# A Shaky Voice Is Not Always a Dodge: Benchmarking Textual and Vocal Evasion Detection in Earnings Calls

**Mirae Kim**[*1], **Seonghun Jeong**[*1,2], **Youngjun Kwak**[†1]
[1]Financial Tech Lab, KakaoBank Corp. [2]Yonsei University
{melissa.kim, bentley.j, vivaan.yjkwak}@lab.kakaobank.com

## Abstract

Existing approaches to evasion detection in earnings calls focus on textual transcripts, treating evasion as a single-dimensional phenomenon. We argue that evasion in spoken communication is inherently multidimensional: beyond *what* executives say, *how* they say it carries independent and complementary information. To study these dimensions jointly, we introduce **DualEvasion**, a benchmark for evasion detection across text and audio in earnings call Q&A. The benchmark contains 505 annotated question-answer pairs from 60 earnings calls, each with two independent labels: textual evasion (direct vs. evasive) and vocal cues operationalized as speaker confidence (confident vs. unconfident). Our experiments show that state-of-the-art multimodal models struggle to detect vocal confidence, particularly on unconfident responses. Our analysis suggests these models interpret acoustic cues in isolation rather than relative to each speaker's baseline. Providing speaker-level references yields modest improvements, but a substantial gap with human performance remains.

## 1 Introduction

Earnings calls offer a rich source of managerial communication, combining prepared remarks with interactive Q&A. Prior work has examined diverse signals from these calls, including textual transcripts, acoustic features, and market time-series data, to predict financial outcomes such as stock volatility and return direction (Qin and Yang, 2019; Li et al., 2020; Sawhney et al., 2020; Yang et al., 2022; Cao et al., 2024; Yu et al., 2025). Q&A sessions are particularly informative because they are less scripted than prepared remarks and require executives to respond to analysts' questions in real time. This interactive setting has motivated research on evasive behavior in executive responses,

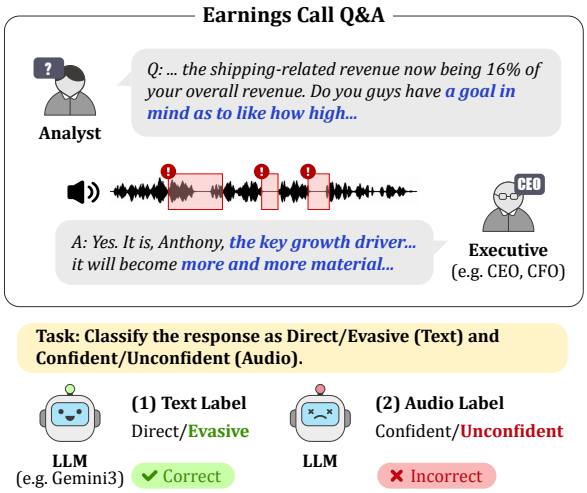


Figure 1: The challenge of **DualEvasion**: an LLM detects textual evasion but misses vocal unconfidence in the same response.

showing that evasiveness and semantic deflection can provide early signals of future performance issues and negative market reactions (Chen et al., 2025; Nuaimi et al., 2025; Hynes et al., 2026).

However, existing work treats evasion as a purely textual phenomenon, focusing on *what* executives say while overlooking *how* they say it (Table 1). We argue that these two channels carry independent information: an executive may give a textually complete answer yet sound hesitant, or skillfully deflect a question while sounding fully composed. In our data, 32.1% of instances show cross-modal disagreement—textually direct yet vocally unconfident, or vice versa (Table 3)—indicating that text alone misses an entire dimension of evasive communication. In an exploratory analysis, this missed dimension carries market-relevant signal beyond the transcript (Section 4.3).

To capture the vocal dimension, we operationalize it through speaker confidence (confident vs. unconfident), grounded in observable prosodic cues that listeners can reliably perceive (Smith and

[*] Equal contribution.
[†] Corresponding author.



*Findings of the Association for Computational Linguistics: EMNLP 2026*, pages 22366–22377
October 24-29, 2026 ©2026 Association for Computational Linguistics

Clark, 1993; Swerts et al., 2003; Jiang and Pell, 2017; Goupil et al., 2021). Confidence serves as a concrete, annotatable entry point into the vocal dimension of evasive communication, though it does not exhaust it. A key challenge is speaker dependence: what sounds hesitant for one executive may reflect another's habitual style, making speaker-aware evaluation essential (Gat et al., 2022; Van Rijn and Larrouy-Maestri, 2023).

We introduce **DualEvasion**, the first benchmark that provides independent textual and vocal labels for evasion-related behavior in earnings call Q&A, containing 505 annotated question-answer pairs from 60 calls. Our main contributions are:

- **A dual-label evasion benchmark.** We construct the first benchmark for evasion-related behavior in earnings calls with two independent annotation dimensions: textual evasion and vocal confidence. The two dimensions are only weakly correlated, supporting the need to study them jointly.

- **Speaker calibration as a key challenge.** Frontier models struggle with vocal confidence detection (Figure 1), relying on absolute acoustic cues without accounting for each speaker's baseline. Speaker-level calibration yields only modest improvements, revealing a substantial gap with human performance.

## 2 Related Work

### 2.1 Textual Evasion

Prior work has proposed various taxonomies of evasive responses. Rasiah (2010) categorizes responses as direct, intermediate, or fully evasive, while Bavelas et al. (1990) and Bull (1998) identify finer-grained strategies such as omission, vagueness, and agenda shifting. More recent work has framed evasion detection as a classification task with varying label schemes, in political discourse (Thomas et al., 2024; Sage and Greco, 2026) and in earnings calls (Gow et al., 2021; Nuaimi et al., 2025); however, these studies rely exclusively on textual features. In earnings calls, where analyst questions are goal-directed and executives are expected to provide firm-specific information, we adopt a binary formulation—direct versus evasive—informed by these finer-grained taxonomies but simplified for annotation reliability.

| | SubjECTive-QA | Nuaimi | EvasionBench | DualEvasion |
|---|---|---|---|---|
| Period | '07–'21 | '19–'22 | Multi-yr | **'23–'25** |
| Audio | ✗ | ✗ | ✗ | ✓ |
| Evasion | △ | ✓ | ✓ | ✓ |
| Label | Text | Text | Text | **Text+Vocal** |
| Annotation | Human | H+LLM | LLM | **Domain Expert** |

Table 1: Comparison with earnings call Q&A datasets (Pardawala et al., 2024; Nuaimi et al., 2025; Ma et al., 2026). Existing datasets label evasion from text only; ours additionally captures vocal cues annotated by finance professionals. △ = partial.

### 2.2 Vocal Cues and Confidence

While textual evasion has been studied across several domains, vocal signals in financial communication remain largely unexplored. Research on vocal confidence, however, provides a useful foundation. Smith and Clark (1993) show that filled pauses and response delays are associated with lower speaker confidence, while Swerts et al. (2003) and Goupil et al. (2021) demonstrate that listeners can reliably perceive speaker uncertainty from prosodic cues. Pon-Barry and Shieber (2010) further classify spoken utterances as certain or uncertain using vocal features in dialogue settings. These studies establish that confidence-related vocal cues are both perceptually salient and systematically measurable, making speaker confidence a natural operationalization for studying the vocal dimension of earnings call communication.

## 3 DualEvasion

### 3.1 Data Collection

We construct **DualEvasion** from earnings call recordings obtained via the EarningsCall API[1] under a commercial license. We target 300 tickers spanning 2023–2025, selected using FinanceDataReader[2] to ensure diversity across stock exchanges, market capitalization tiers, and industry sectors. To segment each call into question-answer pairs with speaker attribution, we run WhisperX (Bain et al., 2023) on the raw audio for sentence-level timestamps and align the result with transcripts from DefeatBeta[3], which provide speaker identity metadata. Starting from 3,592 calls with available audio, we apply three filtering criteria: (1) calls in which all analyst questions are

[1] earningscall.biz
[2] github.com/FinanceData/FinanceDataReader
[3] github.com/defeat-beta/defeatbeta-api

| Model | Direct/Confident | | | Evasive/Unconfident | | | Overall | |
|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | Precision | Recall | F1 | Macro F1 | Accuracy |
| *Textual Evasion: Direct vs. Evasive* | | | | | | | | |
| Qwen2-7B-Instruct | 79.3 | 91.6 | 85.0 | 60.8 | 35.3 | 44.7 | 64.8 | 76.4 |
| Llama-3.1-8B-IT | 84.9 | 80.8 | 82.8 | 53.9 | 61.0 | 57.2 | 70.0 | 75.4 |
| Gemini-2.5-Flash | 93.2 | 74.8 | 83.0 | 55.5 | 85.3 | 67.2 | 75.1 | 77.6 |
| Gemini-3-Flash | 88.7 | **97.6** | **92.9** | **90.9** | 66.2 | 76.6 | 84.7 | 89.1 |
| GPT-5 | **95.9** | 89.4 | 92.6 | 75.8 | **89.7** | **82.2** | **87.4** | **89.5** |
| *Vocal Cues: Confident vs. Unconfident* | | | | | | | | |
| Qwen2-Audio-7B-Instruct | 88.0 | 11.1 | 19.7 | 22.4 | 94.4 | 36.2 | 28.0 | 28.9 |
| Audio-Flamingo-3 | 80.4 | 66.2 | 72.7 | 24.7 | 40.7 | 30.8 | 51.7 | 60.8 |
| Gemini-2.5-Flash-lite | 80.6 | 66.0 | 72.6 | 25.0 | 41.7 | 31.2 | 51.9 | 60.8 |
| Gemini-2.5-Flash | 83.7 | 72.0 | 77.4 | 32.0 | 48.4 | 38.5 | 57.9 | 66.9 |
| Gemini-3-Flash | 84.0 | 67.8 | 75.0 | 30.7 | 52.5 | 38.7 | 56.8 | 64.5 |
| Gemini-3.5-Flash | 88.6 | 33.2 | 48.4 | 25.6 | 84.3 | **39.2** | 43.8 | 44.2 |
| GPT-Audio-mini | **89.0** | 10.2 | 18.4 | 22.4 | **95.4** | 36.3 | 27.3 | 28.4 |
| GPT-Audio | 83.5 | **74.6** | **78.8** | **32.8** | 45.7 | 38.2 | **58.5** | **68.4** |

Table 2: Zero-shot performance on textual evasion and vocal confidence detection. Results for commercial audio models are averaged over three runs, while all other results are from a single run. **Bold** indicates best, underline indicates second best per column within each section.

| Text | Vocal | Count | % |
|---|---|---|---|
| Direct | Confident | 302 | 59.8 |
| Direct | Unconfident | 67 | 13.3 |
| Evasive | Confident | 95 | 18.8 |
| Evasive | Unconfident | 41 | 8.1 |
| **Cross-modal disagreement** | | **162** | **32.1** |

Table 3: Joint distribution of textual evasion and vocal confidence labels. Shaded rows indicate cross-modal disagreement.

answered by a single executive, to avoid ambiguity in vocal cue assessment (110 calls); (2) at least four Q&A pairs per call, to ensure sufficient per-speaker reference material (96 calls); and (3) all responses under 300 seconds, to fit within model input constraints (93 calls). From these, we select 60 calls across 49 unique tickers. Detailed statistics on the dataset composition are provided in Appendix A.

### 3.2 Annotation and Validation

The final benchmark contains 505 question-answer pairs from the 60 selected calls, each annotated along two dimensions: textual evasion and vocal confidence. Two financial domain experts independently annotate all pairs, followed by a separate reliability check on a subset. Annotation details are in Appendix B.

**Textual Evasion Labels.** Each response is labeled as either *direct* or *evasive* based on whether the executive substantively addresses the analyst's question, with responses that avoid, deflect, or only partially address the core issue classified as evasive (Cohen's $\kappa$[4] $= 0.866$, 94.7% agreement).

**Vocal Confidence Labels.** Vocal confidence labels are assigned in a separate pass: annotators listen to audio responses only, without reference to the textual labels or transcript content. Each response is labeled as either *confident* or *unconfident* based on prosodic cues such as filled pauses, hesitation, and intonation patterns. To account for speaker-specific vocal characteristics, annotators first listen to the executive's other responses within the same call to establish a per-speaker baseline before making judgments (Cohen's $\kappa = 0.774$, 87.5% agreement).

**Validation.** To check whether agreement generalizes beyond the two primary annotators, five financial domain experts—including the original two, using their pre-adjudication labels—independently label a random subset of 52 responses for vocal confidence, yielding Fleiss' $\kappa$[5] of 0.713 (substantial agreement; pairwise $\kappa$ in Table 9). These validation labels serve only this reliability check and do not affect the final labels, which are produced by the two experts through adjudication. The resulting label distribution is shown in Table 3; the two dimensions are only weakly correlated, with 32.1% of instances showing cross-modal disagreement.

[4]Cohen's $\kappa$ (Cohen, 1960) measures inter-rater agreement beyond chance, ranging from $-1$ (systematic disagreement) to 1 (perfect agreement).

[5]Fleiss' $\kappa$ (Fleiss, 1971) extends Cohen's $\kappa$ to three or more raters.

| | Text | Vocal |
|---|---|---|
| LLM–Annotator | 0.813 | 0.225 |
| Annotator–Annotator | 0.866 | 0.774 |

Table 4: Cohen's $\kappa$ agreement with human annotators, averaged over two annotators, and human–human agreement. GPT-5 is used for text and Gemini-3-Flash for vocal confidence.

## 4 Experiment

We evaluate models across both dimensions in a zero-shot setting, using the human annotations as ground truth. For textual evasion, we test open-source and commercial language models on transcript text. For vocal confidence, we test audio-capable models on response audio. Main results are reported in Table 2, with all prompts in Appendix G.

### 4.1 A Modality Gap in Model Performance

Table 2 reveals a clear modality gap. For textual evasion, frontier models perform well: GPT-5 and Gemini-3-Flash reach 87.4 and 84.7 macro F1, respectively. Vocal confidence detection, however, proves far more challenging—the best audio model achieves only 58.5 macro F1. As Table 4 shows, GPT-5's agreement with annotators on textual evasion approaches human–human levels, whereas Gemini-3-Flash's agreement on vocal confidence remains substantially lower.

The major bottleneck lies in the unconfident class, where F1 ranges from 30.8 to 39.2—far below the confident class (up to ∼79 F1). This weakness cannot be explained by class imbalance alone: the strongest models exceed the 44.0 macro-F1 all-confident majority baseline (e.g., GPT-Audio at 58.5), while weaker models fail in the opposite direction, overpredicting the unconfident class with recall as high as 95.4. Neither pattern is consistent with a simple majority-class bias. Threshold tuning and supervised fine-tuning likewise fail to close the gap (Appendix C). These results point to a persistent difficulty in detecting unconfidence, which we investigate further in Section 4.2.

### 4.2 Speaker Calibration

**Acoustic analysis.** One explanation for this difficulty is that models judge confidence from raw acoustic values rather than relative to each speaker's baseline. To test this, we measure the point-biserial correlation between four acoustic features and the unconfident label under two conditions: absolute feature values and speaker-relative deviations (each response minus the speaker's mean). We report the mean absolute correlation across features before ($|r_{\text{raw}}|$) and after ($|r_{\text{cent}}|$) speaker-centering, with significance estimated by a cluster bootstrap over the 60 calls (Table 5).

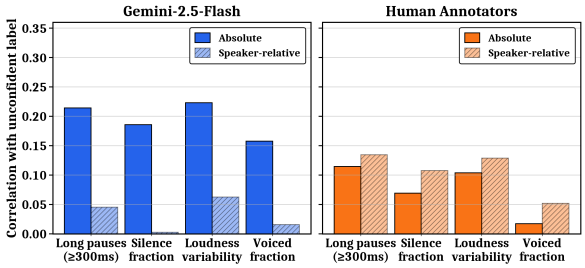


Figure 2: Correlation between acoustic features and unconfident predictions for Gemini-2.5-Flash, the model with the largest gap, measured using absolute values and speaker-relative deviations (response value minus the speaker's mean).

| | $\lvert r_{\text{raw}}\rvert$ | $\lvert r_{\text{cent}}\rvert$ | **Gap** ($p$) |
|---|---|---|---|
| *Models* | | | |
| Gemini-2.5-Flash | .195 | .032 | **.164 (.005)** |
| Gemini-3-Flash | .179 | .045 | **.135 (.003)** |
| GPT-Audio | .142 | .030 | **.113 (.025)** |
| Gemini-2.5-Flash-lite | .093 | .013 | .080 (.234) |
| Audio-Flamingo-3 | .040 | .015 | .025 (.522) |
| *Human annotators* | | | |
| A1 | .074 | .099 | −.025 (.510) |
| A2 | .079 | .112 | −.034 (.398) |
| Consensus | .088 | .104 | −.016 (.654) |

Table 5: Mean absolute point-biserial correlation between acoustic features and the unconfident label, before ($|r_{\text{raw}}|$) and after ($|r_{\text{cent}}|$) speaker-centering. Gap $= |r_{\text{raw}}| - |r_{\text{cent}}|$; $p$-values from a cluster bootstrap over the 60 calls. **Bold** indicates gaps significant at $p < .05$.

Across the three strongest models, speaker-centering sharply reduces the correlation—for Gemini-2.5-Flash, from .195 to .032 ($p < .01$)—indicating reliance on raw acoustic values rather than each speaker's baseline (Figure 2). Human annotators show no such drop; their correlations remain stable or increase slightly, consistent with speaker-aware judgment. Weaker models show smaller gaps, but their raw correlations are already near zero, so little remains to remove. Feature definitions and extraction details are in Appendix D.

**Calibration experiments.** Given this apparent reliance on absolute cues, we test whether providing explicit speaker context helps. In our main evaluation, models receive only a single response

| Model | Baseline | +Speaker Norm. | ΔF1 |
|---|---|---|---|
| Gemini-2.5-Flash | 38.5 | 43.2 | **+4.7** |
| Gemini-3-Flash | 38.7 | 41.6 | **+2.9** |
| GPT-Audio | 38.2 | 39.2 | **+1.0** |

Table 6: Effect of speaker-level normalization on unconfident F1. The model infers the speaker's vocal baseline before classifying the target response.

audio, with no access to other utterances from the same speaker. We first provide same-speaker reference audio as few-shot examples, but this yields no consistent improvement (results in Appendix E). We then prompt models to first infer the speaker's habitual vocal baseline (such as pitch, tempo, and prosody) from three same-speaker reference utterances—selected sequentially and excluding the target—before judging whether the target response deviates from it (prompt template in Table 19). As shown in Table 6, this modestly improves unconfident F1 (e.g., +4.7 for Gemini-2.5-Flash), suggesting that speaker-level calibration is a promising direction, though the gap with human performance remains large.

### 4.3 Market Relevance

To examine whether our labels carry market-relevant information, we regress call-level evasive and unconfident ratios against post-earnings stock volatility over 3- to 30-day windows (Figure 3). Because **DualEvasion** contains only a subset of Q&A pairs for some calls, we repeat the analysis while progressively restricting the sample to calls with more complete Q&A coverage. Across these coverage settings, audio-based unconfidence generally explains more variance than textual evasion, and combining both yields the highest $R^2$, suggesting that the two dimensions capture complementary signals. This association is strongest over shorter horizons ($v_3$, $v_7$) and weakens by $v_{30}$. Given the limited sample of 60 calls, we treat these results as exploratory rather than causal; regression and volatility details are provided in Appendix F.

## 5 Conclusion

We introduced **DualEvasion**, a benchmark for evasion detection across text and audio in earnings call Q&A, containing 505 annotated question-answer pairs with independent textual evasion and vocal confidence labels. Our experiments reveal that models approach human-level performance on textual evasion but fall far short on vocal confidence

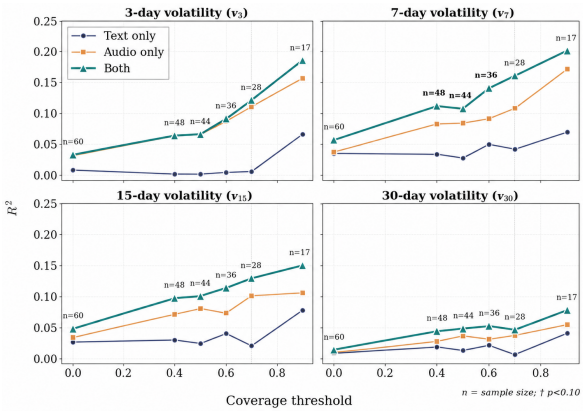


Figure 3: $R^2$ of text-only, audio-only, and combined regression models for post-earnings stock volatility across Q&A coverage thresholds. Coverage is the fraction of each call's Q&A pairs included in **DualEvasion**; at each threshold, only calls with coverage at or above that level are retained. Annotations indicate the number of retained calls ($n$); †: $p < 0.10$.

detection. Acoustic analysis shows that models rely on absolute vocal features rather than speaker-relative deviations, conflating habitual speaker traits with uncertainty cues. Speaker-level calibration yields modest improvements but does not close the gap, suggesting that speaker-aware audio understanding remains an open challenge for current models.

## 6 Limitations

Our benchmark contains 505 question-answer pairs from 60 calls. The dual-annotation design—requiring independent textual and vocal labeling, with vocal annotation further requiring per-speaker calibration across all responses in the same call—makes annotation substantially more time-intensive than standard labeling tasks, limiting the scale we could achieve. Our vocal dimension is operationalized through speaker confidence, which captures one aspect of vocal cues but does not exhaust them. Finally, our stock volatility analysis is based on a small sample and should be interpreted as exploratory.

## References


Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. Whisperx: Time-accurate speech transcription of long-form audio. *INTERSPEECH 2023*.

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.

Yupeng Cao, Zhi Chen, Qingyun Pei, Nathan Lee, K. P. Subbalakshmi, and Papa Momar Ndiaye. 2024. Ecc analyzer: Extracting trading signal from earnings conference calls using large language model for stock volatility prediction. In *Proceedings of the 5th ACM International Conference on AI in Finance*, ICAIF '24, page 257–265, New York, NY, USA. Association for Computing Machinery.

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

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Joseph Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Itai Gat, Hagai Aronowitz, Weizhong Zhu, Edmilson Morais, and Ron Hoory. 2022. Speaker normalization for self-supervised speech emotion recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7342–7346. IEEE.

Louise Goupil, Emmanuel Ponsot, Daniel Richardson, Gabriel Reyes, and Jean-Julien Aucouturier. 2021. Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature. *Nature communications*, 12(1):861.

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

Liam Hynes, Henry Chiang, John Garvey, and Fergal O'Brien. 2026. The language of evasion: How semantic similarity between questions and answers predicts stock returns. *Journal of Behavioral Finance*, pages 1–12.

Xiaoming Jiang and Marc D Pell. 2017. The sound of confidence and doubt. *Speech Communication*, 88:106–126.

Jiazheng Li, Linyi Yang, Barry Smyth, and Ruihai Dong. 2020. Maec: A multimodal aligned earnings conference call dataset for financial risk prediction. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 3063–3070, New York, NY, USA. Association for Computing Machinery.

Shijian Ma, Yan Lin, and Yi Yang. 2026. Evasionbench: A large-scale benchmark for detecting managerial evasion in earnings call q&a. *Preprint*, arXiv:2601.09142.

Khaled Al Nuaimi, Gautier Marti, Alexis Marchal, and Andreas Henschel. 2025. Detecting evasive answers in financial Q&A: A psychological discourse taxonomy and lightweight baselines. In *Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing*, pages 191–196, Suzhou, China. Association for Computational Linguistics.

Huzaifa Pardawala, Siddhant Sukhani, Agam Shah, Veer Kejriwal, Abhishek Pillai, Rohan Bhasin, Andrew DiBiasio, Tarun Mandapati, Dhruv Adha, and Sudheer Chava. 2024. Subjective-qa: measuring subjectivity in earnings call transcripts' qa through six-dimensional feature analysis. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA. Curran Associates Inc.

Heather Pon-Barry and Stuart M Shieber. 2010. Recognizing uncertainty in speech. *EURASIP journal on advances in signal processing*, 2011(1):251753.

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

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

Archie Sage and Salvatore Greco. 2026. Kclarity at semeval-2026 task 6: Encoder and zero-shot approaches to political evasion detection. *arXiv preprint arXiv:2603.06552*.

Ramit Sawhney, Piyush Khanna, Arshiya Aggarwal, Taru Jain, Puneet Mathur, and Rajiv Ratn Shah. 2020. VolTAGE: Volatility forecasting via text audio fusion with graph convolution networks for earnings calls. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8001–8013, Online. Association for Computational Linguistics.

Vicki L Smith and Herbert H Clark. 1993. On the course of answering questions. *Journal of memory and language*, 32(1):25–38.

Marc Swerts, Emiel Krahmer, Pashiera Barkhuysen, and Lennard van de Laar. 2003. Audiovisual cues to uncertainty.

Konstantinos Thomas, Giorgos Filandrianos, Maria Lymperaiou, Chrysoula Zerva, and Giorgos Stamou. 2024. "I never said that": A dataset, taxonomy and baselines on response clarity classification. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5204–5233, Miami, Florida, USA. Association for Computational Linguistics.

Pol Van Rijn and Pauline Larrouy-Maestri. 2023. Modelling individual and cross-cultural variation in the mapping of emotions to speech prosody. *Nature Human Behaviour*, 7(3):386–396.

Linyi Yang, Jiazheng Li, Ruihai Dong, Yue Zhang, and Barry Smyth. 2022. Numhtml: Numeric-oriented hierarchical transformer model for multi-task financial forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11604–11612.

Ding Yu, Zhuo Liu, and Hangfeng He. 2025. Same company, same signal: The role of identity in earnings call transcripts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18403–18422, Vienna, Austria. Association for Computational Linguistics.

# A Dataset Statistics

We summarize the dataset composition in Table 7 and response-level statistics in Table 8. The benchmark spans 11 sectors and three market groups, with response durations ranging from under 1 second to nearly 5 minutes.

| **Composition** | |
|---|---|
| Earnings calls | 60 |
| Q&A pairs | 505 |
| Unique tickers | 49 |
| Call period | 2023–2025 |
| Market cap range | \$56M–\$189B |
| **Market Group** | |
| NYSE | 27 (45%) |
| NASDAQ | 17 (28%) |
| S&P 500 | 16 (27%) |
| **Sector** | |
| Technology | 13 |
| Industrials | 10 |
| Financial Services | 8 |
| Energy | 5 |
| Consumer Cyclical | 5 |
| Healthcare | 4 |
| Basic Materials | 4 |
| Communication Svcs. | 4 |
| Real Estate | 3 |
| Consumer Defensive | 2 |
| Utilities | 2 |

Table 7: Dataset composition.

| | Med. | Min | Max |
|---|---|---|---|
| Response duration (sec) | 66.9 | 0.2 | 296.9 |
| Response length (tokens) | 186 | 2 | 918 |
| Q&A pairs per call | 10 | 4 | 10 |
| Total audio duration | | | 10.9 hrs |

Table 8: Response duration and length statistics.

# B Annotation Details

## B.1 Annotation Guidelines

The following guidelines were provided to all annotators for both textual evasion and vocal confidence labeling (Tables 11 and 12).

## B.2 Annotation Interface

Figure 4 shows the annotation interface used for vocal confidence labeling. Annotators listen to each response audio and select one of the two labels. A comment field is provided for ambiguous cases.

## B.3 Inter-Annotator Agreement

Table 9 reports pairwise Cohen's $\kappa$ among the five validation annotators for vocal confidence. We conducted this additional validation only for vocal confidence, as textual evasion already achieved 94.7% raw agreement in the main annotation. The comment field was optional; nine comments were provided in total, all for unconfident judgments, citing cues such as a large sigh, slower-than-usual speech, or pauses within the utterance. For comparison, Table 10 reports pairwise agreement among the three strongest audio models, which is markedly lower than human inter-annotator agreement, further illustrating the difficulty of vocal confidence detection.

| | **A1** | **A2** | **A3** | **A4** | **A5** |
|---|---|---|---|---|---|
| **A1** | – | .915 | .706 | .651 | .917 |
| **A2** | | – | .623 | .651 | .834 |
| **A3** | | | – | .442 | .713 |
| **A4** | | | | – | .662 |
| **A5** | | | | | – |

Table 9: Pairwise Cohen's $\kappa$ for vocal confidence labels on the validation subset ($N$ = 52).

| | **Gem-2.5** | **Gem-3.0** | **GPT-Aud** |
|---|---|---|---|
| **Gem-2.5** | – | .483 | .309 |
| **Gem-3.0** | | – | .407 |
| **GPT-Aud** | | | – |

Table 10: Inter-model agreement on vocal confidence labels measured by pairwise Cohen's $\kappa$ ($N$ = 52). Gem-2.5 denotes Gemini-2.5-Flash, Gem-3.0 denotes Gemini-3-Flash, and GPT-Aud denotes GPT-Audio.

**Textual Evasion Annotation Guidelines**

Evaluate whether the executive substantively addresses the analyst's question. Assess only the textual content, not vocal delivery.

**Direct**: The response substantively addresses the core question with a relevant and specific explanation, even if no exact figures are provided.

**Evasive**: The response fails to address the core question. Label as evasive if the response exhibits any of the following patterns:

*General forms*:
- Omission of key information
- Vague or non-specific language
- Non-sequitur (unrelated response)
- Restatement of the question or prior remarks

*Specific strategies*:
- Avoidance or deflection
- Acknowledging the question without answering it
- Explicit refusal to answer
- Shifting to a different topic
- Claiming ignorance or lack of information
- Providing only a partial answer or selectively disclosing information
- Interpreting the question too literally
- Repeating previously stated material
- Challenging the premise of the question
- Questioning the question itself or the questioner
- Attributing issues to external factors

**Important:**
- Explicitly declining to answer (e.g., "we can't disclose," "I don't have that number") is *evasive*.
- If uncertain, choose the closer label and leave a comment.

Table 11: Annotation guidelines for textual evasion.

**Vocal Confidence Annotation Guidelines**

Evaluate the executive's vocal attitude. The content of the answer is **not** the subject of evaluation. Assess only the speaker's confidence level as expressed through their voice.

**Confident**:
- Stable and clear voice
- Consistent speech rate and rhythm
- Minimal fillers or stuttering
- Fluent and natural delivery
- Calm and controlled overall tone

**Unconfident**:
- Unstable or trembling voice
- Frequent or prolonged pauses
- Frequent fillers (um, uh)
- Irregular speech rate or stuttering
- Tense overall tone

**Important:**

**1. Consider speaker characteristics.**
Listen to all responses from the same speaker before rating.
- If a speaker habitually speaks slowly, do not label them unconfident solely for slow tempo.
- If a speaker frequently uses fillers (e.g., "um"), do not label them unconfident for that reason alone.

**2. Evaluate voice, not content.**
Focus on *how* the executive speaks, not *what* they say.
- A negative message delivered in a stable voice is *confident*.
- A positive message delivered with trembling or stuttering is *unconfident*.

**3. Always choose one label.**
If uncertain, choose the closer option and leave a comment.

Table 12: Annotation guidelines for vocal confidence.

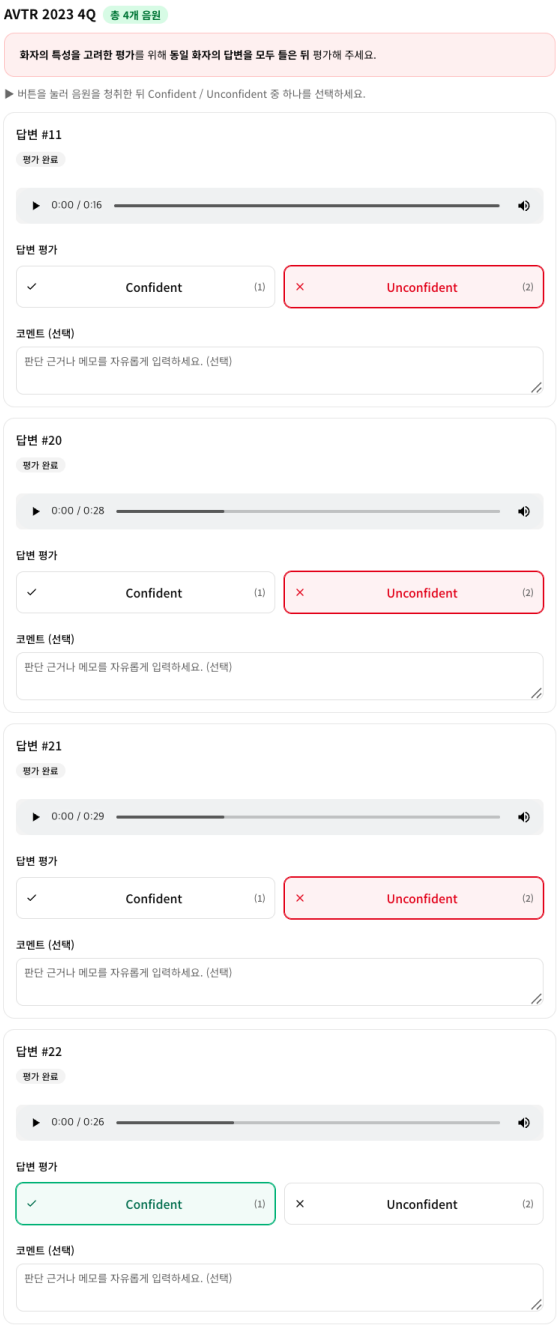


Figure 4: Annotation interface for vocal confidence.

## C Additional Baselines and Fine-Tuning

### C.1 Baseline and Threshold Analysis

A majority (all-confident) baseline yields only 44.0 macro F1 (confident F1 88.0, unconfident F1 0.0; accuracy 78.6). The strongest audio models exceed this baseline, while others fall below it by overpredicting the unconfident class, so the low scores are not a simple majority-class artifact.

We further check whether the decision threshold explains the weak unconfident performance. Commercial models return only discrete labels, so we sweep the decision threshold from 0.05 to 0.95 for the two open models that expose logits. Table 13 reports, for each model, the threshold maximizing macro F1 and the threshold maximizing unconfident F1. Even at the threshold that maximizes unconfident F1, performance reaches only 36.7 and 35.5, barely above the untuned scores of 36.2 and 30.8. Further prioritizing unconfident detection sharply degrades confident-class performance and drives macro F1 below the 44.0 majority baseline. Threshold adjustment thus shifts the precision–recall trade-off but does not resolve the weak unconfident performance.

### C.2 Supervised Fine-Tuning

To test whether the weak vocal confidence performance can be overcome with task-specific supervision, we fine-tune three open speech models on **DualEvasion**. Given the limited dataset size (505 pairs), we use five-fold cross-validation, ensuring that every instance is evaluated on a held-out fold. Table 14 summarizes the results.

Fine-tuning improves overall macro F1 for some models, most notably Qwen2-Audio (+20.6), but these gains do not translate into better detection of the unconfident class. For Qwen2-Audio, the improvement largely comes from recovering confident-class performance from its near-collapse in the zero-shot setting, while unconfident F1 decreases from 36.2 to 32.4. Audio-Flamingo-3 shows only a modest gain in macro F1 (+1.8), with unconfident F1 remaining low at 34.7. VibeVoice-ASR similarly fails on the minority class, reaching only 9.9 unconfident F1 despite a macro F1 of 48.8. Across all fine-tuned models, unconfident F1 remains at or below 34.7, suggesting that task-specific supervision alone does not close the gap in unconfident detection.

| Model | Thr. | Conf. F1 | Unconf. F1 | Macro F1 |
|---|---|---|---|---|
| Audio-Flamingo-3 | $0.40^{M}$ | 82.7 | 30.0 | **56.4** |
| | $0.95^{U}$ | 2.5 | **35.5** | 19.0 |
| Qwen2-Audio | $0.05^{M}$ | 65.2 | 34.3 | **49.7** |
| | $0.45^{U}$ | 24.4 | **36.7** | 30.5 |

Table 13: Threshold sweep for the two open models that expose logits. $^{M}$: threshold maximizing macro F1; $^{U}$: threshold maximizing unconfident F1. **Bold** marks the maximized metric.

| Model | Setting | Unconf. F1 | Macro F1 |
|---|---|---|---|
| Qwen2-Audio | zero-shot | 36.2 | 28.0 |
| | fine-tuned | 32.4 (-3.8) | 48.6 (+20.6) |
| Audio-Flamingo-3 | zero-shot | 30.8 | 51.7 |
| | fine-tuned | 34.7 (+3.9) | 53.5 (+1.8) |
| VibeVoice-ASR | fine-tuned | 9.9 | 48.8 |

Table 14: Supervised fine-tuning with five-fold cross-validation. Values in parentheses indicate changes from zero-shot performance. Fine-tuning improves overall performance but does not consistently improve unconfident detection.

## D Acoustic Feature Extraction

All features are extracted from 16 kHz mono audio using `librosa` and `parselmouth` (Praat). Frame-level RMS energy (root-mean-square amplitude) is computed with 25 ms windows and 10 ms hops. For each feature, we compute both the **absolute** value and the **speaker-relative** value (response value minus the speaker's mean across all responses in the same call).

**Silence fraction.** RMS energy is converted to dB and peak-normalized. Frames below $-35$ dB are marked as silent. Silence fraction is the proportion of silent frames.

**Long pauses.** From the silence mask above, we identify contiguous silent segments $\geq$300 ms (30 frames). Long pause fraction is the total duration of these segments divided by response duration.

**Loudness variability.** Standard deviation of the peak-normalized RMS energy in dB across all frames.

**Voiced fraction.** Fundamental frequency (F0) is estimated using Praat's autocorrelation method (`to_pitch`, pitch floor = 75 Hz, ceiling = 400 Hz, time step = 10 ms). Voiced fraction is the proportion of frames where F0 $>$ 0, i.e., where vocal fold vibration is detected.

| Model | Methods | Confident | | | Unconfident | | | Macro-F1 |
|---|---|---|---|---|---|---|---|---|
| | | Precision | Recall | F1 | Precision | Recall | F1 | |
| Gemini-2.5-Flash | baseline | **83.7** | 72.0 | 77.4 | 32.0 | **48.4** | 38.5 | 57.9 |
| | fewshot($n$=1) | 83.6 | 77.1 | 80.2 | 34.5 | 44.4 | **38.9** | **59.5** |
| | fewshot($n$=2) | 81.1 | 85.6 | 83.3 | 34.1 | 26.9 | 30.1 | 56.7 |
| | fewshot($n$=3) | 81.1 | **91.9** | **86.2** | **41.8** | 21.3 | 28.2 | 57.2 |
| Gemini-3-Flash | baseline | 84.0 | 67.8 | 75.0 | 30.7 | **52.5** | 38.7 | 56.8 |
| | fewshot($n$=1) | **84.2** | 81.9 | 83.0 | 39.5 | 43.5 | **41.4** | **62.2** |
| | fewshot($n$=2) | 82.5 | 88.2 | **85.3** | **42.0** | 31.5 | 36.0 | 60.6 |
| | fewshot($n$=3) | 82.2 | **88.4** | 85.2 | 41.0 | 29.6 | 34.4 | 59.8 |
| GPT-Audio | baseline | **83.5** | 74.6 | 78.8 | 32.8 | **45.7** | **38.2** | **58.5** |
| | fewshot($n$=1) | 80.9 | 84.1 | 82.5 | 31.5 | 26.9 | 29.0 | 55.7 |
| | fewshot($n$=2) | 81.1 | **86.4** | **83.7** | **34.1** | 25.9 | 29.5 | 56.6 |
| | fewshot($n$=3) | 79.8 | 84.4 | 82.0 | 27.1 | 21.3 | 23.8 | 52.9 |

Table 15: Effect of few-shot same-speaker reference audio on vocal confidence detection. **Bold** indicates the best score and shaded indicates the lowest score within each model.

## E Few-Shot Speaker Calibration

A natural strategy for improving vocal confidence detection is to mimic how human annotators work: before judging a target response, first listen to several other responses from the same speaker to establish a personal baseline. We operationalize this by prepending $n \in \{1, 2, 3\}$ same-speaker reference audio clips from the same earnings call before the target clip and asking the model to judge the target relative to these references. The corresponding prompt is provided in Table 18.

Table 15 shows that this approach yields no consistent improvement. While Gemini-2.5-Flash and Gemini-3-Flash see modest gains in macro-F1 at $n$=1 (+1.6 and +5.4, respectively), performance on the unconfident class degrades as more reference samples are added: unconfident F1 drops from 38.5 to 28.2 for Gemini-2.5-Flash at $n$=3, and from 38.2 to 23.8 for GPT-Audio. Across all models, additional reference audio generally increases confident recall relative to the baseline, while reducing unconfident recall, suggesting that the added context shifts predictions toward the confident class rather than improving sensitivity to speaker-relative deviations.

One possible explanation is that current audio LLMs struggle to form a stable speaker-level baseline from in-context examples. Unlike human annotators, who can integrate multiple utterances into a robust sense of a speaker's typical delivery, models may remain sensitive to the particular reference clips provided. As a result, simply adding more same-speaker examples may not provide a consistent basis for identifying subtle deviations in the target response.

## F Stock Volatility Analysis

**Return and volatility computation.** Daily returns are computed from closing prices as:

$$r_t = \frac{P_t - P_{t-1}}{P_{t-1}} \tag{1}$$

where $P_t$ is the closing price on trading day $t$. Volatility over a $k$-day window is then:

$$v_k = \log\left(\sqrt{\frac{\sum_{i=1}^{k}(r_i - \bar{r})^2}{k}}\right) \tag{2}$$

where $\bar{r}$ is the mean return over the window.

**Trading Day 0.** We define Trading Day 0 based on the earnings call timing relative to market hours. For calls released before market open (BMO), the announcement date is Trading Day 0. For calls released after market close (AMC) or during market hours (DMH), the next trading day is Trading Day 0. This convention ensures a full post-call trading session and avoids mixing pre-call and post-call price movements for DMH calls.

**Coverage threshold.** We define Q&A coverage as the fraction of all Q&A pairs in a call that are included in **DualEvasion**. For each coverage threshold, we retain only calls with coverage at or above that level and rerun the regression analysis. For example, a threshold of 0.6 includes only calls for which at least 60% of the Q&A pairs are represented in our benchmark. Higher thresholds therefore provide more complete call-level estimates but reduce the number of calls available for analysis (Figure 3).

## G Prompt Templates

Table 16 and Table 17 show the prompts used for the main textual evasion and vocal confidence evaluations, respectively. Table 18 and Table 19 show the prompts used for the few-shot and speaker-normalized calibration experiments.

**Textual Evasion Classification**

You are a financial analyst evaluating evasiveness in earnings call Q&A.

**Input**
You will receive a Question and an Answer as text. Evaluate semantic content — specificity, relevance, and completeness of the response.

**Classification**
• `direct`: Addresses the core question with a relevant and substantive explanation, even if no exact figures are provided.
• `evasive`: Fails to address the core question. This includes: shifting to a different topic, omitting key details, giving vague or generic statements, or explicitly refusing to answer.

**Output Format**
Classification: `<direct | evasive>`

Now classify the following:
Question: `{question}`
Answer: `{answer}`

Table 16: Prompt for textual evasion classification.

**Vocal Confidence Classification**

You are a financial analyst evaluating vocal confidence in earnings call Q&A.

**Input**
You will receive an audio recording of an executive's response during an earnings call. Evaluate the speaker's vocal confidence based on prosodic cues such as tone, pace, hesitation, and delivery style.

**Classification**
• `confident`: Clear, decisive, and stable vocal delivery.
• `unconfident`: Shows vocal uncertainty, hesitation, or wavering in delivery.

**Output Format**
Classification: `<confident | unconfident>`

Now classify the speaker's vocal confidence in the provided audio.

Table 17: Prompt for vocal confidence classification (audio-only).

**Few-Shot Speaker Calibration**

You are a financial analyst evaluating vocal confidence in earnings call Q&A.

**Input**
You will receive `{baseline_count}` same-speaker reference Answer audios from the same call, followed by a target Answer audio. Use the reference audio as context when judging the target's vocal confidence.

**Classification**
• `confident`: Shows a clear and stable delivery, with no notable signs of increased uncertainty relative to the reference audio.
• `unconfident`: Shows increased uncertainty relative to the reference audio, such as greater hesitation, shakiness, or instability in delivery.

**Output Format**
Classification: `<confident | unconfident>`
Reasoning: `<single sentence describing the vocal cues supporting the decision>`

Reference Answer audios:
Target Answer audio:

Table 18: Prompt for few-shot vocal confidence classification with same-speaker reference audio.

**Speaker-Normalized Vocal Confidence Classification**

You are a financial analyst evaluating vocal confidence in earnings call Q&A.

**Input**
You will receive three same-speaker reference Answer audios from the same call, followed by one target Answer audio.

**Step 1.** Infer the speaker's typical delivery from the reference audios, considering speech rate, pitch/volume stability, articulation, prosody, fillers/pauses, and cadence.
**Step 2.** Judge whether the target shows reduced vocal confidence relative to this speaker-specific baseline.

**Classification**
• `confident`: The target remains broadly consistent with the speaker's typical delivery, without notable increases in hesitation or instability.
• `unconfident`: The target shows increased uncertainty relative to the speaker's baseline, reflected in changes such as hesitation, shakiness, unstable prosody, or disrupted delivery.

**Output Format**
Classification: `<confident | unconfident>`
Reasoning: `<single sentence identifying the main vocal cues and comparing the target to the speaker's baseline>`

Reference Answer audios:
Target Answer audio:

Table 19: Prompt for speaker-normalized vocal confidence classification with explicit speaker-baseline inference.