Replication Report: Detecting Non-Answers in Earnings Conference Call Q&A

Shutong Dong

Minghao Yan

February 01, 2026

Abstract

This report documents a partial replication of de Kok (2024), focusing on the construction of the earnings-call Q&A sample and the evaluation and incidence of alternative non-answer classifiers corresponding to Tables 1 and 2 in the paper. Extracting sample firms from the Capital IQ database, we use (i) the rule-based Gow et al. (2021) non-answer measure, (ii) two zero-shot generative large language model (GLLM) classifiers (Spark Pro and Spark Max), and (iii) a hybrid keyword pre-filter followed by Spark Max. I evaluate classification performance against 100 manually labeled Q&A pairs (19% non-answers) and report Type I and Type II error rates, precision, recall, and F1 scores. Relative to the Gow et al. baseline, the GLLM approaches increase recall and improve non-answer F1, while the keyword–LLM hybrid substantially reduces false positives. Performance remains below the paper's fine-tuned ChatGPT pipeline, consistent with differences in data source, sample size, and model selection.

1. Introduction

De Kok (2024) provides a practical approach for using GLLMs in accounting academic research and demonstrates his approach through a case study on identifying "non-answers" in earnings conference call Q&A pairs. The detection of non-answers is considered to be challenging because the detection is highly contextual and often hinges on subtle statements of inability or unwillingness to answer (Hollander et al., 2010; de Kok, 2024). The paper uses a widely recognized rule-based detection approach as baseline (Gow et al., 2021) and shows that a carefully engineered, multi-step ChatGPT pipeline can materially reduce both false positives and false negatives, achieving 96% accuracy and a non-answer F1 score of 0.87 on a 500-pair evaluation set (de Kok, 2024, Table 1). In addition, the cost of detection can be largely reduced through the combination of a keyword filter and ChatGPT. More broadly, the paper complements prior methodological guidance in accounting textual analysis (Anand et al., 2020; Bochkay et al., 2022).

This replication project has two objectives. First, we provide our own sample construction procedure that mirrors the sample construction process of the original paper but starts from a different data source (CapitalIQ) and applies slightly different screening rules. Second, I reproduce the paper's Table 1/2-style reporting in a small-scale setting by comparing a rule-based baseline (Gow et al., 2021) to alternative GLLM-based classifiers (Spark Pro/Max) and a keyword filter-plus-GLLM hybrid. Due to our very limited budget and time constraint, there are only 100 manually-labelled Q&A pairs of data as a benchmark for model evaluation.

2. Data and Methodology

2.1 Data

The firms in our sample are drawn from the CapitalIQ database for calendar years 2013–2022. Starting from 12,614 unique firms (CIKs), I remove firms with missing industry information and firms in the financial and public utility industries (GICS classification code: 40 & 55). I further exclude firms without any earnings conference call Q&A transcripts and firms whose transcripts all contain fewer than five Q&A responses. Finally, to mitigate the meaningless Q&A pairs, I exclude firms for which none of their Q&A pairs satisfy minimum-length criteria: the question contains at least 30 characters, the answer at least 10 characters, and the combined length at least 75 characters. The resulting sample contains 5,471 firms and 166,848 Q&A pairs (Table 1). Finally, 100 Q&A pairs are randomly drawn and manually labelled for model evaluation. Moreover, 1000 Q&A pairs are randomly drawn for large sample analysis.

The construction process of ours differs from de Kok (2024)'s in several ways. De Kok (2024) retrieves transcripts via Finnhub, restricts to CIKs in the Gow et al. (2021) universe, and requires that each call contain a presentation section and at least three Q&A pairs, yielding 63,959 calls and 1,152,505 Q&A pairs (de Kok, 2024, Section 4.1). It is possible that these deviations can contribute to differences in the main results.

2.2 Methodology

As mentioned in previous sections, we manually evaluate whether the manager's responses in 100 Q&As are non-answers. We defined non-answers as those where the response includes a statement, explanation, or justification indicating an inability or unwillingness to answer the question. If a response meets this definition, it is coded as 1; otherwise, it is coded as 0.

Furthermore, four computer-based prediction approaches are implemented. The first approach is Gow et al. (2021) rule-based approach**.** A response is flagged as a non-answer if it matches one of the Gow et al. regular-expression non-answer categories (e.g., REFUSE, UNABLE, AFTERCALL). The second approach is the GLLM approach. Each Q&A pair is classified using the Spark Pro model with a prompt adapted from de Kok's Online Appendix. The prompt requests a structured JSON response containing both a written assessment and a binary no-answer classification. The third approach is the same as the second approach, but the model implemented in the third approach is more advanced. Spark Max model is used in the third approach. The fourth approach is a keyword filter-plus-GLLM hybrid approach. A keyword-matching filter is first applied to identify candidate non-answers. If no keyword match is found, the observation is classified as an answer (0) without a GLLM call. If a match is found, Spark Max is called using the structured prompt. This mirrors the paper's emphasis on pipeline design (de Kok, 2024).

3. Evaluation Metrics

We evaluate the model performance based on the manually labelled Q&A pairs. For each computer-based approach, I compute the confusion matrix elements—true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN)—treating "non-answer" as the positive class. I report: Accuracy = (TP + TN) / N; Type I error = FP / (FP + TN); Type II error = FN / (FN + TP); Precision = TP / (TP + FP); Recall = TP / (TP + FN); and F1 = 2TP / (2TP + FP + FN). This matches the metrics reported in de Kok (2024, Table 1).

4. Replication Results

Table 2 reports performance on the 100 manually coded Q&A pairs (81 answers and 19 non-answers). The Gow et al. baseline correctly identifies 5 of the 19 non-answers (recall 0.26) and produces 4 false positives (Type I error 0.05), resulting in a non-answer F1 score of 0.36. Both LLM approaches improve recall materially: Spark Pro reaches a recall of 0.47 (F1 0.51), and Spark Max reaches a recall of 0.53 (F1 0.59). The hybrid Keyword + Spark Max approach yields the highest non-answer precision (0.82) and the lowest Type I error (0.02), with a non-answer F1 of 0.60. These patterns are consistent with the idea that an initial keyword filter can reduce over-classification by a zero-shot LLM, trading off some recall for substantially higher precision.

Table 3 uses all 1,000 Q&A pairs to estimate the incidence of non-answers under each approach. The implied non-answer rate ranges from 8.2% for the Gow baseline to 13.3% for Spark Pro (Spark Max: 11.9%; Keyword + Spark Max: 8.7%).

5. Comparison with de Kok (2025)

Several qualitative similarities emerge. First, in both the original paper and my replication, the rule-based Gow et al. method exhibits a high Type II error relative to the best-performing GLLM-based approach. de Kok (2024) reports a baseline accuracy of 0.86 and non-answer F1 0.49 (Type II error 0.57) on a 500-pair evaluation set (Table 1). My replication finds a similar precision (0.56) but substantially lower recall (0.26), producing a lower F1 of 0.36. Second, the general direction of improvements from incorporating a GLLM is consistent: Spark Max raises recall to 0.53 and F1 to 0.59, and the hybrid keyword–GLLM design delivers very low false-positive rates, echoing the paper's emphasis on pipeline design rather than relying on a single zero-shot prompt.

The magnitude of performance differs sharply from the original study's best approach. de Kok's final fine-tuned ChatGPT method achieves an accuracy of 0.96 and non-answer precision/recall of 0.87/0.87 (F1 0.87) (de Kok, 2024, Table 1). The best-performing approach in my replication (Keyword + Spark Max) reaches an accuracy of 0.88 and a non-answer F1 of 0.60. Several factors likely explain the gap: (i) I use different GLLMs (Spark Pro/Max) rather than ChatGPT/ChatGPT-4 and do not implement a fine-tuning step; (ii) the evaluation set is smaller (100 vs. 500), increasing sampling variability; and (iii) differences in data sources can change the results and performance.

Large-sample incidence estimates also differ. de Kok (2024) finds that the ChatGPT method flags 13.9% of Q&A pairs as non-answers and the Gow method 12.3% in the full sample of 1.15 million pairs (de Kok, 2024, Table 2 and Section 4.1). In my 1,000-pair sample, the analogous incidence is lower for the Gow baseline (8.2%) and ranges up to 13.3% for Spark Pro. These differences are not necessarily contradictory: de Kok (2024) emphasizes that method disagreement can be large even when aggregate rates look similar, and small samples can shift the implied base rate materially.

6. Conclusion

This replication reproduces the paper's main empirical results at a smaller scale: identifying non-answers is a context-dependent classification task where rule-based methods can miss many true non-answers, and where GLLM-based approaches can improve recall and overall F1—especially when embedded in a sensible pipeline with keyword pre-filters. At the same time, the replication underscores that performance is sensitive to data source, screening rules, and model choice, and that the highest performance in de Kok (2024) is tied to both careful prompt design and fine-tuning.

References

Anand, V., Bochkay, K., Chychyla, R., & Leone, A. (2020). Using Python for text analysis in accounting research. Foundations and Trends® in Accounting, 14(3–4), 128–359.

Bochkay, K., Brown, S. V., Leone, A. J., & Tucker, J. W. (2022). Textual analysis in accounting: What's next? Contemporary Accounting Research.

De Kok, T. (2024). ChatGPT for Textual Analysis? How to Use Generative LLMs in Accounting Research. Management Science 71(9):7888-7906.

Gow, I. D., Larcker, D. F., & Zakolyukina, A. A. (2021). Non-answers during conference calls. Journal of Accounting Research, 59(4), 1349–1384.

Hollander, S., Pronk, M., & Roelofsen, E. (2010). Does silence speak? An empirical analysis of disclosure choices during conference calls. Journal of Accounting Research, 48(3), 531–563.

Appendix A. Spark Prompt Template

The Spark Pro/Max scripts implement a structured zero-shot prompt adapted from de Kok (2024, OA 2.3). The key design choice is to (i) provide both the question and the full answer, (ii) frame the task as evaluating a research assistant's comment, and (iii) force a single JSON object output that includes an explanation and a binary classification. In this replication, the "comments" field is set to "N/A" rather than passing matched candidate sentences from upstream filters.

```
Investor question:
{question}

Manager response:
{answer}

A research assistant has marked the above response as including a
statement that reflects unwillingness or inability to answer (part) of the
analysts' question, because of the following comment(s):
> {comments}

Based on the question and full response above, provide a detailed
assessment whether the manager's response includes a statement,
explanation, or justification indicating an inability or unwillingness to
answer the question. If you classify the response as reflecting inability
or unwillingness to answer, justify your classification with specific
phrases or sentences from the manager's response. If there's no such
indication, explain why not.

IMPORTANT OUTPUT RULES:
1) Output MUST be exactly ONE valid JSON object.
2) Do NOT include markdown code fences.
3) Do NOT include any extra text before or after the JSON.

Return JSON in this exact format:
{
  "assessment": "a detailed assessment unique to this evaluation",
  "your_classification": 1
}
```

Appendix B. Replicated Tables

TABLE 1. SAMPLE CONSTRUCTION

Total # of firms (CIK) shown in CapitalIQ database from 2013-2022 calendar year	12,614
Exclude firms that lack industry information and firms in finance and public utility industry (GIC classification code: 40 & 55)	-6,988
Exclude firms that lack earnings conference call Q&A transcripts and firms whose earnings call transcripts consistently contain fewer than five question–answer exchanges.	-69
Exclude firms for which none of their earnings call Q&A pairs satisfy the following criteria: the question contains at least 30 words, the answer contains at least 10 words, and the combined length of the question and answer is at least 75 words.	-86
Total firms and Q&A pairs in our sample	5,471 166,848 Q&A pairs

TABLE 2. NONANSWER ANALYSIS (100 OBS)

PANEL A. CONFUSION EVALUTION

Method	True Positive (TP)	False Positive(FP)	True Negative (TN)	False Negative (FN)	N
	(1)	(2)	(3)	(4)	(5)
Gow et al.	5	4	77	14	100
Spark Pro	9	7	74	10	100
Spark Max	10	5	76	9	100
Keyword + Spark Max	9	2	79	10	100

PANEL B. METHOD EVALUATION

	Manual	Gow et al. (2021)	Spark Pro	Spark Max	Keyword + Spark Max
	(1)	(2)	(3)	(4)	(5)
Answer	81	91	84	85	89
Non-answer	19	9	16	15	11
Accuracy		0.820	0.830	0.860	0.880
Type I error		0.050	0.090	0.060	0.020
Type II error		0.740	0.530	0.470	0.530
Non-answers: Precision		0.560	0.560	0.670	0.820
Non-answers: Recall		0.260	0.470	0.530	0.470
Non-answers: F1 score		0.360	0.510	0.590	0.600
Total: Precision		0.820	0.830	0.860	0.880
Total: Recall		0.820	0.830	0.860	0.880
Total: F1 score		0.82	0.83	0.86	0.88
Total cost (in RMB)	0	0	0.361	1.595	0.864
N	100	100	100	100	100

Note: The scores are based on the number of True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). Accuracy = (TP + TN) / (TP + TN + FP + FN), Type I error = FP / (FP + TN), Type II error = FN / (FN + TP), Precision = TP / (TP + FP), Recall = TP / (TP + FN), and F1 score = TP / (TP+ 0.5 * (FP+FN))

TABLE 3. PAIR-LEVEL ANALYSIS (1000 OBS)

	Obs	Mean	Std_Dev	P5	P50	P95
	(1)	(2)	(3)	(4)	(6)	(8)
Gow et al. - % non-answer	1,000	0.082	0.275	- .000	- .000	1.000
Spark Pro - % non-answer	1,000	0.133	0.340	- .000	- .000	1.000
Spark Max - % non-answer	1,000	0.119	0.324	- .000	- .000	1.000
Keyword + Spark Max - % non-answer	1,000	0.087	0.282	- .000	- .000	1.000