Detecting Evasive Answers in Financial Q&A: A Psychological Discourse Taxonomy and Lightweight Baselines

Khaled Al Nuaimi
Khalifa University, UAE
khaled.alnuaimi@adia.ae

Gautier Marti
ADIA, UAE

Alexis Marchal
ADIA, UAE

Andreas Henschel
Khalifa University, UAE
andreas.henschel@ku.ac.ae

Abstract

Q&A segments of earnings calls and central bank press conferences often contain evasive answers that avoid, obscure, or reframe the question asked. We introduce the task of *evasive answer detection* in financial Q&A and propose a multi-level taxonomy grounded in discourse pragmatics and deception psychology. Using earnings call transcripts, we curate an annotated subset and evaluate simple, interpretable baselines (surface cues, hedge detection, tense, and embedding alignment). Evasive answers show consistent linguistic and semantic signatures (e.g., present-tense bias, lower question–answer semantic alignment), providing practical signals for transparency-aware financial NLP.

1 Introduction

Transparency, the availability of firm-specific information to external stakeholders (Bushman et al., 2004), is central to market efficiency. In unscripted Q&A, however, executives can strategically avoid direct answers, distorting downstream tasks such as sentiment, risk, and event prediction. We formalize **evasive answer detection** for financial dialogue and make three contributions: (i) a discourse- and psychology-informed taxonomy of evasive strategies; (ii) an annotated subset of Q&A exchanges; and (iii) lightweight baselines that reveal robust linguistic signals of evasiveness distinct from sentiment and veracity.

Unlike sentiment or factuality, our focus is *answer responsiveness*—whether and how a reply addresses the informational intent of the question. This lens surfaces strategies such as omission, vagueness, and reframing that polarity or claim-checking may miss, and it complements topic-drift measures (Chen et al., 2025) and evidence on non-answers in earnings calls (Gow et al., 2021).

2 Related Work

2.1 Evasion vs. Sentiment and Veracity

Financial NLP has emphasized sentiment and factuality, using domain lexicons (Loughran and McDonald, 2011) and pretrained models like FinBERT (Yang et al., 2020; Liu et al., 2020). These approaches capture polarity or correctness but not whether a question was actually answered. Evasion is a pragmatic choice to be vague, tangential, or incomplete. Related deception work—e.g., *BERTective* (Fornaciari et al., 2021), explainable detectors (Ilias et al., 2022), and weakly supervised veracity frameworks (Leite et al., 2025; Irnawan et al., 2025)—generally presumes a checkable claim; evasive answers may avoid making one at all.

2.2 Strategic Communication in Financial Discourse

In earnings-call Q&A, topic divergence predicts market-relevant outcomes (Chen et al., 2025) and is consistent with strategic communication theory (Crawford and Sobel, 1982; Milgrom, 1986). Prior work documents non-answers and links language to misreporting risk (Larcker and Zakolyukina, 2012; Gow et al., 2021). We complement these scalar or outcome-focused measures with a psychologically grounded taxonomy that explains *how* evasion is executed, aligning with management obfuscation hypotheses (Bushman and Smith, 2005; Khalmetski et al., 2017).

2.3 Theoretical Foundations

Our taxonomy draws on discourse pragmatics and equivocation theory: violations of Gricean maxims signal non-responsiveness (Grice, 1975); Bavelas-style forms capture omission, vagueness, and topic shifting (Bavelas et al., 1990); and Bull/Rasiah provide institutional tactics and response types (Bull, 1998; Rasiah, 2010).

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

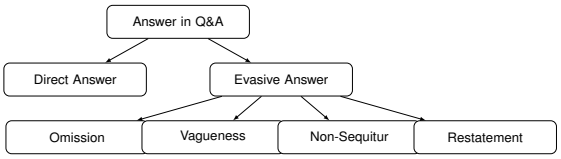


Figure 1: Mid-level taxonomy (Bavelas). Level 3 further refines each branch (Bull).

3 Task and Taxonomy

We define an evasive answer as a response that fails to directly address the core informational intent of a question via omission, ambiguity, reframing, or selective disclosure. Our taxonomy integrates discourse pragmatics (Grice, 1975), psychological equivocation (Bavelas et al., 1990), and political interview tactics (Bull, 1998; Rasiah, 2010):

Level 1 (Rasiah-style): *Direct*, *Intermediate*, *Fully Evasive*.

Level 2 (Bavelas forms): *Omission*, *Vagueness*, *Non-Sequitur*, *Restatement*.

Level 3 (Bull subtypes): *Avoidance/Deflection*, *Acknowledging Without Answer*, *Refusal to Answer*, *Agenda Shifting*, *Claiming Ignorance*, *Partial Answer/Selective Disclosure*, *Literal Interpretation*, *Repetition of Prior Material*, *Challenge Premise*, *Attack Question*, *Attack Questioner*, *External Blame*.

Table 1 provides illustrative examples annotated with our full three-level taxonomy.

4 Data and Annotations

We use a large collection of earnings call transcripts (2019–2022) originally scraped from The Motley Fool and hosted on Kaggle.[1] The corpus covers 18,755 transcripts across 2,876 companies. We extract and annotate a high-quality subset of ∼4,600 Q&A pairs (semi-automatic extraction + manual checks, we obtain 60% agreement between human and LLM annotators) across 521 transcripts from 398 companies. On average, each transcript contains nine Q&A pairs (Figure 2). Each pair receives: (i) taxonomy labels (Levels 1–3), (ii) surface features (lengths, hedge counts, lexical overlap/entropy, tense), and (iii) embedding-based similarities (cosine between question/answer).

5 Baselines

We evaluate interpretable, compute-light signals: **(a) Surface cues**: hedge counts, lexical entropy/overlap, lengths, and answer-to-question (A/Q) ratios; **(b) Tense features**: dominant answer tense and question→answer shifts; **(c) Embedding alignment**: cosine similarity between Q and A, and between the original question Q and $\hat{Q}$, a question inferred from the executive answer A using a powerful LLM, namely Claude 3.7 Sonnet.

[1] https://www.kaggle.com/datasets/gautiermarti/earnings-calls-qa-evasive-answers

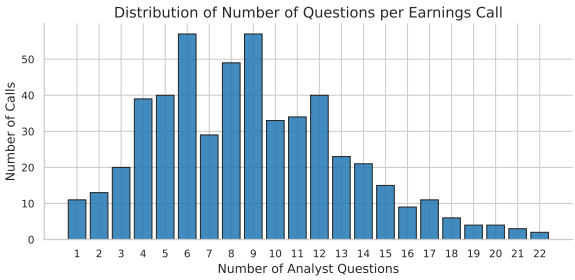


Figure 2: Distribution of the number of analyst questions per earnings call.

6 Results

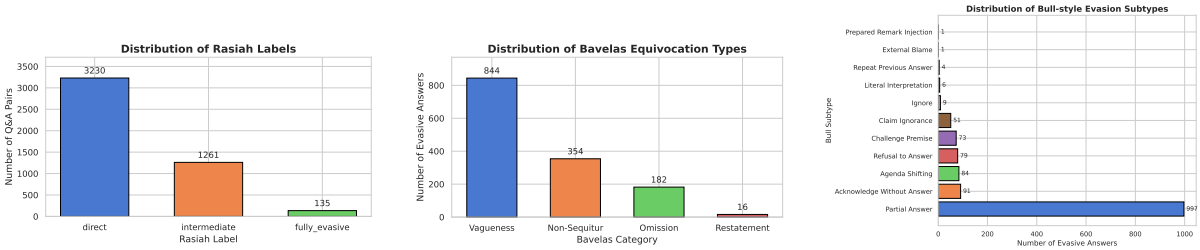
Prevalence and types. Roughly 30% of answers are evasive (labeled *intermediate* or *fully evasive*), with the remaining ∼70% *direct*. Figure 3 summarizes distributions across our ∼4,600 annotated (Q, A) pairs and three taxonomy levels. At the Bavelas level (evasive only), *Vagueness* dominates, followed by *Non-Sequitur* and *Omission*; *Restatement* is rare. At the Bull subtype level, *Partial Answer / Selective Disclosure* accounts for the largest share, with *Agenda Shifting*, *Challenge Premise*, and *Acknowledging Without Answer* also prevalent; outright *Refusal to Answer* appears but remains in the single digits. Taken together, these patterns suggest firms often offer responses that are technically informative yet pragmatically non-committal, reinforcing the value of a multi-level taxonomy beyond binary non-answer detection and aligning with prior evidence from earnings calls (Gow et al., 2021).

Surface signatures. Verbose, hedged tactics such as *Agenda Shifting* and *Partial Answer* use more hedges and longer answers, while *Refusal* is shorter with minimal hedging (Table 2).

Temporal framing. Fully evasive answers prefer future-oriented or unanchored framing: they exhibit roughly twice as many *forward* shifts (e.g., past→future or present→future; ∼29%) as direct answers (∼14%), and the highest share of *no shift* (mostly present→present; ∼49%). By contrast, direct answers engage more with the question's

Question	Answer	Rasiah	Bavelas	Subtype (Bull)
Will you revise earnings guidance for next quarter?	We remain focused on delivering long-term value to shareholders.	`fully_evasive`	Omission	Avoidance / Deflection
What explains the decline in margins?	There are several interacting macroeconomic factors, including supply chain volatility and input costs.	`intermediate`	Vagueness	Partial Answer (Obfuscation)
How will the new regulations affect your Q3 profits?	What's important to highlight is that our revenues have grown 10% this quarter.	`fully_evasive`	Non-Sequitur	Agenda Shifting (Misdirection)

Table 1: Example annotations using our three-level taxonomy: Rasiah (response type), Bavelas (evasion form), and Bull (evasion subtype).



(a) Rasiah (top-level) (b) Bavelas (evasive only) (c) Bull (evasive only)

Figure 3: Distributions across the three taxonomy levels: (a) Rasiah response types, (b) Bavelas equivocation categories, and (c) Bull-style evasion subtypes.

Subtype	Hed.	Overlap	Ans.	A/Q
Agenda Shift.	3.20	0.47	346	7.53
Partial Ans.	2.30	0.41	206	3.38
Chall. Prem.	2.11	0.39	201	3.12
Ack. w/o Ans.	1.43	0.33	111	1.84
Claim Ignor.	1.08	0.32	78	1.81
Refusal	0.78	0.29	73	2.06

Table 2: Surface cues by Bull subtypes (means). Longer, hedged answers track deflective tactics.

Figure 4: Distribution of tense shift types across Rasiah-style evasiveness levels. Fully evasive answers emphasize forward shifts; direct answers show more mixed and backward transitions.

temporal anchor—a concrete time reference conveyed by tense and/or explicit markers such as *"Q2 2022", "last quarter", "in November"*—showing higher *mixed* (∼25% vs. ∼12% for fully evasive) and *backward* (∼9% vs. ∼6%) transitions. Intermediate answers fall in between with the largest mixed share (∼29%). Intuitively, evasion either defers to the future or stays in a vague present instead of addressing past-specific, time-anchored details (Figure 4).

Semantic alignment. Cosine similarity (Q vs. A) outperforms lexical Jaccard for detecting fully evasive answers (AUC ≈ 0.79 vs. 0.68 in a direct vs. fully-evasive logistic regression setup). In a 3-class setting (direct/intermediate/fully), cosine yields higher macro-F1 than Jaccard; combining both marginally improves recall. Figure 5 shows that fully evasive responses exhibit lower Q–A similarity, whereas direct and intermediate overlap substantially.

Question recoverability ($Q \to \hat{Q}$ distance). We infer the "most" likely analyst question ($\hat{Q}$) implied by an executive answer A using an LLM (Claude 3.7 Sonnet), then measure cosine distance between the original question Q and $\hat{Q}$. Larger Q–$\hat{Q}$ distances indicate that the answer implicitly addresses a different question. Evasive subtypes (*Non-Sequitur*, *Omission*, *Challenge Premise*, *Agenda Shifting*) show the largest distances; *Direct* and *Restatement* are smallest. Q–A similarity correlates with Q–$\hat{Q}$ similarity ($\rho \approx 0.66$), suggesting a consistent semantic notion of responsiveness.

Firm Size and Transparency We quantify transparency per (Q, A) pair p for firm i by mapping

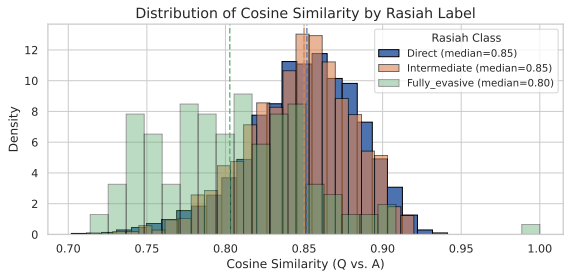


Figure 5: Distribution of cosine similarity (Q vs. A) by Rasiah label. Fully evasive answers show lower alignment.

Rasiah labels to numeric scores:

$$r_{i,p} = \begin{cases} 1 & \text{if direct,} \\ 0 & \text{if intermediate,} \\ -1 & \text{if fully evasive.} \end{cases}$$

Averaging across N pairs in a call yields the per-event score:

$$R_i = \frac{1}{N} \sum_{p=1}^{N} r_{i,p}.$$

We regress R_i on log market capitalization X_i:

$$R_i = \alpha + \beta X_i + \varepsilon_i.$$

Results (Table 3) show that the estimated coefficient β is positive and highly significant ($p < 0.01$, $t > 50$); the model fit is strong ($R^2 > 0.8$). This indicates that larger firms are more transparent, consistent with Bushman et al. (2004).

Transparency Persistence Within Calls We split each call into a first half and second half, computing $R_i^{1/2}$ and $R_i^{2/2}$, respectively. Figure 6 shows a strong linear relationship between the two. An OLS regression:

$$R_i^{2/2} = \alpha + \beta R_i^{1/2} + \varepsilon_i$$

yields $\beta \approx 0.84$ and $p < 0.01$ (Table 4), indicating that transparency (or evasiveness) early in the Q&A strongly predicts behavior later in the call.

7 Future Work

Building on our taxonomy and lightweight signals, we see six priorities: **Multimodal cues**—add prosody, hesitations, and disfluencies via audio embeddings to capture content-independent markers of evasion (Ahbabi et al., 2025); **Explainability & supervision**—apply SHAP/counterfactuals to expose drivers and run human–LLM annotation studies with agreement metrics for both binary and fine-grained labels; **Corpus-scale expansion**—extend coverage from our annotated subset to the full earnings call corpus (estimated ∼5M Q–A pairs) using top-performing LLMs, with stratified human audits to ensure quality, enabling large-scale psychological and market-behavior analyses; **Robustness & generalization**—evaluate across executives, sectors, geographies, and time, probe transfer to adjacent domains (e.g., FOMC meetings), and assess sensitivity to model/prompt choices; **Downstream integration**—plug evasiveness features into analyst tools, risk/market models; **Benchmarking & release**—expand labels with guidelines, and release code with enriched data alongside a minimal reproducible pipeline.

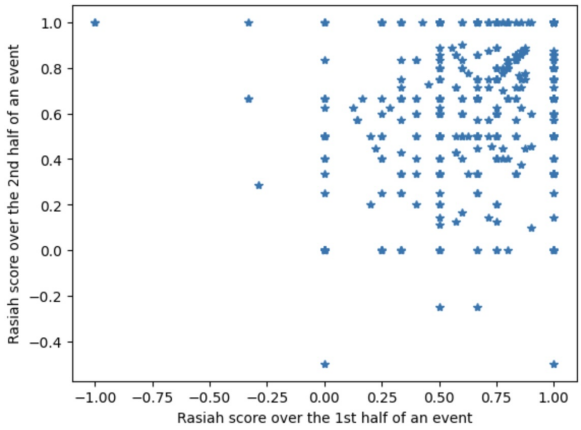


Figure 6: Average Rasiah score during the first half of an earning call against the average score during the second half.

8 Conclusion

We frame evasive answer detection as a practical task for financial Q&A, grounded in a psychology-informed taxonomy and supported by a newly annotated dataset. Our analysis shows that lightweight, interpretable features—hedges, lengths, tense, and embedding alignment—capture robust evasiveness signals distinct from sentiment or veracity. Beyond academic interest, these cues can directly enhance transparency-aware pipelines for analyst tools, regulatory triage, and explainable market surveillance. By releasing our taxonomy, annotations, and baseline models, we aim to catalyze further work on scaling detection to millions of Q–A pairs, integrating multimodal signals, and probing the role of strategic communication in financial markets.

References

Hamdan Al Ahbabi, Gautier Marti, Saeed AlMarri, and Ibrahim Elfadel. 2025. Residual speech embeddings for tone classification: Removing linguistic content to enhance paralinguistic analysis. *arXiv preprint arXiv:2502.19387*.

Janet Beavin Bavelas, Alex Black, Nicole Chovil, and Jennifer Mullett. 1990. *Equivocal communication*. Sage Publications, Inc.

Peter Bull. 1998. Equivocation theory and news interviews. *Journal of Language and Social Psychology*, 17(1):36–51.

Robert Bushman, Joseph Piotroski, and Abbie Smith. 2004. What determines corporate transparency? *Journal of Accounting Research*, 42(2):207–252.

Robert M Bushman and Abbie J Smith. 2005. Financial accounting information and corporate governance. *Journal of Accounting and Economics*, 32(1-3):237–333.

Yanzhen Chen, Huaxia Rui, and Andrew B Whinston. 2025. Conversation analytics: Can machines read between the lines in real-time strategic conversations? *Information Systems Research*, 36(1):440–455.

Vincent P Crawford and Joel Sobel. 1982. Strategic information transmission. *Econometrica*, 50(6):1431–1451.

Tommaso Fornaciari, Federico Bianchi, Massimo Poesio, Dirk Hovy, and 1 others. 2021. Bertective: Language models and contextual information for deception detection. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: Main volume*. Association for Computational Linguistics.

Ian D Gow, David F Larcker, and Anastasia A Zakolyukina. 2021. Non-answers during conference calls. *Journal of Accounting Research*, 59(4):1349–1384.

Herbert P Grice. 1975. Logic and conversation. In *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press.

Loukas Ilias, Felix Soldner, and Bennett Kleinberg. 2022. Explainable verbal deception detection using transformers. *arXiv preprint arXiv:2210.03080*.

Bassamtiano Renaufalgi Irnawan, Sheng Xu, Noriko Tomuro, Fumiyo Fukumoto, and Yoshimi Suzuki. 2025. Claim veracity assessment for explainable fake news detection. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4011–4029.

Kiryl Khalmetski, Bettina Rockenbach, and Peter Werner. 2017. Evasive lying in strategic communication. *Journal of Public Economics*, 156:59–72.

David F Larcker and Anastasia A Zakolyukina. 2012. Detecting deceptive discussions in conference calls. *Journal of Accounting Research*, 50(2):495–540.

João A Leite, Olesya Razuvayevskaya, Kalina Bontcheva, and Carolina Scarton. 2025. Weakly supervised veracity classification with llm-predicted credibility signals. *EPJ Data Science*, 14(1):16.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4513–4519. International Joint Conferences on Artificial Intelligence Organization. Special Track on AI in FinTech.

Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of finance*, 66(1):35–65.

Paul R Milgrom. 1986. Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics*, 17(3):380–391.

Parameswary Rasiah. 2010. A framework for the systematic analysis of evasion in parliamentary discourse. *Journal of Pragmatics*, 42(3):664–680.

Yi Yang, Mark Christopher Siy Uy, and Allen Huang. 2020. FinBERT: A Pretrained Language Model for Financial Communications. *arXiv preprint arXiv:2006.08097*.

A Appendix

Dep. Variable:	Rasiah_numeric_mean_per_doc	**R-squared (uncentered):**	0.847
Model:	OLS	**Adj. R-squared (uncentered):**	0.846
Method:	Least Squares	**F-statistic:**	2716.
Date:	Fri, 08 Aug 2025	**Prob (F-statistic):**	1.93e-202
Time:	14:38:48	**Log-Likelihood:**	-68.952
No. Observations:	493	**AIC:**	139.9
Df Residuals:	492	**BIC:**	144.1
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
log_market_cap_usd	0.0309	0.001	52.114	0.000	0.030	0.032

Omnibus:	151.920	**Durbin-Watson:**	1.913
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	561.721
Skew:	-1.373	**Prob(JB):**	1.06e-122
Kurtosis:	7.451	**Cond. No.**	1.00

Table 3: OLS Regression Results: Explaining the Rasiah score of an event with the firm size (log market capitalization). The high R^2 is a consequence of taking the log of the market capitalization. Without the log, the beta remains positive and significant. In general, cross-sectional distribution of market capitalization is highly skewed and exponentially distributed. Taking the log mitigates this issues and reduces the impact of outliers.

Dep. Variable:	Rasiah_numeric_second_half	**R-squared (uncentered):**	0.749
Model:	OLS	**Adj. R-squared (uncentered):**	0.748
Method:	Least Squares	**F-statistic:**	1517.
Date:	Fri, 08 Aug 2025	**Prob (F-statistic):**	8.38e-155
Time:	14:21:05	**Log-Likelihood:**	-202.19
No. Observations:	510	**AIC:**	406.4
Df Residuals:	509	**BIC:**	410.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Rasiah_numeric_first_half	0.8359	0.021	38.950	0.000	0.794	0.878

Omnibus:	43.456	**Durbin-Watson:**	2.015
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	167.363
Skew:	0.256	**Prob(JB):**	4.54e-37
Kurtosis:	5.759	**Cond. No.**	1.00

Table 4: OLS Regression Results: Forecasting the Rasiah score over the second half of a call by using the Rasiah score during the first half.