discourse & Communication*, 4(4):343–359.

Giulia D'Agostino, Chris Reed, and Daniele Puccinelli. 2024a. Segmentation of Complex Question Turns for Argument Mining: A Corpus-based Study in the Financial Domain. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14524–14530, Torino, Italia. ELRA and ICCL.

Giulia D'Agostino, Ella Schad, Eimar Maguire, Costanza Lucchini, Andrea Rocci, and Chris Reed. 2024b. Superquestions and some ways to answer them. *Journal of Argumentation in Context*. In press.

Angela K. Davis, Weili Ge, Dawn Matsumoto, and Jenny Li Zhang. 2015. The effect of manager-specific optimism on the tone of earnings conference calls. *Review of Accounting Studies*, 20(2):639–673.

Maria do Carmo Leite de Oliveira and Silvia Maura Rodrigues Pereira. 2017. Formulations in Delicate Actions: A Study of Analyst Questions in Earnings Conference Calls. *International Journal of Business Communication*, 55(3):293–309.

Maria Paz Garcia Villalba and Patrick Saint-Dizier. 2012. Some facets of argument mining for opinion analysis. *Frontiers in Artificial Intelligence and Applications*, 245(1):23–34. ISBN: 9781614991106.

John Heritage. 2003. Designing Questions and Setting Agendas in the News Interview. In Phillip Glenn, Curtis D. LeBaron, and Jenny Mandelbaum, editors, *Studies in Language and Social Interaction: In Honor of Robert Hopper*, pages 57–90. Lawrence Erlbaum, Mahwah, NJ.

John Heritage and Steven E. Clayman. 2013. The changing tenor of questioning over time tracking a question form across us presidential news conferences, 1953-2000. *Journalism Practice*, 7(4):481–501.

Colm Kearney and Sha Liu. 2014. Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, 33:171–185.

Katherine A. 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. ArXiv: 1906.02868.

Veronika Koller and Xiaoxi Wu. 2023. Analysts' identity negotiations and politeness behaviour in earnings calls of US firms with extreme earnings changes. *Corporate Communications: An International Journal*, 28(5):769–787.

Klaus Krippendorff. 1995. On the Reliability of Unitizing Continuous Data. *Sociological Methodology*, 25:47.

B. Shravan Kumar and Vadlamani Ravi. 2016. A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, 114:128–147.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool. ISSN: 1947-4040.

Costanza Lucchini and Giulia D'Agostino. 2023. Good answers, better questions. Building an annotation scheme for financial dialogues. Technical report. Ark:/12658/srd1326777.

Costanza Lucchini, Andrea Rocci, and Giulia D'Agostino. 2022. Annotating argumentation within questions. Prefaced questions as genre specific argumentative pattern in earnings conference calls. In *Proceedings of the 22nd Edition of the Workshop on Computational Models of Natural Argument (CMNA 22)*, volume vol. 3205, pages 61–66, Cardiff. CEUR.

Rudi Palmieri, Andrea Rocci, and Nadzeya Kudrautsava. 2015. Argumentation in earnings conference calls. Corporate standpoints and analysts' challenges. *Studies in communication sciences*, 15, 2015(1):120–132.

Andrea Pazienza, Davide Grossi, Floriana Grasso, Rudi Palmieri, Michele Zito, and Stefano Ferilli. 2020. An abstract argumentation approach for the prediction of analysts' recommendations following earnings conference calls. *Intelligenza Artificiale*, 13(2):173–188.

Andrea Rocci and Carlo Raimondo. 2017. Conference calls: A communication perspective. In Alexander V. Laskin, editor, *The Handbook of Financial Communication and Investor Relations*, pages 293–308. John Wiley & Sons, New York, NY. Https://doi.org/10.1002/9781119240822.ch26.

Andrea Rocci, Carlo Raimondo, and Daniele Puccinelli. 2019. Evidentiality and Disagreement in Earnings Conference Calls : Preliminary Empirical Findings. In *Proceedings of the 3rd Workshop on Advances In Argumentation In Artificial Intelligence co-located with the 18th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2019)*, pages 1–5.

Andrzej Wiśniewski. 2015. Semantics of Questions. In Shalom Lappin and Chris Fox, editors, *The Handbook of Contemporary Semantic Theory*, 1 edition, pages 271–313. Wiley.

A Examples of QCS attribution

Showcasing both the single feature-level scores and the overall cornering score of the MIU.

(2) Firstly is just on the strategy again. Obviously, you had a very detailed presentation in December last year and probably we were talking about the 10% to 12% and the investing for growth. Should we assume by this fact the strategy stands at 10% to 12% RoTE or should we expect, as you indicated, post the Investment Banking review and the new chairman arriving that we will get a new strategic update?

Example (2) (analyst Anke Reingen, CS Q1 2021): Framing preface: 1 (preface present; the preface has a neutral-to-positive tone); Complexity: 0 (one question); Directness: 1 (no hedging or modalization); Assertiveness: 1 (closed formulation of the question); Request type: 2 (confirmation of an inference); Time orientation: 1 (future-oriented); **QCS: 6** → *the MIU is cornering*

(3) Sorry, on the prime brokerage business, you indicated two-thirds of the, I think, balances down, but you also gave an indication of 600 million of revenues, 400 million of cost to be reduced in 2022. I just wonder – I assume some of that is already in the numbers or should we think about the numbers getting bigger?

Example (3) (analyst Kian Abouhossein, CS Q4 2021): Framing preface: 2; Complexity: 0; Directness: 0; Assertiveness: 1; Request type: 1; Time orientation: 1; **QCS: 5** → *the MIU is slightly cornering*

(4) The first one, just trying to get a sense, I appreciate you don't prejudge the outcome, but – so the strategic review, just if we can get a bit more color in terms how the process works, how that's being conducted, how decisions will be made and the kind of trade-offs and the processes involved in that, that would be helpful.

Example (4) (analyst Amit Goel, CS Q2 2021): Framing preface: 1; Complexity: 0; Assertiveness: 0; Request type: 0; Time orientation: 0; **QCS: 1** → *the MIU is not cornering*

B GPT prompts

Framing preface messages=[
{"role": "system", "content": "You are a helpful assistant designed to output JSON."},
{"role": "system", "content": "You take as an input some text drawn from a question turn of an earnings conference call and output a quantitative assessment called 'framing preface', relative to the presence and the nature of background statements that precede or follow questions in the input. A background statement is an assertive sentence used to give some background, context or justification to the question proper. The quantitative assessment must come in the form of a single integer number in the range {0, 2}, where 0 means that there is no background statement, 1 means that there is at least one background statement that presents a positive or neutral situation, 2 means that there is at least one background statement that presents a negative situation."},

{"role": "system", "content": "The scores must be considered as in ascending order of importance: if there were two background statements, of which one positive and one negative, only the negative one will be considered."},
{"role": "user", "content": *text*}]

Complexity messages=[
{"role": "system", "content": "You are a helpful assistant designed to output JSON."},
{"role": "system", "content": "You take as an input some text drawn from a question turn of an earnings conference call and output a score called 'complexity', which refers to the number of interrogative sentences present in the input. The score 'complexity' must come in the form of a single integer number in the range {0, 2}, where 0 means that the input displays one interrogative sentence, 1 means that the input displays two to three interrogative sentences, 2 means that the input displays four or more interrogative sentences."},
{"role": "system", "content": "The answer must be based on the total number of interrogative sentences, including multiple instances of the same one."},
{"role": "user", "content": *text*}]

Directness messages=[
{"role": "system", "content": "You are a helpful assistant designed to output JSON."},
{"role": "system", "content": "You examine the text of questions asked by financial analysts in earnings conference calls. Sometimes analysts ask questions in an indirect, tentative and polite manner, sometimes they are blunt and to the point."},
{"role": "system", "content": "You take as an input some text drawn from a question turn of an earnings conference call and output a score called 'directness'. The score 'directness' must come in the form of a single integer number in the range {0, 1}, where 0 means that the input contains at least one indicator of politeness, tentativeness, indirectness or hedging, 1 means that the input does not contain any indicator of politeness, tentativeness, indirectness or hedging."},
{"role": "user", "content": *text*}]

N.B. the second "system" message can be omitted and the results, i.e., the F1 measure of the performance over the 111 MIUs, is not affected up to the 15th decimal position.

Assertiveness messages=[
{"role": "system", "content": "You are a helpful assistant designed to output JSON."},
{"role": "system", "content": "You take as an input some text drawn from a question turn of an earnings conference call and output a quantitative assessment called 'assertiveness', which refers to the formulation of questions in the input. The quantitative assessment 'assertiveness' must come in the form of a single integer number in the range {0, 1}, which rates only once the element with the highest score in the input."},
{"role": "system", "content": "Score 0 means that the input contains only open questions. Open questions are wh- questions (beginning with what, why, how) or questions asking to describe, elaborate or explain something in an open ended way. Score 1 means that the input contains polar questions or choice questions. Polar questions are yes/no questions or questions asking whether someone can confirm or agrees with a statement, a comment, a forecast, an explanation or a piece of reasoning. Choice questions posit a closed list of alternatives asking to choose between them, it could be alternative descriptions, evaluations, outlooks, explanations or reasons"},
{"role": "user", "content": *text*}]

Request type messages=[
{"role": "system", "content": "You are a helpful assistant designed to output JSON."},
{"role": "system", "content": "You take as an input some text drawn from a question turn of an earnings conference call and output a quantitative assessment called 'request type', which refers to the kind of answer that is sought by the questions in the input. The quantitative assessment 'request type' must come in the form of a single integer number in the range {0, 2}, which rates only once the element with the highest score in the input."},
{"role": "system", "content": "For the quantitative assessment 'request type', follow these criteria: questions that challenge the respondent to provide a justification grant the score 2; questions that seek a commitment to action from the respondent grant the score 2; questions that ask the respondent to confirm or disconfirm a hypothesis, inference, guess or calculation grant the score 2; questions that ask for an evaluative or predictive opinion or some kind of assessment grant the score 1; questions that seek an explanation, query about the causes of an event or the motives of an action grant the score 1; questions that request a clarification of what has been said grant the score 1;

questions that ask to confirm material data grant the score 1; questions that ask to elaborate on a topic or ask for details grant the score 0; questions that merely ask for data grant the score 0."},
{"role": "user", "content": *text*}]

Time orientation messages=[
{"role": "system", "content": "You are a helpful assistant designed to output JSON."},
{"role": "system", "content": "You take as an input some text drawn from a question turn of an earnings conference call and output a quantitative assessment called 'time orientation'. The quantitative assessment 'time orientation' must come in the form of a single integer number in the range {0, 2}, which rates only once the element with the highest score in the input."},
{"role": "system", "content": "Score 0 is attributed when the entire input asks questions about the present; score 1 is attributed when the input contains questions about the future; score 2 is attributed when the input contains questions about the past."},
{"role": "user", "content": *text*}]