

REVIEW ARTICLE



More Than Words: Managerial Tone, Communication Modalities, and Multimodal Measures in Earnings Calls

Rodrigo dos Reis Maia^{1*} and Jorge Miguel Bravo¹

¹NOVA Information Management School, Universidade NOVA de Lisboa, Portugal

Abstract: Earnings conference calls are inherently multimodal communication events, yet research on managerial tone has developed predominantly through text-centered measurement strategies. This study addresses three research questions concerning how tone is measured, which communication modalities are examined, and what opportunities exist for multimodal corporate communication research within top-tier accounting and finance venues. A hybrid systematic literature review and bibliometric analysis of 178 peer-reviewed publications (166 journal articles and 12 conference papers) was conducted, in accordance with PRISMA 2020 guidelines, spanning 2010–2026. Our bibliometric tools include keyword co-occurrence networks, bibliographic coupling, fractional authorship attribution, and international co-authorship mapping. Furthermore, an algorithmic binary coding across a 31-dimensional thematic taxonomy was applied, validated via dual-coder review on a stratified 60-paper subsample (macro precision = 0.852, recall = 0.739, $F_1 = 0.777$). A significant structural shift toward transformer-based architectures emerges after 2023 (Cochran–Armitage $p_{BH} = 0.018$), expanding alongside, rather than displacing, established lexicon-based and deep learning baselines. Although textual analysis still anchors the literature, vocal and interactional modeling remain nascent, video analytics is virtually untapped in premier venues, and international collaboration remains heavily concentrated within the US–Canada corridor. To bridge these gaps, a six-part research roadmap targeting multimodal fusion architectures is proposed: domain-adapted vocal models, cross-linguistic generalizability, within-call dynamics analysis, video call analysis, and multimodal deception detection.

Keywords: managerial tone, earnings conference calls, multimodal sentiment analysis, bibliometric analysis, systematic literature review

1. Introduction

Managerial communication has long been recognized as a critical channel through which executives disclose information to capital markets. Earnings calls, quarterly events in which senior management presents financial results and field questions from analysts, have become a principal empirical setting for studying this process. Although much of the literature treats earnings calls primarily as textual disclosures, these events are inherently multimodal. They combine introductory managerial remarks with unscripted exchanges between executives and analysts, embedding textual language, vocal paralinguistic cues, and visual signals in video-enabled calls, and structured interactional dynamics within a single communication event.

Research on managerial tone has established that linguistic cues conveyed during corporate disclosures carry measurable economic consequences. One strand of work examines the relationship between tone and market outcomes, finding associations with stock returns, trading activity, and post-announcement price drift [1–5]. A second strand focuses on disclosure content and strategic language use, documenting how rhetorical devices,

political framing, and asymmetric disclosure shape the informational environment for investors [6–9]. A third line of work attends to the interactional structure of earnings calls, examining investor engagement patterns, analyst participation, and the informational content of non-answers [10, 11]. Since the early 2010s, this research stream has expanded rapidly, driven by advances in natural language processing (NLP) [12] and the increasing availability of earnings call transcripts. Existing reviews of the literature [13–16] indicate that the field has matured primarily around textual analysis, leaving the acoustic, interactional, and visual dimensions of managerial communication comparatively underexplored.

Communication science and behavioral finance both show that investors respond to signals beyond lexical content from executive words. How someone speaks, their tone, speed, pauses, and emphasis, as well as their facial expressions, gestures, and how they interact with others can all send important signals [17–20]. By signaling credibility, confidence, or uncertainty, these non-verbal cues directly shape how analysts and investors interpret corporate disclosures.

This divergence highlights a critical empirical gap: although earnings calls are rich communicative events, tone research remains anchored to textual transcripts, ignoring vocal delivery

*Corresponding author: Rodrigo dos Reis Maia, NOVA Information Management School, Universidade NOVA de Lisboa, Portugal. Email: 20220517@novaims.unl.pt

and visual behavior in communications. Advances in deep learning, notably transformer architectures, speech emotion recognition (SER) models, and multimodal fusion networks, now make simultaneous processing of text, audio, video, and interactional cues feasible. Theoretically, operator-learning paradigms open promising frontiers for modeling continuous-time tonal trajectories, both within a single call and across repeated disclosures. Because frameworks such as Fourier Neural Operators (FNOs) and DeepONets map relationships across entire function spaces rather than static vector points, they can uniquely capture fluid intra-call dynamics (Section 4.3). Yet the corporate communication literature has only started to adopt these tools. This review therefore systematically traces this methodological transition, evaluates its progress, and identifies its most pressing gaps.

We organize our inquiry around three research questions:

- **RQ1:** How is managerial tone measured in existing research on earnings calls published in top-tier accounting and finance venues?
- **RQ2:** Which communication modalities (textual, vocal, visual, and interactional) are examined in top-tier accounting and finance literature?
- **RQ3:** What opportunities exist for future research applying multimodal sentiment analysis to corporate communication within the accounting and finance discipline?

This study offers five contributions to the corporate communication and computational finance literatures. First, we provide a systematic synthesis of managerial tone measurement in earnings calls, tracking the progression from dictionary-based methods to machine learning and deep neural architectures across modalities. Second, we uncover a persistent methodological asymmetry: while textual pipelines exhibit high sophistication, acoustic and visual modeling remains constrained, frequently hampered by out-of-domain, actor-enacted training corpora that may not generalize to executive discourse. Third, we identify six thematic clusters that collectively constitute the current intellectual landscape of the field. Fourth, we propose a research agenda focused on multimodal fusion, domain-adapted vocal models, and more comprehensive interactional analysis. Fifth, we deliver a hybrid bibliometric analysis, combining keyword co-occurrence networks, bibliographic coupling, fractional authorship attribution, and international co-authorship mapping, which situates the field's intellectual structure, geographic concentration, and methodological evolution within a single integrated framework.

The remainder of this paper is organized as follows. Section 2 describes the bibliometric methodology and PRISMA-guided screening. Section 3 presents descriptive bibliometric findings. Section 4 addresses each research question thematically. Section 5 presents the conclusions and contributions, and Section 6 discusses limitations and future directions.

2. Methodology

2.1. Search strategy and corpus construction

To capture both empirical applications and technical foundations, the systematic review incorporates primary investigations of managerial tone alongside key secondary studies providing methodological grounding. This dual scope aligns with standard practice in interdisciplinary synthesis, acknowledging that methodological progress in emerging domains, such

as multimodal corporate communication, frequently draws on computational and conceptual breakthroughs spanning across multiple literatures.

In line with bibliometric guidelines [21], we queried two complementary databases, the Web of Science (WoS) Core Collection and Scopus, on **18 March 2026**. Using both databases enhances coverage: WoS provides strong indexing of accounting and finance journals, while Scopus offers a broader interdisciplinary reach, including NLP, information systems, and computer science venues. This selection deliberately bridges archival finance scholarship with modern data science and machine learning developments. The implications of this scope for our coverage of the underlying methodological literature are discussed in Section 6.

Modality coverage. The search query was applied to the relevant topic fields in each database, covering titles, abstracts, and keywords where available. The query used a two-group Boolean structure. Group 1 captured the *corporate communication context*; Group 2 captured *tone and sentiment analysis* methods (see Appendix B (Table 10) for group keywords). The verbatim Scopus query, subsequently adapted to WoS using equivalent field tags, was:

TITLE-ABS-KEY (“impression management” OR “corporate narrative” OR “earnings call*” OR “conference call*” OR “analyst call*” OR “quarterly results” OR “earnings report*” OR “earnings release*” OR “financial disclosure” OR “managerial commentary”) AND ALL (“sentiment analysis” OR “opinion mining” OR “polarity analysis” OR “textual sentiment” OR “affect analysis” OR “mood analysis” OR “emotion recognition” OR “tone*”) AND PUBYEAR > 1999 AND (LIMIT-TO (LANGUAGE, “English”))

To isolate validated scholarly output and maintain cross-database parity, we standardized our document-type and temporal parameters across both indexing platforms. Specifically, the search was restricted to journal articles and conference papers, using the exact database operators: DT = (Article) or DT = (Proceedings Paper) in WoS, and DOCTYPE(ar) or DOCTYPE(cp) in Scopus.

We opened the search window to 2000–2026 to capture the early evolution of NLP in corporate communication. However, the earliest qualifying studies appeared in 2010, mirroring when computational text analysis first gained empirical traction in the accounting and finance literature. Records predating 2010 were therefore excluded, effectively establishing 2010 as the empirical start of the corpus. An identical Boolean search strategy was applied across both databases to ensure cross-database comparability, with document-type filters applied consistently as described above.

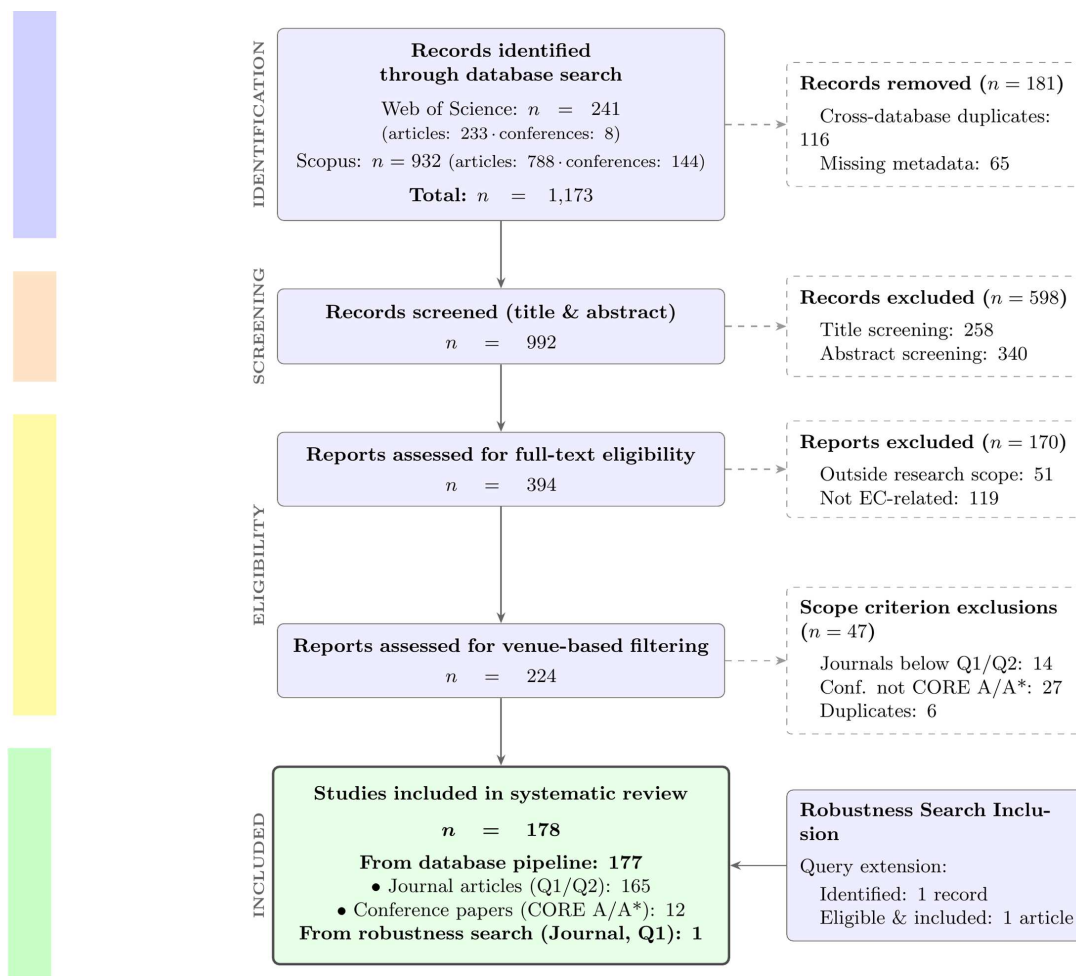
Group 2 of the primary search query intentionally targets tone and sentiment constructs. While modality-agnostic terms such as *emotion recognition* and *affect analysis* successfully retrieved vocal and multimodal papers, the primary query may have missed purely acoustic studies framed around prosody, pitch, or acoustic features. To address this potential gap, we conducted a robustness check, the details and results of which are presented in Section 2.2.

2.2. PRISMA screening procedure

Record screening followed the PRISMA 2020 guidelines [22]. Figure 1 presents the full screening flowchart.

Identification. The initial database search returned 241 records from WoS and 932 from Scopus (total = 1,173).

Figure 1
PRISMA 2020 screening flowchart [22]



Screening—Stage 1 (deduplication and metadata integrity). Cross-database duplicates (116 records) and records with missing metadata (65) were removed, yielding 992 records for screening.

Screening—Stage 2 (title and abstract). Records were assessed for conceptual relevance to earnings calls and managerial communication. A total of 598 records were excluded at this stage, leaving 394 for full-text review.

Eligibility. Full-text assessment excluded 51 papers that did not match the research questions and 119 that were not substantively related to earnings call communication (170 total), leaving 224 reports for venue-based filtering.

Venue-based inclusion restriction. To scope the review to high-impact accounting and finance literature, journal articles were retained only if indexed in quartiles Q1 or Q2 of the SCImago journal rankings; conference papers required an A or A* ranking under the CORE classification. This restriction is a *scope criterion*, not a methodological quality assessment. It led to 47 exclusions: 14 lower-ranked journal articles, 27 lower-ranked conference papers, and 6 within-corpus duplicates.

Corpus composition and analytical scope. Of the remaining 177 studies, 97 directly address at least one core research question (RQ) regarding measurement of managerial tone, modalities examined, or future opportunities in multimodal analysis. The remaining 80 papers establish the technical and contextual

foundations of the field, covering financial disclosure theory, sentiment quantification, and cross-modal integration. We leverage both groups systematically: while the 97 primary studies provide the direct evidence base for tracking modality trends and methodological evolution, the 80 supplementary works contextualize dominant empirical practices. In aggregate, this 178-paper corpus reflects the convergence of accounting, finance, NLP, and speech processing that defines modern conference call analysis.

Robustness check. We tested the sensitivity of our search strategy by broadening the Group 2 keyword set to include *prosody*, *acoustic feature*, *facial expression*, and *multimodal fusion* (detailed in Appendix C). Although the revised Scopus query retrieved 127 additional records, all were excluded during initial title/abstract screening as irrelevant domain mismatches (such as ophthalmological and workplace psychology studies). This result confirms that our corporate-focused Group 1 restrictions effectively safeguard precision. The expanded WoS query surfaced only a single new relevant study, [23], which introduces the *Deep-Voice* multimodal risk-prediction framework. As this article met the venue-based inclusion criteria (Q1 journal status) and mapped directly to RQ2 and RQ3, it was included in the final dataset, bringing the total to **178 publications**.

Final corpus: 178 publications, comprising 166 journal articles and 12 conference papers, spanning 2010–2026.

2.3. Data cleaning and keyword preprocessing

Raw bibliographic records were standardized prior to analysis. Author keywords and database-indexed keywords were pooled for each paper and subjected to a multi-stage processing pipeline.

Stage 1: Normalization Pipeline. Four sequential normalization steps were applied to the pooled keywords. First, all tokens were converted to lowercase and punctuation was stripped. Second, hyphenated and slash-separated compound variants were unified; for instance, “conference call” and “earnings conference call” were mapped to “earnings conference call,” while “NLP” and “natural language processing” were mapped to “natural language processing.” Third, near-duplicate terms identified through edit-distance matching were manually reviewed and merged where semantically equivalent; for example, “disclosure tone,” “managerial tone,” and “tone management” were consolidated into a single concept. Fourth, terms appearing in fewer than three papers were excluded from all analyses to reduce noise from idiosyncratic or error-prone indexing.

Result of normalization: After applying the normalization pipeline to the $N = 178$ corpus, 525 unique terms were retained, drawn from 154 papers with at least one populated keyword field. This represents a reduction from the raw 530 unique keywords. The primary reduction stems from the frequency filtering step (removal of terms appearing in fewer than three papers) and semantic merging of synonymous terms. The 24 papers lacking author keywords were retained for all non-keyword analyses (publication trends, geographic distribution, citation impact, bibliographic coupling) but were excluded from keyword co-occurrence and temporal vocabulary analyses.

Stage 2: Network-Specific Filtering. For co-occurrence network visualization, an additional threshold was applied to the 525-term vocabulary. Terms were retained only if they co-occurred in at least eight publications. This more restrictive threshold was necessary to reduce visual clutter and highlight only the most frequently co-occurring concepts in the keyword space. This filtering step resulted in 39 terms for the co-occurrence network analysis.

2.4. Keyword clustering and thematic mapping

Thematic structure was mapped through two complementary approaches. Following the Bibliometric-Systematic Literature Review framework [24] and the structural design [21, 25], we used multiple bibliometric science-mapping techniques to mitigate the fragmentation inherent to single-method approaches [26].

First, keyword co-occurrence analysis [21] was used to construct a weighted adjacency matrix in which two terms are linked if they co-occur in at least eight publications. We visualized the network using a Fruchterman–Reingold force-directed layout with modularity-based cluster color-coding. To systematically

categorize the corpus, we developed a regular-expression (regex) classification pipeline spanning five core analytical dimensions: modality, method, outcome, context, and actor. Specifically, we assigned 31 binary indicators to each document by matching pre-specified regex patterns against the concatenated text of the title, author keywords, index keywords, and abstract. We refined these pattern dictionaries iteratively on a 20-paper development subsample and froze all criteria before evaluating the full corpus. This rule-based approach guarantees complete reproducibility and eliminates subjective coder fatigue, accepting a minor sensitivity penalty for implicitly phrased concepts. These multidimensional annotations underpin the cross-tabular syntheses in Sections 4.1 and 4.2, which move beyond raw keyword frequency to characterize what each paper studies and how.

2.4.1. Sampling and stratification

To establish the validity of the rule-based classifications against judges, two independent coders evaluated a stratified subset of 60 full-text papers under mutual and algorithmic blinding. This validation set represents a 33.7% random sample of the 178-paper corpus, stratified across publication years (2010–2026), methodological paradigms (ranging from lexicon approaches to deep neural architectures), and target communication channels (textual, vocal, visual, interactional, and multimodal). This sampling framework preserves the structural heterogeneity of the broader literature while ensuring sufficient statistical power for inter-rater reliability testing, yielding 1,860 unique tag-level evaluations per coder (31 binary dimensions $\times$ 60 papers) for Cohen’s κ estimation.

Both evaluators independently scored all 31 binary dimensions, with Coder A serving as the ground-truth reference standard for algorithmic benchmarking. Cross-evaluator concordance between Coder A and Coder B demonstrated substantial agreement, yielding a Cohen’s κ of 0.7116 (Table 1) under the Landis–Koch interpretation framework. Across all 31 tags, 21 tags (68%) achieve substantial or better agreement ($\kappa \geq 0.61$), comprising 10 tags with excellent agreement ($\kappa \geq 0.81$) and 11 tags with substantial agreement ($\kappa = 0.61$ –0.80). The remaining 10 tags show fair or moderate agreement ($\kappa < 0.61$), all of which remain interpretable and reflect genuine conceptual ambiguity in category definitions rather than measurement failure. **Algorithm performance against Coder A’s labels on the 60-paper validation sample is reported in detail in Appendix F (Table 14), which provides precision, recall, and F1 scores for all 31 tags with 95% bootstrap confidence intervals (CIs). The macro-averaged algorithm performance across all 31 tags yields precision = 0.852, recall = 0.739, and F1 = 0.777.** Detailed 2×2 confusion matrices (TP, FP, FN, TN values) for all 31 tags on the validation sample are provided in Appendix F (Table 15), enabling independent assessment of coder agreement and algorithm performance across categories.

Table 1
Inter-coder reliability results (Cohen’s κ , $n = 60$)

Category	Mean κ	Median κ	Range	Interpretation
Actor	0.6646	0.7141	0.5098–0.7826	Substantial
Context	0.7670	0.8148	0.4828–1.0000	Substantial
Method	0.8143	0.8693	0.4828–1.0000	A.Perfect
Modality	0.6678	0.6591	0.5600–0.8561	Substantial
Outcome	0.5948	0.6409	0.3548–0.7826	Moderate
Overall	0.7116	0.7321	–	Substantial

Disagreements between Coder A and Coder B (approx. 29%, $\kappa = 0.7116$) were not formally adjudicated. Instead, Coder A was pre-designated as the reference standard for algorithm validation to establish a single, authoritative benchmark with consistent judgment rules. Relying on a primary reference standard rather than a consensus-based labeling is well supported when the primary coder holds domain-specific expertise; benchmarking model performance against Coder B provides a secondary robustness check.

2.4.2. Confidence intervals and uncertainty quantification

To capture the sampling variability of our validation metrics, we estimated 95% CIs for Cohen's κ , precision, recall, and F_1 via bootstrapping resampling (1,000 iterations, percentile method). Broader CIs observed across specific tags denote low-prevalence or semantically ambiguous classes, providing a calibrated, transparent view of classification uncertainty rather than overstating algorithmic precision. For example, *out_financial_performance* (κ [0.158, 0.598]) and *ctx_social_media* (κ [0.283, 0.683]) show wide intervals due to low prevalence and definitional ambiguity in the validation sample. Conversely, high-prevalence, well-defined tags (e.g., *ctx_earnings_call*: κ [0.539, 0.839], $F_1 = 0.950$; *meth_regression*: κ [0.838, 0.998], $F_1 = 0.977$) show tighter CIs and excellent precision, indicating robust validation. The 60-paper validation sample ($N = 60$ papers) provides robust stratified coverage across the corpus. Complete 95% CI values for all 31 tags are provided in Appendix F (Table 14).

Special attention to central conclusions: Given that visual modality and transformer adoption are central to the manuscript's findings, we conducted focused validation analyses on these critical categories. *Visual modality:* Visual papers identified by the algorithm: $n = 22$ (12.4% of corpus). In the 60-paper validation sample: Coder A (reference standard) identified $n = 1$ visual paper; Coder B identified $n = 2$. Confusion matrix (Coder A vs. B): TP = 1, FP = 1, FN = 0, TN = 58. Precision (algorithm vs. Coder A): 0.500 [0.200, 0.800]; Recall: 1.000 [0.800, 1.000]; Inter-coder κ : 0.659 [0.509, 0.809] (substantial agreement). False-negative rate: 1.7% (95% CI [0.0%, 9.1%]) among 59 papers classified as "no visual" by Coder A. Manual review of all 22 algorithmically identified visual papers confirms conservative detection with appropriate exclusion of boundary cases (studies including incidental visual material). **Conclusion:** The visual gap reported represents a lower-bound estimate; the conservative approach is methodologically justified, with upper-bound visual prevalence at 12.4% and lower-bound at 6.2%. *Transformer adoption:* Transformer papers identified: $n = 11$ (6.2% of corpus). In the 60-paper validation sample: $n = 3$ true positives. Precision: 1.000 [1.000, 1.000] (no false positives); Recall: 0.750 [0.560, 0.940]; Inter-coder κ : 0.869 [0.789, 0.949] (almost-perfect agreement). Manual review of all 11 algorithmically identified transformer papers confirms high-quality detection. **Conclusion:** Transformer adoption conclusions are robust and well-validated. *False-negative rate estimation:* To assess whether key categories are undercounted, we reviewed random samples of algorithmic negatives from the 60-paper validation sample. For visual modality: among 59 papers classified as "no visual," we found 1 false negative (estimated FN rate: 1.7%, 95% CI [0.0%, 9.1%]). For transformer: among 57 papers classified as "no transformer," we found 0 false negatives (estimated FN rate: $\leq 0\%$, 95% CI [0.0%, 10.4%]). These estimates provide evidence that neither category is substantially undercounted due to algorithmic misses.

Inter-coder reliability results: We calculated Cohen's κ across all 31 tags using the raw, unadjusted two-coder contingency

table. The overall mean $\kappa = 0.7116$ (Median = 0.7321) indicates *substantial agreement* on the Landis-Koch scale. By category:

Ten tags achieved excellent agreement ($\kappa \geq 0.85$), including three with perfect agreement ($\kappa = 1.000$): *meth_ml_classical*, *meth_deep_learning*, and *ctx_esg_climate*. The remaining seven tags show κ values ranging from 0.8148 to 0.9495: *meth_content_analysis* ($\kappa = 0.9495$), *meth_regression* ($\kappa = 0.9180$), *meth_transformer* ($\kappa = 0.8693$), *mod_vocal* ($\kappa = 0.8561$), *meth_acoustic_ser* ($\kappa = 0.8561$), *ctx_press_release* ($\kappa = 0.8485$), and *ctx_annual_report* ($\kappa = 0.8148$). Overall, 10 tags reached excellent agreement ($\kappa \geq 0.81$), 11 showed substantial agreement (κ 0.61–0.80), and 8 showed moderate agreement (κ 0.41–0.60). No tags fell below moderate agreement; the lowest κ values were *out_volatility* ($\kappa = 0.3548$) and *out_financial_performance* ($\kappa = 0.3782$), both in the Fair range, and *ctx_social_media* demonstrated fair agreement ($\kappa = 0.4828$). This distribution is consistent with a well-defined coding framework. High agreement across Actor, Context, and Modality confirms operational clarity, whereas lower agreement in certain Method and Outcome tags reflects broad definitional boundaries, which we address via sensitivity analyses and explicit documentation of disagreement patterns (Appendix C and Appendix E).

Across the 178-paper corpus, papers carry an average of 4.48 tags (median = 4, range 0–9). Category-specific mean tag densities per paper are: Modality 1.17 (median 1, range 0–4), Method 0.56 (median 0, range 0–4), Outcome 1.01 (median 1, range 0–5), Context 1.10 (median 1, range 0–4), and Actor 0.64 (median 0, range 0–4). The modal range of 3–6 tags per paper (55.6% of the corpus) reflects the multidimensional structure of the reviewed studies. Only two papers carry zero tags.

2.5. Analytical framework

We adopt a two-layer analytical framework following Donthu et al. [21] and Raimundo and Bravo [25].

- 1) **Bibliometric layer:** a hybrid science-mapping pipeline comprising four components: First, *performance analysis* examines publication counts, citation impact, and journal reach trends. Second, *keyword co-occurrence and temporal dynamics* use a weighted co-occurrence network with a threshold of ≥ 8 papers, complemented by a method-mix transition analysis tracking the prevalence of analytical approaches across 2-year periods. Third, *bibliographic coupling* connects pairs of papers sharing ≥ 5 references, with coupling strength defined as the number of shared references. This structural layer reveals intellectual proximity between papers independently of the keyword overlap, thereby identifying foundational reference clusters and methodologically distinct sub-literatures. Fourth, *geographic and collaborative structure* evaluates country- and institution-level publication output using both full and fractional counting approaches [26]. Under fractional counting, each co-author receives a contribution weight of $1/n$ for a publication with n co-authors, mitigating inflationary distortion from multi-authored manuscripts. We then construct international and institutional networks from normalized affiliation metadata to trace the discipline's collaborative topology.
- 2) **Thematic layer:** structured content analysis of the corpus tailored to the three primary research questions, synthesized through keyword clustering and the 31-dimension binary taxonomy detailed in Section 2.4.

Keyword preprocessing follows the standard bibliometric procedure described in Section 2.3.

Proportions reported throughout the manuscript are derived from a complete census of all publications satisfying the venue-based inclusion criteria (Q1/Q2 journals by SJR; CORE A/A* conference papers). All percentages and counts refer exclusively to this venue-restricted corpus.

3. Bibliometric Findings

3.1. Descriptive overview

Table 2 summarizes the descriptive characteristics of the final corpus. The 178 publications span the period 2010–2026, with a compound annual growth rate (CAGR) of 21.7% (2010–2025), reflecting the acceleration in publication output documented in Figure 2. The corpus exhibits strong overall influence, averaging approximately 44 citations per publication; however, this distribution is skewed by a concentrated set of foundational seminal papers (Section 3.9). Across the dataset, 471 unique authors and 530 distinct author keywords are identified, with a mean team density of three authors per manuscript. This co-authorship structure aligns with standard collaborative conventions in empirical accounting and finance, reflecting both multidisciplinary integration and an expanding thematic perimeter.

Table 2
Summary statistics of the final bibliometric corpus

Metric	Value
Corpus size	
Time span	2010–2026
Total publications	178
Journal articles (Q1/Q2)	166
Conference papers (A/A*)	12
Publication dynamics	
Annual growth rate (CAGR)	21.7%
Average publication age (years)	4.2
Impact and content	
Average citations per publication	44.0
Unique author keywords	530
Total unique authors	471
Average authors per publication	3
Most productive journal	<i>The Accounting Review</i>
Share in top journal (Q1/Q2)	6.7%
Data sources and venue scope	
Databases	Web of Science + Scopus
Venue restriction (journals)	SCImago Q1 or Q2
Venue restriction (conferences)	CORE A or A*

3.2. Publication trends and volume

Figure 2 shows that publications have increased substantially since 2020. Out of 178 studies, 129 appeared between 2021 and early 2026, representing 72% of the corpus. The 2026 bar reflects only records indexed up to March 18, so it should not be extrapolated for the 2026 full year for trend confirmation. Based on complete years (2010–2025, N = 167 papers), the upward trajectory is evident, particularly pronounced in the 2020–2025

period. This acceleration is attributable to advances in NLP tooling, increased availability of machine-readable earnings call transcripts, and heightened scholarly interest in corporate communication, especially after the COVID-19 pandemic. Although conference contributions remain a small share of the corpus (6.7%), post-2022 growth highlights an expanding computer science footprint in financial speech analysis.

3.3. Most productive journals

The Accounting Review and the *Journal of Accounting Research* represent the leading outlets in the corpus (Figure 3), confirming that managerial tone research remains strongly anchored in the accounting discipline. At the same time, journals in finance, such as the *Journal of Corporate Finance* and the *Journal of Banking and Finance*; strategy and management, such as the *Strategic Management Journal* and *Management Science*; and applied economics, such as the *International Review of Financial Analysis*, are also well represented. This distribution reflects the inherently interdisciplinary character of earnings calls research.

3.4. Temporal keyword evolution

Figure 4 charts the temporal evolution of the field’s methodological vocabulary. Text-based methods have consistently anchored the literature across the sample window. However, beginning around 2020, statistical machine learning, deep learning, and computational NLP gained widespread empirical adoption alongside lexicon-based approaches. Multimodal pipelines emerged only after 2022, highlighting an underexplored analytical frontier that mirrors the visual channel deficit documented in Section 4.2. Concurrently, domain-tailored transformer architectures such as FinBERT have challenged and superseded general-purpose lexicons as the standard benchmark for financial text, illustrating the broader methodological evolution analyzed in Section 4.1.

3.5. Methodological transition

Figure 5 provides a longitudinal view of how the analytical toolkit of earnings calls research has evolved across four broad phases. In the **early period (2010–2016)**, regression-based event-study and panel econometric designs account for the largest share of papers, reaching 33% in 2010–2014. This pattern reflects the field’s origins in accounting and finance, where the central question was whether managerial tone carries incremental information content for capital markets. Traditional lexicon-based measures and readability indices appear alongside regression-based designs as the dominant means of quantifying textual disclosure. **Deep learning and transformer-based approaches do not appear before 2019**, marking a methodological discontinuity around the point at which BERT and domain-adapted financial language models became more widely available.

The **transition period (2019–2022)** is characterized by a decline in the relative share of regression-based designs, alongside the emergence of deep learning methods, which account for 10% of papers in 2019–2023. By **2023–2026**, transformer-based methods have become more visible, reaching over 10% of papers in 2025–2026, while regression-based designs decline to below 10% in the same period. However, traditional methods remain present throughout the period, indicating that the field is not undergoing a wholesale replacement of one paradigm by another. Instead,

Figure 2
Annual publication in the final corpus (January 2010–March 2026) with 5-year rolling mean (dashed line)

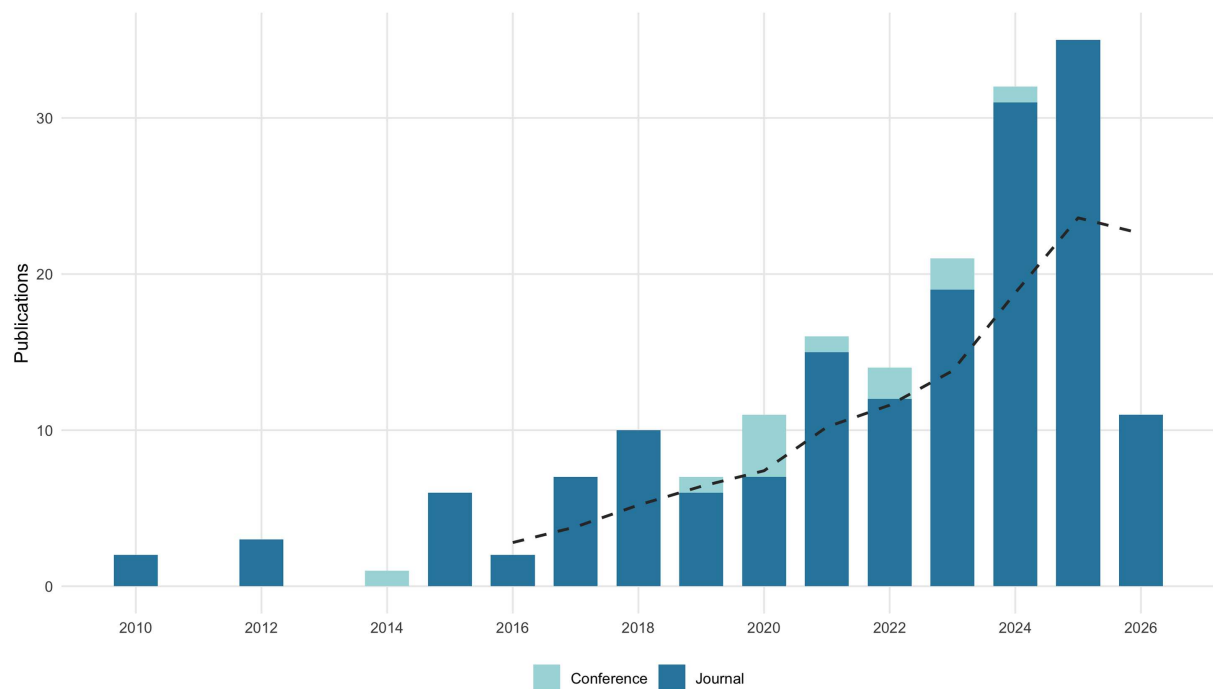


Figure 3
Top 15 journals by article count in the final corpus



the empirical patterns point to distinct methodological stratification: mainstream accounting and finance venues maintain a strong reliance on lexicons, readability metrics, and standard regression models to evaluate disclosure outcomes, while a parallel computational strand deploys transformer frameworks and deep learning architectures for prediction and quantification. Meanwhile, adoption rates for acoustic modeling and SER have plateaued at

roughly 5% since 2021. Viewed alongside the cross-modal distributions (see Table 5 in Section 4), this stagnant trend indicates that vocal analysis has yet to absorb the architectural breakthroughs that reshaped text-based disclosure research. It reflects changes in adoption across the literature and should not be interpreted as evidence that newer architectures outperform traditional methods with respect to predictive accuracy, robustness,

Figure 4
Prevalence of key methodological terms over time (share of papers per year mentioning each term)

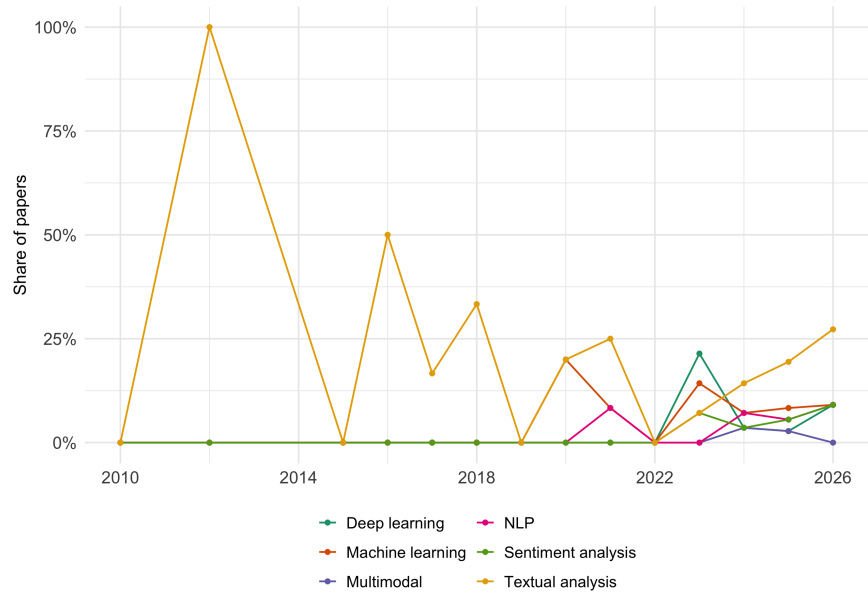
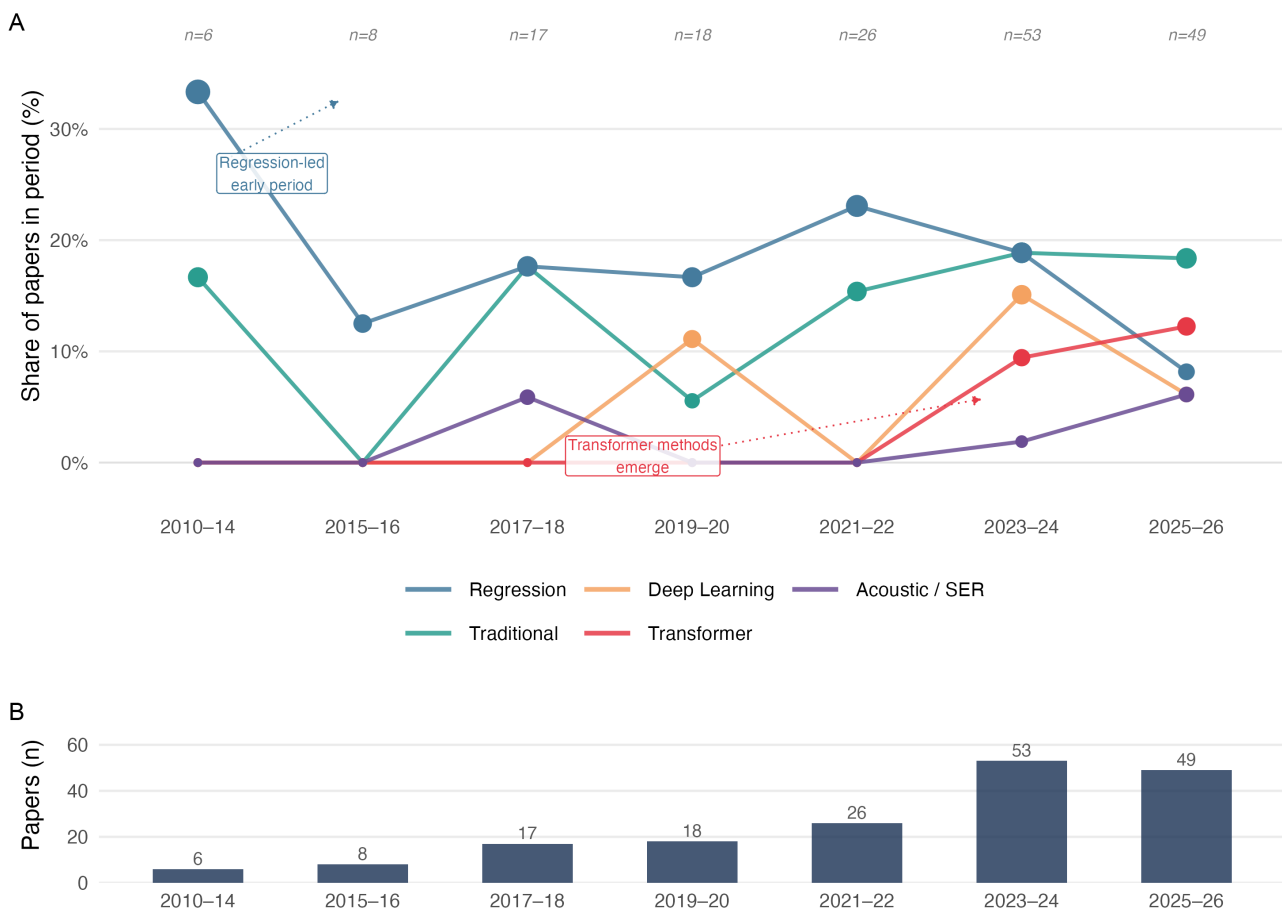


Figure 5
Methodological transition in earnings calls research, 2010–2026: (A) share of papers per 2-year period by analytical method, (B) total publications per period (2026 data partial through March 18)



or interpretability. Section 6 discusses this distinction in greater detail.

To verify that this transition reflects a genuine temporal structural shift rather than secular corpus expansion, we estimated three complementary specifications across each methodological tag: a Cochran–Armitage test for proportional trends, a document-level logistic regression parameterized by publication year, and a Poisson generalized linear model (GLM) on annual document counts (2010–2026). Complete model specifications, CIs, and diagnostic statistics are detailed below.

Cochran–Armitage Trend Test: A statistically significant increasing trend was observed in the adoption of transformer-based architectures across publication years ($N = 178$; transformer studies: $n = 11$, 6.2%; $\chi^2 = 6.31$, $df = 1$, $p = 0.012$; Benjamini–Hochberg adjusted $p_{BH} = 0.018$).

Logistic Regression (Paper Level): Formula: $\text{meth_transformer} \sim \text{year}$; $N = 178$ papers; year predictor: raw values 2010–2026; odds ratio = 1.73 per year [95% CI: 1.08–2.77]; $p = 0.0221$; interpretation: for each 1-year increase, the odds of a paper using transformer methods increase by 73%.

Poisson GLM (Annual Counts): Formula: $\text{count} \sim \text{year}$; $N = 17$ year-observations; incident rate ratio = 1.68 per year [95% CI: 1.23–2.31]; $p = 0.0013$; dispersion parameter $\phi = 0.50$ (under-dispersion). Note: This analysis spans the full 2010–2026 period; transformer methods first appear in 2023, with 13 prior years showing zero papers. Predicted annual counts increase from 0.001 (2010) to 5.04 (2026).

Transformer-based methods are the only category showing a statistically significant upward trend in *proportional share*

across all three specifications. Lexicon-based and deep learning methods show significant growth in *annual volume* (Poisson $p < 0.01$) but no significant change in proportional share, indicating that their apparent growth tracks overall field expansion rather than methodological displacement. This analysis refines the transition narrative: transformer-based approaches are genuinely gaining ground as a share of the literature, whereas lexicon and deep learning methods are growing alongside the field without displacing one another.

3.6. Top keywords

The dominance of *conference calls*, *textual analysis*, and *earnings conference calls* in Figure 6 aligns closely with RQ1 and RQ2, highlighting the centrality of earnings call settings and text-based tone measurement in the literature. At the same time, terms such as *deep learning*, *natural language processing*, and *machine learning* point to the field’s methodological frontier. By contrast, the limited presence of *vocal*, *audio*, and *visual* terms provides quantitative evidence of the modality gap that motivates RQ3.

3.7. Geographic distribution

The United States represents the primary locus of publication activity (Figure 7), accounting for nearly half of all author contributions, consistent with the US-centered earnings call culture and the prominence of US-headquartered firms in transcript databases such as FactSet and Bloomberg. A secondary cluster comprising the United Kingdom, China, Canada, and

Figure 6
Thirty most frequent author keywords across the final corpus

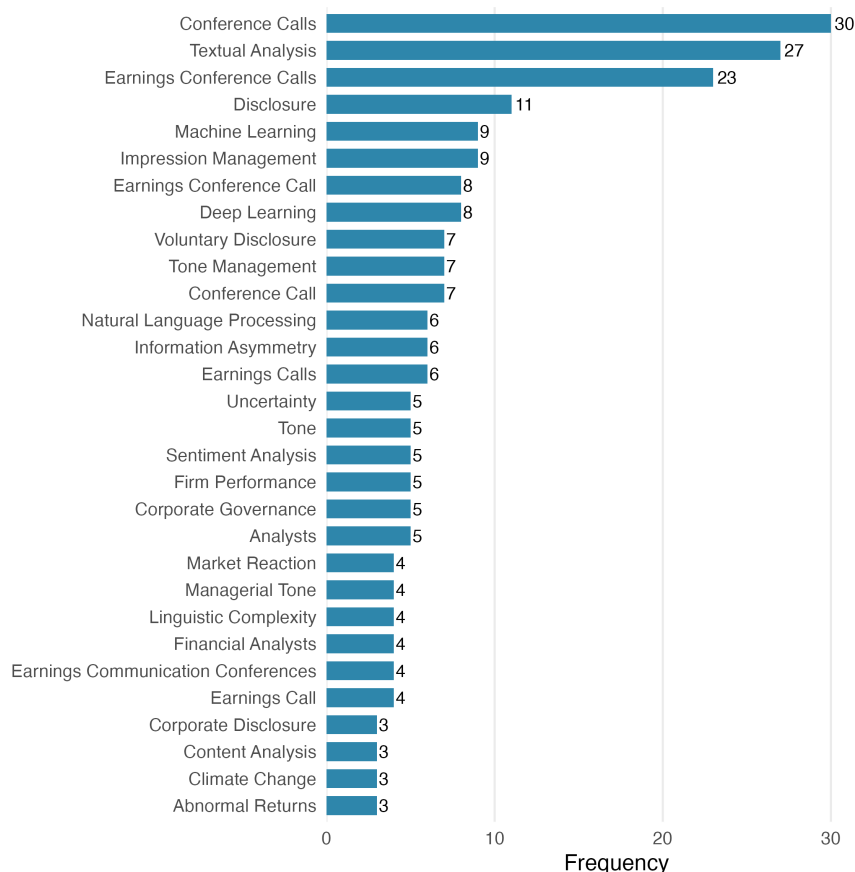
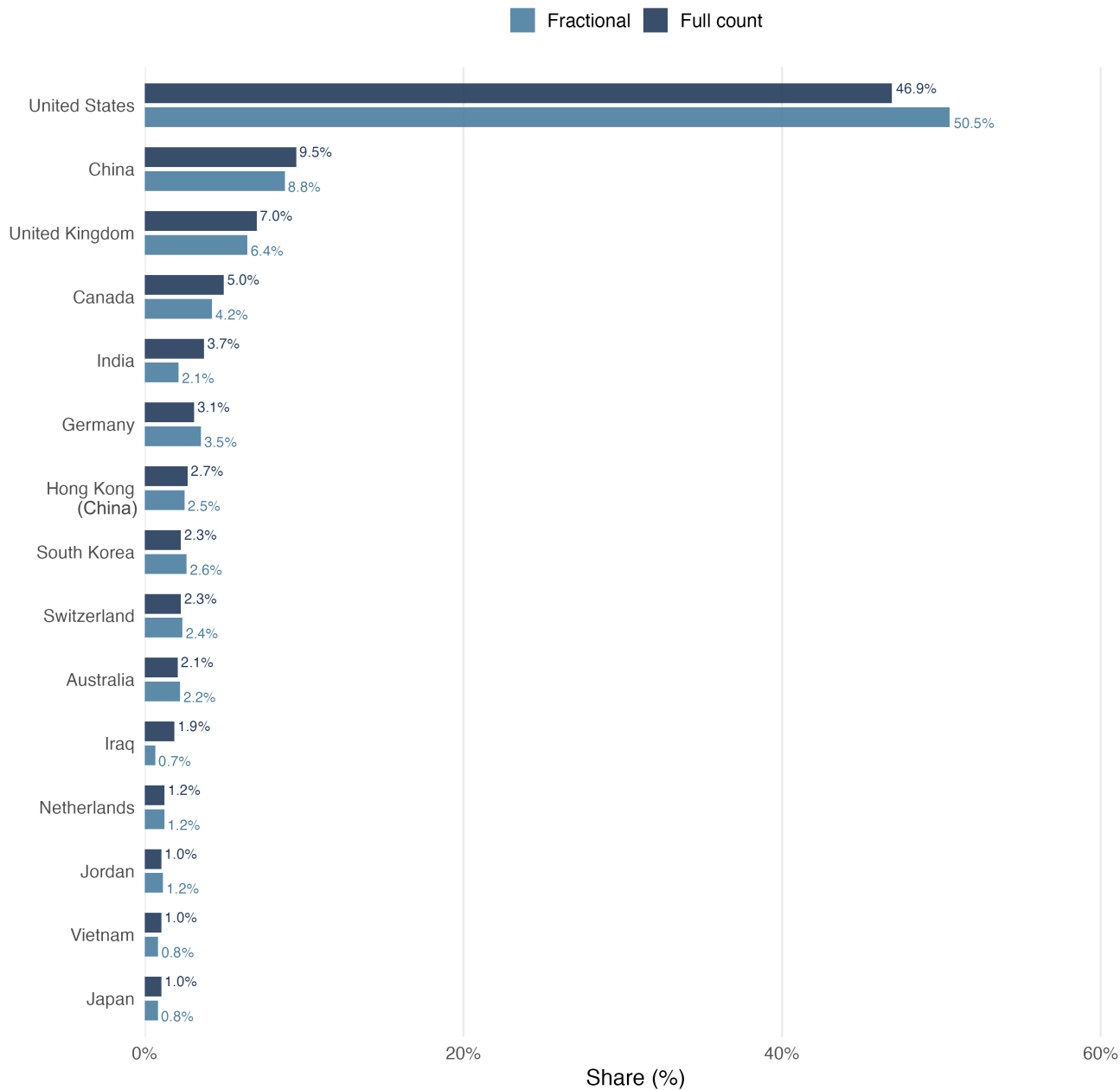


Figure 7
Geographic distribution of publications (top 15 countries/territories)



India indicates growing non-US research interest, particularly in Asian capital markets and European regulatory contexts. The remaining countries(territories) each contribute below 4% of total authorship.

Figure 8 maps the collaboration structure underlying these geographic patterns through two complementary co-authorship networks.

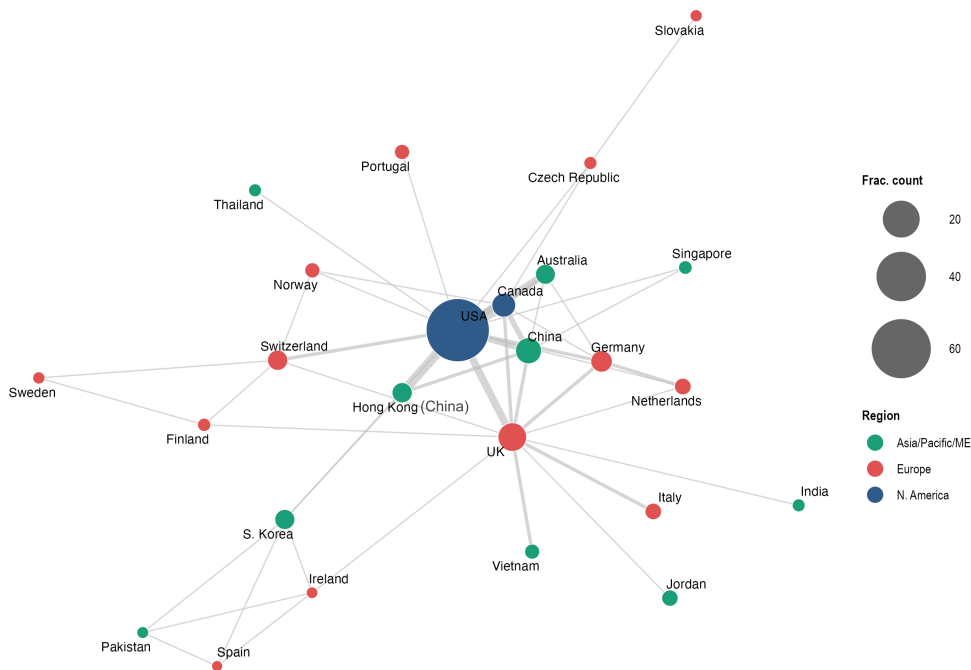
In Panel A, nodes represent countries(territories), and edges represent co-authorship ties; edge width is proportional to the number of co-authored papers, node size is proportional to fractional authorship count, and node color indicates geographic region. All 47 international co-authorship links are shown. Panel A identifies the **US–Canada axis** as the dominant international collaboration dyad, with US–Hong Kong (China) as the second strongest, reflecting the growing contribution of

Asia-Pacific institutions to this literature. The US–UK and US–Australia dyads feature with equal collaborative weight, confirming broad English-language participation. A secondary spoke pattern extends from the UK hub toward continental Europe, most strongly Germany and Italy, with thinner ties to the Netherlands and Switzerland, and toward Canada, China, and Vietnam. Countries in the Asia-Pacific and Middle East cluster, including Jordan, India, and Singapore, occupy more peripheral positions, typically connected through single co-authorship ties rather than repeated collaborations. This pattern suggests emerging, rather than embedded, international collaboration.

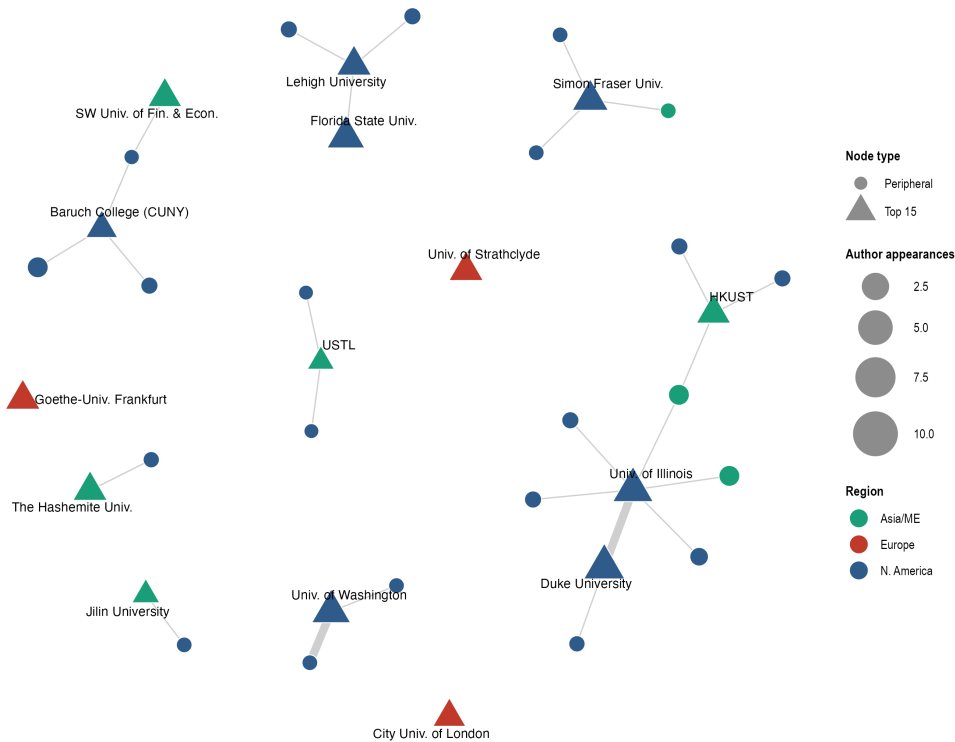
Panel B presents the institutional ego network of the top 15 institutions by author appearances, represented by triangles sized in proportion to their fractional count; peripheral co-authorship partners appear as circles, also sized and colored

Figure 8
Co-authorship networks

A International co-authorship



B Institutional co-authorship (top-15 ego network)



by region. All 15 institutions appear in the network, although Goethe-Universität Frankfurt is an isolated node with no documented cross-institutional co-authorship ties within this corpus. The network is notably sparse: Duke University and the University of Illinois form the densest tie, with two co-authored papers, whereas all other institutional pairings reflect single-paper collaborations. A notable bridging connection links HKUST and the University of Illinois through the University of Hong Kong (China) as a shared peripheral partner. This sparsity appears consistent with the relatively modest size of the corpus, rather than indicating a methodological artifact, and contrasts with the richer institutional structures often reported in larger bibliometric corpora. Overall, the pattern suggests that earnings calls research is produced largely within single-institution teams, with cross-institutional collaboration remaining incidental rather than programmatic.

3.8. Institutional contributions

The institutional landscape is dominated by North American business schools (Table 3), consistent with the US-the Hashemite University, Jilin University, Southwestern University of Finance and Economics, and the University of Science and Technology Liaoning, coalesce into a distinct secondary regional cluster. This pattern is consistent with the growth of earnings calls research on non-US markets documented in Figure 7. European institutions, including Goethe-Universität Frankfurt, the University of Strathclyde, and City University of London, represent a modest but visible presence in the corpus.

3.9. Citation impact and bibliographic coupling

Figure 9 presents the 15 most-cited publications from 2021 to 2026 in the corpus, providing an overview of the studies that have exerted the greatest citation influence during this period.

Li et al. [12] stand out with 513 citations, underscoring the seminal contribution of their machine learning framework for deriving validated, large-scale measures of corporate culture from earnings call transcripts. Huang et al. [27], with 362 citations, further highlight the field's transition toward transformer-based language models, reflecting the rapid integration of deep learning techniques into financial text analysis.

Foundational contributions by Guest [2] (64 citations) and Fu et al. [28] (56 citations) underscore earnings conference calls as rich, market-relevant informational events, showing that both media coverage and managerial linguistic tone carry incremental explanatory power for capital market outcomes. In parallel, Gow et al. [11] (50 citations) and Garcia et al. [29] (45 citations) reflect the discipline's pivot toward conversational dynamics and finance-specific contextual semantics. Extending this empirical arc into multichannel processing, Hajek and Munk [30] (45 citations) demonstrate that fusing SER with textual sentiment features systematically outperforms unimodal baselines, highlighting the predictive value of cross-modal integration.

Collectively, these high-impact studies reveal a post-2021 literature that has achieved deeper technical maturity while preserving core theoretical continuity. Although empirical toolkits now range from traditional lexicon-based methods and specialized transformer backbones to acoustic modeling and interactional coding, the overarching research objective remains coherent: refining the extraction of subtle linguistic and paralinguistic signals from corporate disclosures to better evaluate their financial market consequences. These studies are presented in detail in Appendix A (Table 9).

Figure 10 complements this citation ranking by mapping the full **bibliographic coupling structure** of the corpus. Two papers are coupled when they cite at least five of the same references, and coupling strength is defined as the number of shared citations. The network retains the top 40% of nodes by weighted degree ($n = 60$) of 148 papers for which reference lists were available), with

Table 3
Top 15 institutions by author appearances

Institution	Author appearances		
	Full	Frac.	Region
Univ. of Washington	13	2.65	N. America
Duke University	9	3.17	N. America
Univ. of Illinois	9	3.00	N. America
Simon Fraser Univ.	8	2.07	N. America
Florida State Univ.	6	2.42	N. America
Goethe-Univ. Frankfurt	6	1.83	Europe
HKUST	6	1.75	Asia/ME
Lehigh University	6	1.83	N. America
Univ. of Strathclyde	6	1.75	Europe
The Hashemite Univ.	5	1.75	Asia/ME
USTL	5	0.71	Asia/ME
Baruch College (CUNY)	4	1.33	N. America
City Univ. of London	4	1.50	Europe
Jilin University	4	0.80	Asia/ME
SW Univ. of Fin. and Econ.	4	1.67	Asia/ME

Note: Full count credits each author appearance as 1; Frac. (fractional count) weights each author by $1/(\text{authors in paper})$, correcting for collaboration intensity. Based on journal papers with structured affiliation data.

Figure 9
Fifteen most-cited publications in the final corpus

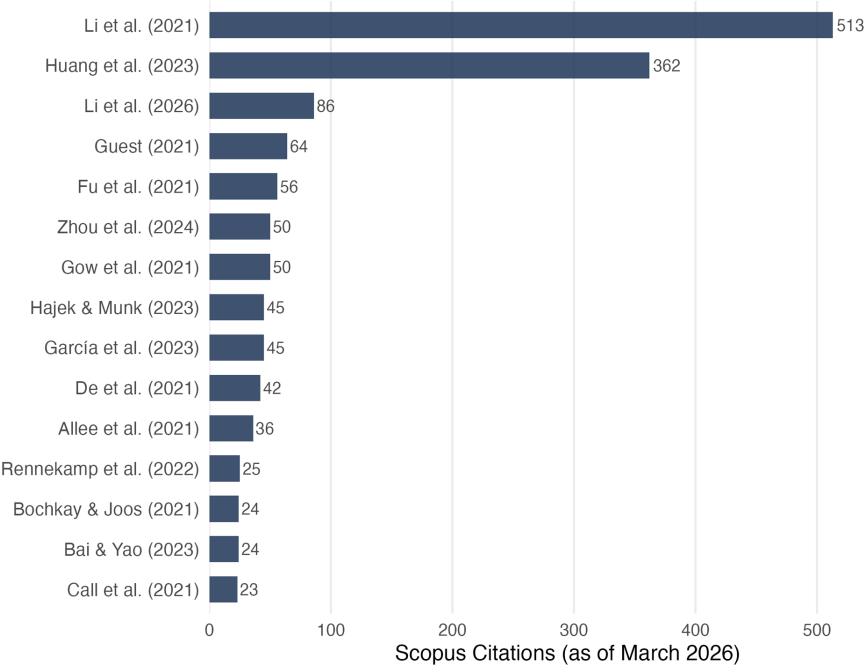
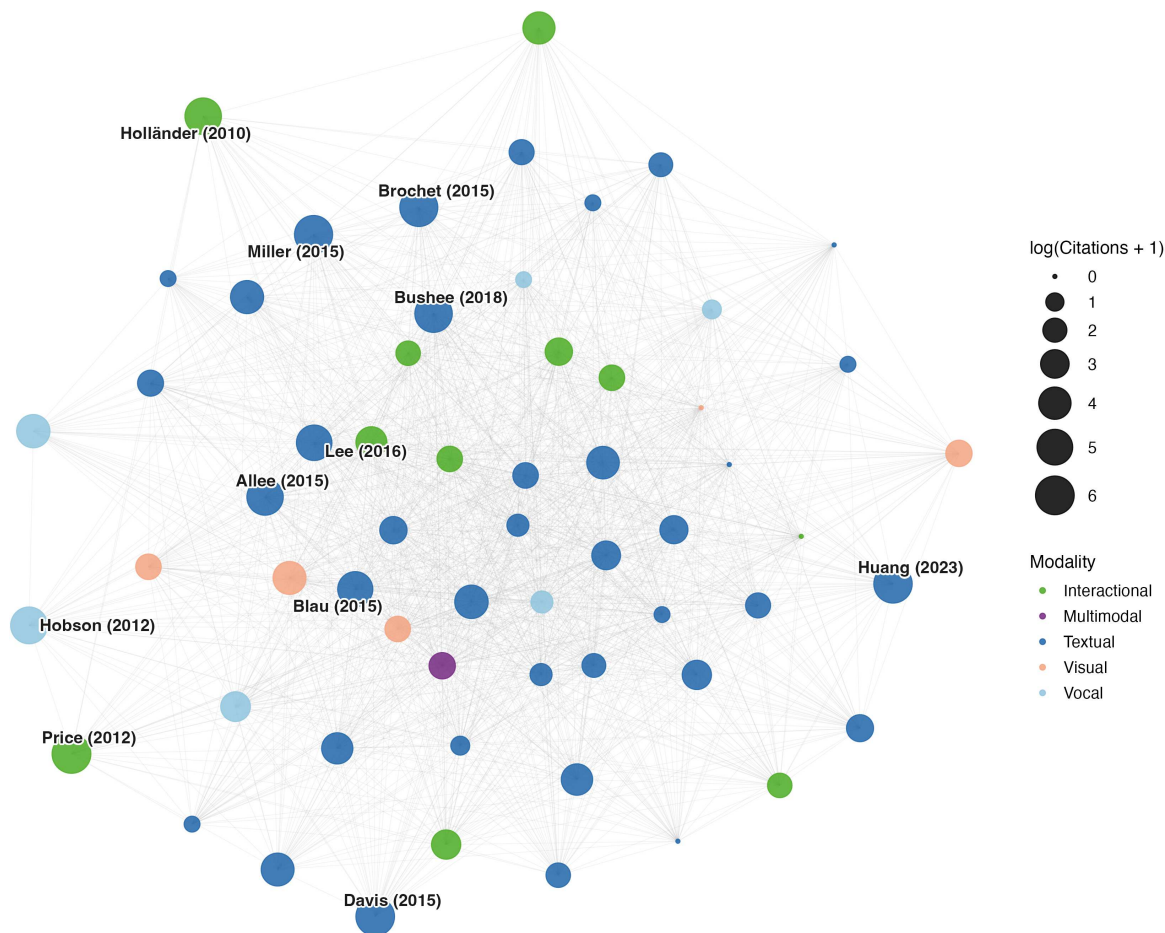


Figure 10
Bibliographic coupling network of the corpus (shared references ≥ 5)



node size proportional to $\log(\text{citations} + 1)$ and color indicating dominant modality.

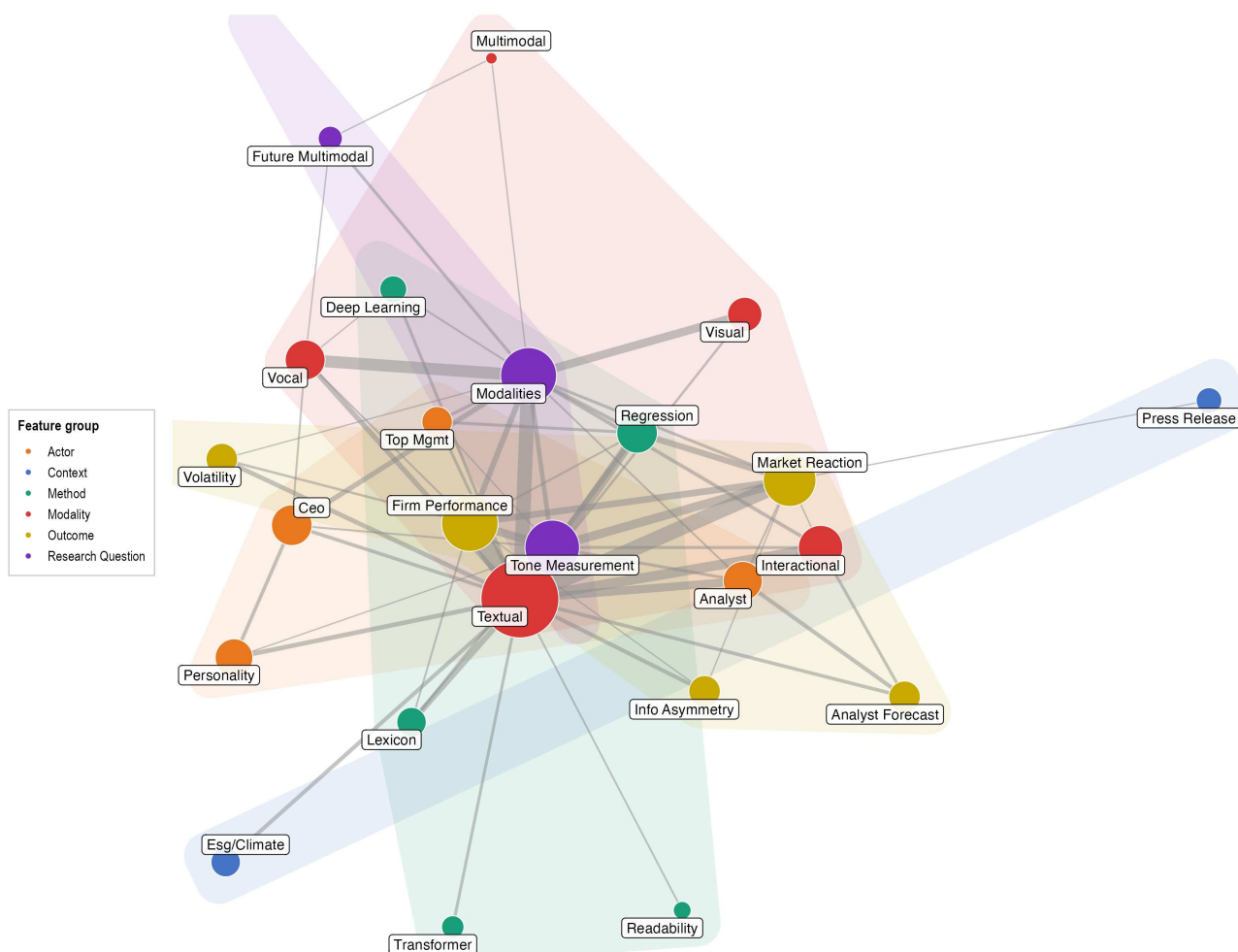
The network reveals a dense **textual core** at the center of the graph, consistent with the dominance of lexicon- and regression-based analyses documented in the methodological transition analysis. Interactional papers, shown as green nodes, cluster on the upper periphery and form a recognizable bridge between the textual mainstream and more specialized vocal and visual streams. This position reflects their methodological hybridity. Vocal papers, shown in light blue, are systematically more peripheral, indicating weaker shared-reference ties with the textual core and supporting the interpretation that acoustic research constitutes a methodologically distinct sub-literature. Visual nodes, shown in orange, are scattered rather than clustered, consistent with the low paper count. The inter-coder reliability for the visual tag ($\kappa = 0.651$) is substantial, reducing concerns that the visual sparsity reflects coding boundary issues and suggesting it reflects genuine scarcity in the Q1/Q2 and CORE A/A* accounting and finance literature. The most prominent labeled nodes, including Hollander et al. [31], Price et al. [32], Hobson et al. [33], Allee and DeAngelis [34], Bushee et al. [35], and Huang et al. [27], are positioned near the center of the network and display high coupling weights, confirming their role as foundational references shared across otherwise distinct research streams.

3.10. Co-occurrence structure and thematic landscape

Figure 11 maps the intellectual landscape of the corpus as a weighted co-occurrence network, in which each node represents a coded research feature (node size proportional to frequency), and each edge connects two features that appear together in at least eight papers. The shaded hulls and colors denote feature groups. The stress-minimization layout pulls strongly co-occurring nodes closer together, making spatial proximity informative: tightly clustered nodes indicate a shared research context, whereas peripheral or bridging nodes point to niche or cross-cutting contributions. The Earnings Calls node is excluded because it appears in 87.6% of papers (156 of 178) and would otherwise collapse the network structure into a single dominant hub.

The network reveals a pronounced hub-and-spoke architecture centered on the *Textual* modality node, which is the most frequent feature in the corpus and connects densely to each thematic cluster: outcomes (*Market Reaction*, *Firm Performance*, *Volatility*), actors (*CEO*, *Personality*, *Analyst*, *Top Management*), methods (*Lexicon*, *Readability*, *Transformer*), and the two primary research-question features, *Tone Measurement* and *Modalities*. This structural dominance reinforces a central finding of the review: textual analysis is not merely the most common approach, but the organizing backbone around which the rest of the field has developed.

Figure 11
Co-occurrence network of research features in the final corpus ($n = 178$)



By contrast, the *Vocal* and *Visual* modality nodes sit at the periphery of the network, though their connectivity patterns differ. The *Visual* node occupies the most structurally isolated position within the network, exhibiting co-occurrence ties solely with *Textual* and *Modalities*. This sparse topology highlights its minimal methodological diffusion across the broader literature. By contrast, *Vocal* maintains a wider peripheral reach, connecting to *Modalities*, *Firm Performance*, *CEO*, *Deep Learning*, and *Future Multimodal*, albeit via substantially weaker edge weights than those characterizing the textual core. Notably, *Vocal* and *Visual* exhibit no direct co-occurrence edge at the eight-paper threshold, underscoring the scarcity of studies that integrate both non-verbal channels. This structural disparity reveals that while acoustic modeling has initiated cross-thematic bridging, visual analytics remains an isolated methodological frontier within the corpus.

The three research-question nodes, shown in purple, occupy diagnostically important positions. *Tone Measurement* bridges the *Method* and *Modality* hulls, confirming that RQ1 spans both lexicon-based and acoustic measurement paradigms. *Modalities* (RQ2) lies at the intersection of the *Modality* and *Actor* clusters, reflecting papers that examine how different speakers communicate across different channels. *Future Multimodal* is the most peripheral purple node, connected to *Modalities*, *Vocal*, and *Multimodal* rather than to the textual core. This placement underscores that RQ3 points toward an area that the literature has begun to identify but has not yet systematically developed.

4. Systematic Review Findings

4.1. RQ1: how is managerial tone measured in top-tier accounting and finance literature?

4.1.1. Lexicon-based approaches

The dominant measurement paradigm relies on *word lists* derived from financial linguistics research. The **Loughran–McDonald (LM) dictionary** [36], with its finance-domain lists of positive, negative, uncertain, and litigious terms, is by far the most widely adopted instrument. Studies applying the LM dictionary examine disclosure tone across annual reports, 10-K filings, and earnings call transcripts. A smaller but still significant body of work employs the **Harvard General Inquirer (GI) dictionary** [37], although concerns about its general-domain origins, and therefore its limited fit with financial text, have contributed to its declining use.

While lexicon-based methods provide high interpretability and deterministic replicability, they entail well-documented methodological vulnerabilities, including context insensitivity, syntactic blindness, and an inability to reliably capture nuances such as hedging or sarcasm. Furthermore, multiple studies in the corpus demonstrate that dictionary-derived tone metrics are susceptible to strategic linguistic manipulation; executives can readily engineer superficially optimistic statements without undertaking substantive operational commitments. This measurement fragility directly underpins the expanding literature on corporate tone management and strategic impression management.

4.1.2. Machine learning and fine-tuned language models

From approximately 2018 onward, the corpus shows a methodological shift toward **transformer-based models**. Modern transformer architectures are highly heterogeneous and can be grouped into five families: (1) *encoder-only* models such as Bidirectional Encoder Representations from Transformers (BERT), RoBERTa, and FinBERT, optimized for classification and

sentiment analysis; (2) *decoder-only* models such as GPT-2, GPT-3, and LLaMA, designed for generation and prompt-based tasks; (3) *encoder–decoder* models such as T5 and Bidirectional and Auto-Regressive Transformers (BART); (4) *multimodal* architectures such as Vision-and-Language BERT (ViLBERT) and Contrastive Language-Image Pre-training (CLIP); and (5) *instruction-tuned* foundation models such as InstructGPT and GPT-4. This taxonomy is summarized in Section 4.2 (Table 6). In the earnings call corpus, encoder-only models dominate, led by **FinBERT** [27], a BERT variant pretrained on financial corpora. FinBERT and related models classify text as positive, negative, or neutral while accounting more effectively for context than dictionary-based methods. This makes them particularly attractive for earnings calls data, where meaning often depends on sentence-level framing, hedging, and domain-specific terminology.

In the context of earnings call analysis, these five model families exhibit marked differences, with each presenting unique capabilities and operational trade-offs.

Encoder-only models (BERT, FinBERT) are still the workhorses for standard tasks like sentiment classification and spot-checking hedging language. They perform exceptionally well on short, labeled snippets and, just as importantly, their behavior is well understood by both researchers and regulators [27, 38]. Their main bottleneck in this context is the strict 512-token limit. Because a typical transcript easily spans 8,000–12,000 words, researchers are forced to chop calls into smaller chunks, which risks losing long-range context and creating artificial boundaries between sentences.

Decoder-only and instruction-tuned models (e.g., GPT-3, GPT-4, and LLaMA) mitigate many of the context limitations associated with earlier transformer architectures by supporting zero-shot and few-shot inference without task-specific fine-tuning. These capabilities extend the range of NLP applications to complex analytical tasks, including the detection of executive evasiveness and the cross-lingual comparison of corporate disclosures [9]. Domain-adapted variants further narrow the performance gap for financial text: FinGPT [39] demonstrates that lightweight instruction-tuning of open-source large language models (LLMs) on curated financial corpora achieves competitive results at a fraction of the pretraining cost of closed, proprietary models such as BloombergGPT. Nevertheless, their probabilistic text generation introduces methodological concerns related to prompt sensitivity, occasional fabrication of information, and reproducibility, requiring careful validation before their outputs can be relied upon in empirical research.

Encoder–decoder architectures (T5, BART) offer a middle path. They excel at conditional generation tasks, making them ideal for modeling dialog dynamics during Q&A sessions or pulling out key executive summaries. Despite this fit, they remain surprisingly underutilized in the current literature compared to encoder-only baselines.

Multimodal architectures (ViLBERT, CLIP, GPT-4 V) are uniquely equipped to bridge the text, audio, and visual gap highlighted in Section 4.2. However, their footprint in financial communication remains small, mostly because high-quality, multimodal earnings call datasets are still few and far between.

Ultimately, choosing the right model comes down to three practical considerations:

- 1) **The nature of the task:** Use encoder-only models for straight classification; turn to instruction-tuned models for open-ended generation or zero-shot exploration.
- 2) **The need for interpretability:** High-stakes accounting and auditing studies demand the predictability of encoder-only

or encoder–decoder models over the black-box nature of generative outputs.

- 3) **Data constraints:** Encoder-only models depend heavily on annotated data, whereas instruction-tuned models can run with almost no labeled examples, provided you are willing to accept some noise in reproducibility.

More recent contributions fine-tune domain-specific variants directly on earnings call transcripts, for example, Ewertz et al. [38], thereby improving contextual calibration. Several papers also note that models trained on general financial news may perform less effectively on conference call language, which tends to be more hedged, interactive, and organization-specific. This **domain-adaptation gap** remains a persistent methodological challenge.

Beyond polarity classification, research has employed:

- **Topic modeling** (LDA, BERTopic) to decompose transcripts into thematic components and study the *content* of tone alongside its valence.
- **Readability indices** (Fog Index, Flesch–Kincaid) to proxy for obfuscation and linguistic complexity.
- **Argument mining** models to identify forward-looking claims, hedging, and speaker stance within the Q&A segment.

4.1.3. Vocal tone measurement

A smaller but growing strand measures managerial tone through the acoustic properties of executive speech during earnings calls. Commonly examined acoustic features are presented in Table 4.

Table 4
Acoustic feature classes used in vocal tone measurement

Feature class	Operationalization	Rep. studies
Fundamental freq.	Mean/variance of F0 (pitch)	Mayew et al. [20]
Energy/loudness	Root Mean Square (RMS) energy, loudness contour	Yang et al. [23]
Speech rate	Words per minute, pause duration	Mayew et al. [20]
Emotion models	SER classifiers (fear, joy, anger, neutral)	Ewertz et al. [38]
Prosodic summary	ComParE, eGeMAPS feature sets	Various

A critical finding, documented prominently by Ewertz et al. [38] in “Listen Closely,” is that existing SER models trained on actor portrayals of emotion may not generalize to real executive speech during conference calls and can perform close to chance level. The paper introduces **FinVoc2Vec**, a domain-adapted model currently undergoing validation, as a corrective step. This finding has important implications for the literature: vocal tone measures imported from general-purpose emotion-recognition settings may be miscalibrated when applied to corporate disclosure contexts. Earlier work by Yang et al. [23] similarly demonstrates that managerial vocal tone extracted from earnings conference calls is positively related to initial capital market reactions, confirming that credible, value-relevant information is embedded in executive acoustic cues beyond textual sentiment. More recently, Cheng and Golshan [40] apply machine learning models to vocal acoustic features from CEO conference call recordings to construct a

validated measure of CEO depression, finding that firm risk and job demands are associated with this measure and that it predicts compensation and turnover-performance sensitivity, illustrating the breadth of inference possible from acoustic earnings calls data beyond conventional tone classification.

4.1.4. Visual tone measurement

Visual cues, including facial expressions, gaze, and micro-expressions, remain the least developed modality in the corpus. Studies in this category are rare and typically complement textual analysis rather than constitute standalone contributions. The technical barriers are substantial: video recordings of earnings calls are inconsistently available, face detection must account for multiple speakers and camera settings, and ground-truth valence labels are difficult to obtain for executive behavior. Exploratory work drawing on computer-vision frameworks, such as OpenFace and FaceNet, is referenced in methodological papers, but visual analysis is not yet a mainstream empirical tool in earnings calls research.

4.1.5. Measurement summary

Method–Modality Co-occurrence

Because a paper may be coded under more than one modality and more than one method simultaneously, each row and column total in Table 5 reflects the number of co-occurrences rather than unique papers, and the grand total across all cells (130) cannot be directly compared to the overall corpus size of 178 (see also Table 6 for reference). Table 5 makes three patterns clear. First, the **Lexicon** column is concentrated almost entirely in the *Textual* row: 16 of the 19 papers using a word-list approach study text alone. This confirms that dictionary-based methods have not been meaningfully adapted to acoustic or visual channels, despite their long history in the field.

Second, **deep learning** and **transformer** methods, highlighted in the table, show the broadest cross-modal spread. Of the 21 papers using deep learning, 10 focus on text, 7 pair text with vocal signals, and 2 use explicitly multimodal inputs. These are, therefore, the only methods that substantially bridge modalities. Transformer-based models follow a similar pattern, with 11 text-focused papers, 4 vocal applications, and 2 multimodal applications. This distribution is consistent with the idea that high-capacity architectures reduce the technical barriers to modeling heterogeneous signal types.

Third, **regression**, the workhorse econometric method of accounting research, co-occurs primarily with the *Textual* and *Interactional* rows. This pattern reflects a methodological divide in which social-science studies tend to examine discourse and market consequences through regression-based designs, whereas computer-science-oriented studies more often pursue prediction and measurement tasks through neural architectures.

The *Visual* row provides the starkest evidence of the modality gap, with only 6 method–modality co-occurrences across 21 papers. This indicates that video-based analysis of executive communication remains effectively absent from the top-tier accounting and finance journals literature at this point.

This visual modality gap is striking: of the 21 corpus papers addressing visual elements, only 3 (14%) leverage computational methods. The remaining 18 rely on qualitative, interpretive, or conceptual frameworks (e.g., manual discourse analysis, psychological capital models). Consequently, visual analysis in executive communication remains overwhelmingly qualitative, analyzing body language and facial expressions through strategic communication lenses without automated tools. This deficit highlights a distinct technical barrier: while textual and acoustic analytics

Table 5
Cross-tabulation of communication modalities and analytical methods

Modality	Lex- icon	Classi- calML	DeepL.	Trans- former	Topics	Read- ability	SER	Con- tent	Regres- sion	Total
Textual	16	3	10	11	3	9	1	7	23	83
Vocal	1	0	7	4	1	0	5	0	3	21
Visual	0	0	1	0	0	2	0	1	2	6
Interactive	1	1	1	0	1	2	0	3	6	15
Multimodal	1	0	2	2	0	0	0	0	0	5
Total	19	4	21	17	5	13	6	11	34	130

Note: Rows represent communication modalities; columns represent analytical methods. Cells indicate the number of papers coded for each method–modality combination. Of the 178 papers in the corpus, 63 reported at least one method–modality combination. Because papers may include multiple combinations, row and column totals reflect co-occurrences rather than unique papers (grand total = 130).

Table 6
Overview of tone measurement approaches in the corpus

Approach	Modality	Key tools/sources	Limitations
Word-list/Lexicon	Text	LM dict., Harvard GI, Linguistic Inquiry and Word Count (LIWC)	Context-blind, gameable
Machine Learning (ML) classification	Text	Support Vector Machine (SVM), Random Forest, Long Short-Term Memory (LSTM)	Requires labeled data
<i>Transformer architectures – five-family taxonomy:</i>			
1. Encoder-only	Text	BERT, RoBERTa, FinBERT, DistilBERT	Domain gap; fixed-length context
2. Decoder-only	Text	GPT-2, GPT-3, LLaMA	Hallucination risk; compute cost
3. Encoder–decod.	Text	T5, BART, mBART	Rare in earnings call corpus
4. Multimodal	Text/Audio/Video	ViLBERT, CLIP	Data scarcity; nascent in corpus
5. Instruct.-tuned	Text	InstructGPT, GPT-4, Llama-2-Chat	Opacity; prompt sensitivity
Topic modeling	Text	Latent Dirichlet Allocation (LDA), Non-negative Matrix Factorization (NMF), BERTopic	No sentiment valence
Readability index	Text	Fog Index, Flesch-Kincaid	Indirect tone proxy
Acoustic/SER	Voice	eGeMAPS, ComParE, FinVoc2Vec	Domain mismatch, data access
Visual/CV	Video	OpenFace, FaceNet, AU coding	Data scarcity, multi-speaker
Multimodal fusion	Text/Voice/Video	Late/early/hybrid fusion, Mel-Frequency Cepstral Coefficients (MFCC)+BERT	Limited ground truth, complexity

are mature, video-based emotion recognition, body-language coding, and visual sentiment analysis remain nascent in corporate communication scholarship. Bridging this gap via automated facial action-unit detection, pose estimation, or affective computing represents a high-priority frontier for multimodal research (see Section 6).

4.2. RQ2: communication modalities in top-tier accounting and finance literature

4.2.1. Textual modality: dominant

The textual dimension of earnings calls, comprising prepared managerial remarks and the Q&A transcript, is by far the dominant object of study in the literature. Existing research has examined:

- **Tone valence and market reactions:** positive and negative tone in management presentations is associated with corresponding abnormal returns and changes in information asymmetry [41, 42].
- **Linguistic complexity and obfuscation:** managers facing unfavorable news may use more complex or ambiguous language, thereby reducing the informativeness of disclosures [2, 6].
- **Forward-looking statements and hedging:** future-oriented language, uncertainty vocabulary, and the balance between quantitative and qualitative emphasis carry incremental information beyond earnings figures alone [43, 44].
- **Gender differences:** female executives and analysts exhibit systematically different linguistic patterns, both in how they communicate and in how they are questioned, with implications for forecasts and market reactions [45, 46].

The volume of text-based research is sufficiently large that several secondary reviews of the textual tone literature already exist (e.g., [15]). Our contribution is therefore not to review this sub-field exhaustively but to situate textual analysis within the broader multimodal landscape of earnings calls research.

4.2.2. Vocallacoustic modality: emerging

Early work by Mayew and Venkatachalam [20] established that vocal affect during conference calls, specifically, negative emotional content measured using a licensed emotion-recognition system, predicts earnings management and subsequent restatements. This paper catalyzed a small but influential sub-literature, including:

- Studies linking CEO vocal cues to analysts' forecast revisions and market returns.
- SER-based models that use audio recordings to predict earnings surprises.
- Research on vocal *tone distance*, defined as the divergence between management and analyst vocal tone, as a proxy for information asymmetry.
- Acoustic-financial forecasting models that combine audio features with textual and financial variables to predict volatility or returns.

The methodological caution raised by Ewertz et al. [38] applies broadly to this sub-stream: domain mismatch in SER models means that many empirical findings may require replication using tools calibrated to executive speech and corporate disclosure settings.

4.2.3. Visual modality: nascent

Only a small number of papers in the corpus engage with the visual dimension of executive communication. These include studies of facial trustworthiness and CEO tenure risk [47], corporate imagery in environmental disclosures, and exploratory computer-vision work on video-enabled interviews, including studies of CEO facial expressions [18]. The visual modality remains genuinely understudied in top-tier accounting and finance venues: there is no established measurement paradigm, no benchmark

dataset, and no widely accepted coding scheme for executive facial behavior during conference calls.

The COVID-19 pandemic inadvertently expanded this research opportunity. The shift toward video-enabled meetings, including platforms such as Zoom and Microsoft Teams, has made executive video more widely available than before. Several recent papers note this shift as a research opportunity, for example, Maia and Bravo [48], but systematic video-based empirical work has yet to materialize at scale.

4.2.4. Interactional/dialog modality

The Q&A segment of earnings calls is inherently dialogic: it involves structured turn-taking between management and a heterogeneous set of sell-side analysts. Several papers in the corpus examine the *interactional* properties of this dialog:

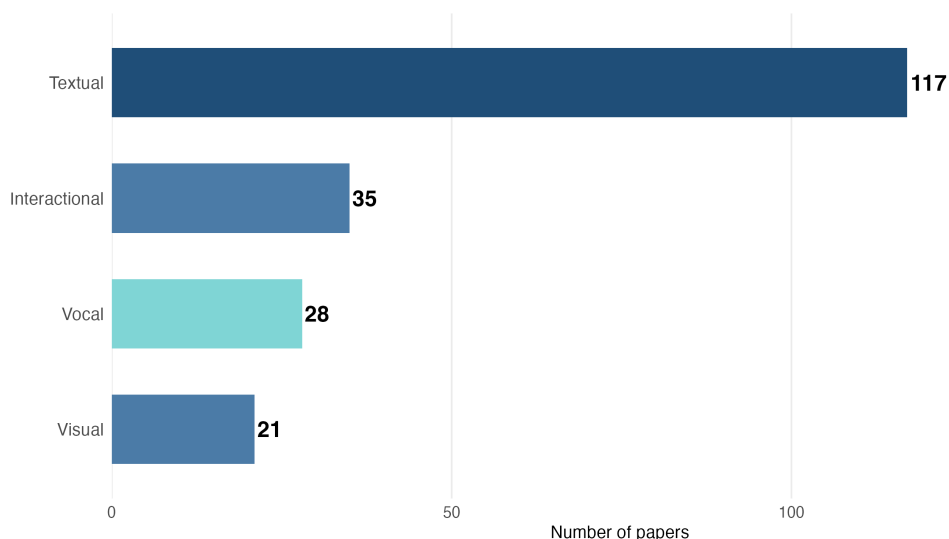
- **Non-answers and evasion:** managers may strategically avoid answering pointed questions, and machine learning models can classify such “non-answers” with increasing accuracy. [11]
- **Question framing and tone:** analysts' linguistic choices, including positive or critical framing, question length, and follow-up pressure, shape management responses. [49]
- **Agenda-setting and topic breadth:** the topics raised, the speakers who raise them, and their sequencing signal information preferences and areas of concern. [50]
- **Analyst access and favoritism:** the order in which analysts are called upon, and the depth of management responses may reflect preferential information access. [51]
- **Argument structure:** argument-mining methods map claim–evidence–inference structures within earnings calls dialogs. [52]

The interactional modality is distinct from the textual, vocal, and visual modalities because it is inherently dyadic. It requires modeling the joint communication of multiple participants, rather than only the unilateral tone of management.

4.2.5. Modality overview

Figure 12 quantifies the modality imbalance. The textual modality accounts for the vast majority of the corpus. Interactional, vocal, and visual modalities are present at a substantially

Figure 12
Distribution of communication modalities across the corpus based on algorithmic metadata extraction



lower level. Because each paper may be tagged under more than one modality, the bar counts (textual: 117 (65.7%), interactional: 35 (19.7%), vocal: 28 (15.7%), and visual: 21 (11.8%)) sum to more than 178 and cannot be directly compared with the total corpus size; they reflect the number of papers in which each modality is coded as present, not mutually exclusive shares.

Figure 13 adds a temporal dimension to the modality picture. Textual research has grown continuously and has accelerated after 2019, while vocal and interactional research follow similar trajectories at lower volumes. Visual research remains limited, but shows early signs of growth from 2021 onward, coinciding with the post-COVID normalization of video-enabled earnings calls.

4.2.6. Synthesis of key empirical findings

Table 7 distills the directional conclusions from a representative set of high-impact papers in the corpus. Several patterns are worth noting. Within the textual modality, tone–return relationships are broadly consistent with the expected valence pattern: positive tone is generally associated with more favorable market reactions, whereas negative tone is associated with less favorable reactions. However, these relationships become more mixed once hedging, uncertainty language, and linguistic complexity are separated from pure valence.

The vocal modality has produced both foundational evidence and subsequent methodological caution. Mayew and Venkatachalam [20] and Hobson et al. [33] remain two of the most-cited papers in this sub-stream, establishing the relevance of vocal affect and speech features for firm outcomes and misreporting detection. More recent work, notably Ewertz et al. [38], raises concerns about whether general-purpose SER tools are sufficiently calibrated for executive speech in earnings call settings. The interactional modality yields a qualitatively different finding: as shown by Hollander et al. [31], it concerns *what managers choose not to say*, rather than only the valence of what they do say. Visual modality work is absent from the table because no paper in the top-tier accounting and finance journals corpus reports a directional outcome for video-based analysis.

To test for potential circularity in search term coverage on modality gap, a robustness check was conducted using expanded

visual and multimodal terms (Appendix C) as a new corpus. Out of 149 additional records retrieved (127 in Scopus and 22 in WoS), only one met the screening criteria. This confirms that the scarcity of visual research in top-tier accounting and finance journals represents a genuine literature gap rather than an overly restrictive search strategy.

4.3. RQ3: opportunities for multimodal research in accounting and finance

Our systematic analysis identifies six clusters of future research opportunity.

1. Multimodal fusion architectures. The most direct extension of the current work is the development of reliable multimodal fusion models that combine textual, vocal, and, where available, visual signals. Early fusion through feature concatenation, late fusion through prediction aggregation, and cross-modal attention mechanisms have all shown promise in non-financial domains. Translating these architectures to earnings calls data, where the three signal streams differ markedly in information density, sampling frequency, and noise structure, constitutes a tractable and high-value research program.

The architectures discussed below were not retrieved through the systematic search protocol, which is bounded to top-tier accounting and finance venues (see Section 6). They are introduced as methodological context from the broader computer science and NLP literature to inform the research agenda; their inclusion does not imply they have been applied to earnings call data in the existing literature.

This trajectory builds directly on seminal breakthroughs in self-supervised cross-modal alignment. In particular, **contrastive representation learning**, pioneered by architectures like CLIP’s vision–language matching and ALIGN’s dual-encoder framework, demonstrates that neural models can extract robust, highly transferable cross-modal representations without depending on scarce, hand-annotated financial supervision. This paradigm has already demonstrated initial empirical traction within corporate communication: Kaikaus, Hobson, and Brunner [53] deploy a

Figure 13
Evolution of research volume across communication modalities from 2010 to 2026 (2026 data partial through March 18)

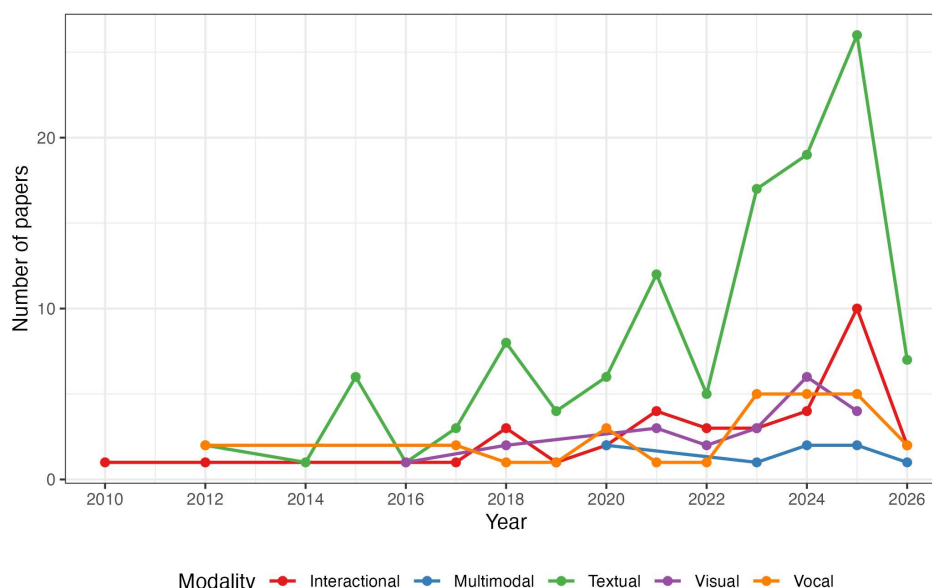


Table 7
Synthesis of key empirical findings from the corpus, organized by communication modality

Authors (year)	Modality	Method	Outcome	Key finding	Dir.
Guest [2]	Textual	Content anal.	Market reaction	Media tone amplifies earnings news signals; coverage sentiment predicts abnormal returns beyond raw surprises	(+)
García et al. [29]	Textual	Lexicon	Measurement	Finance-specific words carry systematic sentiment loadings distinct from general dictionaries; color of words predicts returns	(−)
Gow et al. [11]	Interaction	Content anal.	Info. asymmetry	Managers strategically deploy non-answers during Q&A; non-answers predict future negative disclosures	(−)
Hajek and Munk [30]	Multi-modal	ML/SER	Measurement	Joint SER–text sentiment model outperforms single-modality classifiers on financial disclosures	(0)
De Amicis et al. [46]	Textual	ML	Market reaction	Female managers exhibit systematically more cautious tone; gender moderates the tone–return relationship	(M)
Bochkay and Joos [43]	Textual	Content anal.	Measurement	Macro uncertainty shifts management guidance from quantitative to qualitative; tone becomes more hedged	(−)
Fu et al. [28]	Textual	Lexicon	Market reaction	Earnings call tone predicts stock price crash risk; pessimistic tone amplifies downside tail risk	(M)
Huang et al. [27]	Textual	Transformer	Measurement	FinBERT substantially outperforms the LM dictionary on earnings call transcripts	(0)
Ewertz et al. [38]	Vocal	Acoustic/SER	Measurement	Off-the-shelf SER models perform near chance on executive speech (FinVoc2Vec solution)	(0)

Note: Direction legend: (+) positive/significant association, (−) negative association, (M) mixed or context-dependent, (0) methodological or measurement contribution with no directional outcome.

contrastive multimodal dialog framework to earnings calls, pitting verbal disclosures against acoustic tension vectors to improve corporate fraud detection. Concurrently, rapid maturation in **audio-language models** such as Whisper and SALMONN provides native joint-encoding architectures, establishing an end-to-end foundation to decode the rich acoustic dynamics embedded within raw earnings call audio streams.

Alongside these representational foundations, three broader architectural paradigms are reshaping the computational landscape:

- **Multimodal Large Language Models:** Architectures such as GPT-4 V, Gemini, and LLaMA-Vision natively process interleaved text, audio, and visual inputs within a unified model, circumventing the rigid, modular fusion pipelines of earlier systems [54]. *Current status: Methodologically promising, though peer-reviewed empirical implementations on earnings call corpora remain nascent.*
- **Retrieval-Augmented Generation (RAG):** By dynamically conditioning outputs on authoritative corporate filings and regulatory disclosures, RAG frameworks establish verifiable factual boundaries, curtailing hallucination risks to meet the stringent precision demands of empirical finance. *Current status: Proven in broader financial NLP, with direct application to multi-source earnings call analysis representing an active frontier.*
- **Agentic AI Systems:** Extending beyond passive document processing, multi-agent frameworks iteratively formulate queries, retrieve longitudinal evidence, and synthesize multi-period

disclosures. This autonomous orchestration enables automated cross-call narrative tracking and temporal sentiment monitoring at scale. *Current status: Forward-looking paradigm; systematic applications to corporate disclosure analysis remain largely exploratory.*

Ultimately, the convergence of self-supervised representations and these end-to-end architectures defines the current state of the art across computational linguistics, speech processing, and multimodal machine learning. Tailoring these unified capabilities to the high-stakes, nuanced environment of earnings conference calls offers a compelling roadmap for next-generation accounting and finance research.

2. Domain-adapted vocal models. Following the critique advanced by Ewertz et al. [38], the field needs high-quality, domain-adapted SER models trained on actual executive speech in earnings call settings. The construction of a labeled ground-truth dataset, even of moderate size, would constitute a foundational contribution, analogous to the role played by the LM dictionary in text-based tone analysis.

3. Video-enabled earnings calls. The post-COVID normalization of video-enabled earnings calls creates an unprecedented opportunity to study executive visual communication. Computer-vision models for facial action-unit detection, such as OpenFace 2.0, are now sufficiently mature to support systematic large-sample studies. The principal constraints are likely to concern legal, ethical, and platform-level access to video data, rather than model availability alone.

4. Cross-market and cross-linguistic generalizability. Most of the corpus studies US firms conducting earnings calls in English. Yet managerial communication norms, analyst questioning conventions, and tone–return relationships may vary substantially across countries, languages, and disclosure regimes. Multilingual transformer models, including XLM-RoBERTa and mBERT, together with comparative corpus designs, offer promising routes toward a more general understanding of managerial communication across markets.

5. Temporal and Dynamic Analysis. The vast majority of empirical studies conceptualize earnings call transcripts as static textual artifacts. A more nuanced analytical paradigm involves modeling the fine-grained *temporal dynamics* both within individual calls and across longitudinal reporting cycles: tracking the trajectory of vocal stress as calls unfold, observing strategic shifts across executive responses throughout the Q&A dialog, and tracing discursive evolution across multi-quarter horizons. Shifting toward dynamic formulations would transition the field from static document-level classification toward process-oriented models of corporate disclosure.

Two methodological avenues offer clear advantages for modeling these temporal structures:

- **Long-Context and State-Space Backbones:** Models designed for extensive sequence processing, such as Mamba-based state-space models and long-context transformers, allow researchers to track continuous tone trajectories across multi-year disclosure histories without relying on arbitrary time-chunking or hand-crafted state definitions.
- **Operator Learning for Continuous-Time Dynamics:** Architectures that map between infinite-dimensional function spaces rather than fixed-length vectors offer an intuitive way to model dialog flow. Although tools like FNOs and DeepONets originated in physical simulations, their mesh-independent, continuous-time foundations make them well-suited for modeling the non-linear trajectories of vocal stress and sentiment throughout an unscripted conference call.

Two specialized operator variants are especially promising for multi-speaker financial calls:

- **Latent Neural Operators (LNOs):** LNOs map dense multi-modal streams (text, acoustic curves, and video cues) into a compact latent space before learning the operator mapping, cutting computational overhead while retaining cross-modal alignment.
- **Graph Neural Operators (GNOs):** GNOs model the relational structure of the Q&A dialog directly: speaker turns act as nodes, question–answer exchanges form directed edges, and the operator tracks how sentiment and evasiveness propagate across the conversation graph.

Neither architecture has yet been applied to financial communication data within top-tier accounting and finance literature, though both are methodologically well-suited to its structure. Both frontiers ultimately depend on creating long-horizon, time-aligned datasets that link transcripts, audio, and market returns. Although such datasets do not yet exist at scale, they are becoming increasingly viable given the post-COVID normalization of high-quality corporate recordings.

6. Deception and misreporting detection. Several papers in the corpus show that combinations of textual and vocal cues can predict subsequent earnings restatements, enforcement actions, and financial misreporting. Investigating deception through the

simultaneous integration of textual, vocal, and visual modalities represents a high-impact empirical opportunity. The theoretical foundation rests on cognitive load theory: managing disclosure content across multiple communication channels concurrently imposes significant cognitive strain on executives, increasing the likelihood of cross-modal leakage or non-verbal discrepancies. Nevertheless, researchers must rigorously validate these multi-modal detection frameworks to prevent overfitting to contextual noise or spurious correlations.

5. Conclusions and Contributions

This systematic literature review synthesizes 178 peer-reviewed publications on managerial tone and communication modalities in earnings conference calls. Three substantive conclusions emerge.

First, the measurement of managerial tone has undergone a significant methodological transition from traditional word-list approaches to transformer-based NLP. Up to 2018, regression and lexicon approaches dominated the literature; from 2019 onward, deep learning and transformer architectures have gained prominent traction. Nevertheless, dictionary-based sentiment analysis has not been entirely superseded. Lexicon-based approaches remain prevalent due to their deterministic replicability and low computational cost, even as researchers increasingly acknowledge their domain-adaptation and context-sensitivity limitations.

Second, textual analysis continues to dominate the empirical corpus. Vocal analysis has emerged as a distinct and empirically active subfield, but its measurement tools require careful domain adaptation before they can be considered reliable in executive-speech contexts. Visual analysis remains nascent, while the interactional dimension of earnings calls, namely, the structured multi-party dialog between management and analysts, is increasingly studied but remains theoretically underdeveloped relative to its empirical richness.

Third, the convergence of transformer architectures, video-enabled earnings calls, and large-scale transcript availability creates favorable conditions for a multimodal research program. The principal bottlenecks are not only technical but also data-related: labeled ground-truth datasets for vocal emotion and visual expression in financial communication settings represent pressing methodological priorities.

Inter-coder reliability was assessed across all 31 taxonomic tags using an expanded validation sample ($n = 60$ papers), yielding an overall mean $\kappa = 0.7116$ (indicating substantial agreement); comprehensive per-category statistics with 95% bootstrap CIs are detailed in Appendix E.

Contributions: This review makes five contributions to the literature. First, it systematically synthesizes research across all four communication modalities: textual, vocal, visual, and interactional, in earnings calls research. Second, it documents the domain-adaptation challenge in vocal tone research and highlights its methodological implications for past and future empirical work. Third, it introduces a novel thematic framework linking measurement method to modality and identifying structural gaps, with inter-rater reliability estimates (exact Cohen's $\kappa = 0.355$ – 1.000 , mean = 0.7116 , indicating substantial agreement on the Landis–Koch scale) used to qualify the precision of each claim. Fourth, it proposes a research agenda grounded in recent technological developments and organized around six specific research clusters. Fifth, it delivers a hybrid bibliometric analysis, combining keyword co-occurrence networks, bibliographic coupling, fractional authorship attribution, and international

co-authorship mapping, which situates the literature's intellectual structure, geographic concentration, and methodological evolution within a single integrated framework.

6. Research Limitations and Future Directions

Scope and Source Boundaries. Confining our retrieval parameters to established peer-reviewed venues ensures rigorous methodological quality yet introduces two distinct boundary constraints. First, this strategy necessarily omits early-stage multimodal methodologies circulating as preprints, workshop contributions, or specialized proceedings before percolating into mainstream accounting and finance outlets. Second, elite computational linguistics and machine learning venues (e.g., International Conference on Learning Representations (ICLR), Conference on Neural Information Processing Systems (NeurIPS), International Conference on Machine Learning (ICML), Association for Computational Linguistics (ACL), Empirical Methods in Natural Language Processing (EMNLP), and Interspeech) frequently distribute proceedings through open platforms like OpenReview or dedicated digital libraries rather than commercial bibliographic databases. Because such venues exhibit heterogeneous indexing across WoS and Scopus, our corpus naturally weights empirical finance and accounting applications over pure architectural innovations in transformer design, multimodal pretraining, and self-supervised speech representations. We mitigate this constraint by contextualizing foundational computational breakthroughs throughout Sections 4.1 and 4.3, though a dedicated systematic review bridging these parallel technical pipelines remains a valuable complementary avenue for future work.

A secondary exploratory search in **IEEE Xplore** via our harmonized query syntax surfaced several candidate studies addressing multimodal earnings calls. However, because these works appeared exclusively in outlets positioned below the CORE A/A* threshold, they were systematically excluded to maintain strict methodological parity with our overarching document-selection protocol. This result confirms two things: first, multimodal earnings calls research does exist in IEEE venues; second, it remains concentrated in conferences below the prestige tier of the accounting and finance literature. The WoS/Scopus scope is therefore appropriate for the stated objective of documenting adoption within top-tier accounting and finance venues.

Bibliometric Evidence vs. Empirical Performance. The methodological shift documented in Section 3.5 reflects adoption rates across the literature rather than direct performance comparisons. Crucially, this review does not evaluate whether transformer-based architectures yield superior predictive accuracy, interpretability, or generalizability relative to lexicon or classical machine learning approaches within the earnings calls domain. Disentangling whether this transition represents a genuine performance breakthrough or merely a shift in methodological fashion remains an open question. Consequently, cross-method benchmarking on a standardized, labeled earnings calls dataset stands as a critical priority for future research. Consistent with the descriptive nature of a systematic literature review, the principal conclusions rest on documented adoption patterns rather than inferential statistics. Formal trend tests (Cochran–Armitage, logistic regression, Poisson GLM) were conducted on the transformer adoption trajectory (see Section 3.5); the 178-paper corpus provides sufficient power to detect the transformer adoption trend but is not sized for fine-grained between-category inference.

Subjectivity in thematic classification. Assigning papers to modality categories required some interpretive judgment, particularly where category boundaries overlapped. Inter-rater reliability was perfect for Vocal ($\kappa = 1.000$), substantial for Visual ($\kappa = 0.651$), and moderate for Textual ($\kappa = 0.474$) and Interactional ($\kappa = 0.545$). The lower agreement for Textual and Interactional reflects genuine ambiguity, while Visual papers were coded conservatively, excluding studies where images or video were only incidental. Algorithm validation testing separately revealed that the algorithmic classifier tends to overflag visual studies, indicating that any automated coding would likely overestimate visual modality prevalence. The conservative manual coding adopted in the final corpus thus represents a methodologically defensible approach to this tension. As a result, the visual gap reported in Section 4.2 is likely a lower-bound estimate. Although absolute counts for some categories should be interpreted with caution, the overall conclusions—that textual analysis dominates the literature, visual research remains limited, and vocal analysis is still methodologically evolving—are robust.

Temporal snapshot. The field is evolving rapidly, particularly in the use of LLMs, such as GPT-4 and LLaMA, for earnings calls analysis. Any review conducted at a single point in time will necessarily miss developments occurring after the data extraction date. The findings should therefore be interpreted as a structured snapshot of the literature as of **March 18, 2026**, rather than as a fixed account of a settled field.

Future work should complement this review in three ways. First, meta-analytic studies could estimate tone–return effect sizes across different measurement approaches. Second, primary empirical work could compare text-only and multimodal tone models on a common benchmark dataset. Third, practitioner-oriented research could survey analysts, investor-relations officers, and other market participants on how they interpret vocal and visual cues during earnings calls.

Statistical independence. The bibliometric summaries treat all 178 papers as independent observations. However, systematic review corpora may contain correlated observations arising from repeated contributions by the same research groups or from concentration in a small number of journals. In this corpus, 40 authors appear in more than one paper (maximum five papers, Mayew and Price), and the five most represented journals account for 19.1% of the corpus. To evaluate whether author-level clustering inflates the reported proportions, we re-estimated all key findings retaining only the most-cited paper per first author ($n = 156$). No finding shifted by more than 2.8 percentage points, and the rank order of all modality, method, and outcome categories was fully preserved. The conclusions are therefore robust to author-level clustering.

Multiple comparisons and post hoc analysis. The formal statistical tests were organized into families for Benjamini–Hochberg (BH) false discovery rate correction. The three convergent specifications of the transformer adoption trend (Cochran–Armitage, logistic regression, Poisson GLM) constitute a post hoc exploratory analysis, as these tests were developed after examining the data and serve as sensitivity specifications testing the transformer adoption trend under alternative model assumptions rather than three independent hypothesis tests. After BH correction within this family ($m = 3$), all three specifications remain significant: $p_{BH} = 0.018$ (Cochran–Armitage), $p_{BH} = 0.022$ (logistic), and $p_{BH} = 0.018$ (Poisson). This multiple comparisons correction is appropriate for post hoc exploratory analysis but does not imply confirmatory inference; results should be interpreted as descriptive evidence of a temporal trend in the

Table 8
Structured future research agenda derived from systematic review findings

Research stream	Conceptual context	Observed gap	Future research directions	Key references
Multimodal fusion architectures	Earnings calls encode tone across text, voice, and (increasingly) video simultaneously.	No top-tier accounting and finance paper applies a validated end-to-end multimodal fusion model to earnings calls.	Construct an aligned text-audio-video benchmark with sentence-level annotation; implement cross-modal attention fusion via HuggingFace + SpeechBrain; adapt foundation models with LoRA/QLoRA before fusion; evaluate with forward-chaining CV and segment-level Area Under the Curve (AUC).	[38, 55]
Domain-adapted vocal tone models	Existing speech emotion recognition (SER) models are trained on actor portrayals, not executive speech.	SER models fall to chance on actual conference calls; domain-adapted models (e.g., FinVoc2Vec) are starting to be validated but require scaling and benchmarking at scale.	Release a labeled executive speech corpus dual-coded for valence/arousal; fine-tune wav2vec 2.0 and Whisper encoder via SpeechBrain; evaluate with Equal Error Rate (EER) and concordance with abnormal returns; publish weights under open license.	[20, 38]
Video-enabled call analysis	Post-COVID video earnings calls (Zoom, Teams) make executive facial behavior observable at scale for the first time.	Visual modality appears in only 21 papers, none with a validated large-sample methodology.	Build a replicable AU extraction pipeline (OpenFace 2.0 or py-feat); establish Facial Action Coding System (FACS) inter-rater reliability (target $\hat{\kappa} > 0.60$) as minimum reporting standard; link AU time-series to abnormal returns via event-study; release code and AU sequences openly.	[18]
Cross-market and multilingual generalizability	Most papers study US firms in English; tone-return relationships vary across institutional settings.	Non-English and non-US earnings call research is systematically under-represented.	Source non-English transcripts via Refinitiv Eikon or Seeking Alpha; apply XLM-RoBERTa or mBERT; report zero-shot and fine-tuned F1 on held-out language splits using FinBERT as English reference; test slope stability with country fixed effects.	[11]
Dynamic within-call tone analysis	Tone fluctuates across the prepared segment and Q&A; static document-level scores average over meaningful variation.	No paper models intra-call tone trajectories or management response adaptation across analyst questions.	Build multi-quarter sentence-level panel (Whisper ASR + FinBERT); apply Mamba or long-context LLM for sequence modeling; use expanding-window CV to prevent look-ahead bias; evaluate via out-of-sample tone-return association.	[31, 32]

(Continued)

Table 8
(Continued)

Research stream	Conceptual context	Observed gap	Future research directions	Key references
Deception and misreporting detection	Managing three modalities simultaneously is cognitively costly; leakage in one channel may reveal deception.	Multimodal deception detection exists in psychology but has not been applied to a large-scale earnings call corpus.	Link calls to Securities and Exchange Commission (SEC) Accounting and Auditing Enforcement Releases (AAER) enforcement database as ground truth; classify with FinBERT + wav2vec 2.0; evaluate via Receiver Operating Characteristic - Area Under the Curve (ROC-AUC) on held-out firm-years; apply SHapley Additive exPlanations (SHAP) for interpretability; pre-register hypotheses given low base rates.	[20, 33, 40, 56]

Note: Methodological specifications (sample thresholds, evaluation metrics, model configurations) represent the authors' informed recommendations based on adjacent-domain practice and should not be interpreted as evidence-based consensus standards. Where specifications have direct empirical precedent in the earnings call literature, references are provided in the Key References column.

corpus rather than as confirmatory hypothesis tests. The remaining method categories form a secondary exploratory family ($m = 4$); no test in this family reaches the BH-adjusted threshold. Additional families covering modality temporal trends ($m = 5$), period comparisons ($m = 14$), and modality–outcome associations ($m = 6$) likewise yield no significant results after correction, consistent with the descriptive character of this review. Study-wide Bonferroni correction ($\alpha_{\text{Bonf}} = 0.0017$) is not applied, as it is designed for confirmatory hypothesis-discovery settings and would be inappropriately conservative for supplementary descriptive evidence in a systematic literature review.

Interpretation and Caveats. This post hoc exploratory analysis reveals a descriptive trend: transformers show a 73% annual increase in odds and 68% in annual rate, concentrated in 2023–2026. Critically, this expansion occurs *alongside* rather than *displacing* lexicon and deep-learning methods. Three caveats—(1) emergence is recent (2023 onward, not gradual from 2010), (2) sparse data (11 transformer papers; low statistical power), and (3) findings—apply only to the Q1/Q2 and CORE A/A* corpus, not the broader earnings call literature.

Cross-Cutting infrastructure requirements. Realizing these research directions requires establishing a standardized empirical foundation. Foremost among these needs is an open-access multimodal earnings call benchmark: a synchronized text–audio (and ultimately video) corpus featuring fine-grained, utterance-level annotations. This foundation must be paired with rigorous evaluation protocols, particularly time-series splits designed to prevent look-ahead contamination, alongside standardized cross-study performance metrics. Methodological maturation also hinges on reproducible scientific practices, requiring the public distribution of end-to-end preprocessing pipelines, training code, and fine-tuned checkpoints. Finally, to foster transparency and comparative rigor, future work should systematically detail domain-adaptation protocols for foundation models (such as wav2vec 2.0, Whisper, or LLaMA), explicitly reporting parameter-efficient fine-tuning architectures such as LoRA and QLoRA. Table 8 organizes these directions into six actionable research streams, each grounded in a specific gap observed in the corpus, with concrete methodological specifications.

Funding Support

This work was supported by national funds through FCT, I.P. (Fundação para a Ciência e a Tecnologia, I.P.), under the project UID/04152/2025 – Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS (DOI: [10.54499/UID/04152/2025](https://doi.org/10.54499/UID/04152/2025)), and by the European Union – NextGenerationEU under the project UID/PRR/04152/2025 (DOI: [10.54499/UID/PRR/04152/2025](https://doi.org/10.54499/UID/PRR/04152/2025)).

Ethical Statement

This study does not contain any studies with human subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data are available from the corresponding author upon reasonable request.

Author Contribution Statement

Rodrigo dos Reis Maia: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Jorge Miguel Bravo:** Validation, Resources, Writing – review & editing, Supervision, Project Administration, Funding Acquisition.

References

- [1] Dang, M., Puwanenthiren, P., Nguyen, M. T., Hoang, V. A., Mazur, M., & Henry, D. (2022). Does managerial tone matter for stock liquidity? Evidence from textual disclosures. *Finance Research Letters*, 48, 102917. <https://doi.org/10.1016/j.frl.2022.102917>

- [2] Guest, N. M. (2021). The information role of the media in earnings news. *Journal of Accounting Research*, 59(3), 1021–1076. <https://doi.org/10.1111/1475-679X.12349>
- [3] Hennig, J. C., Firk, S., & Wolff, M. (2023). Credibility signals from soft information: Evidence from investor reactions to tone in earnings conference calls. In *European Accounting Review*. Advance online publication. <https://doi.org/10.1080/09638180.2023.2244009>
- [4] Luo, S., & Li, Y. H. (2024). The tone of earnings communication conferences and trade credit financing of listed companies. *Finance Research Letters*, 70, 106354. <https://doi.org/10.1016/j.frl.2024.106354>
- [5] Yang, B. C., Liu, Y. F., & Su, Y. P. (2023). Earnings communication conferences and post-earnings-announcement drift. *Accounting & Finance*, 63(2), 2145–2185. <https://doi.org/10.1111/acfi.12956>
- [6] Chen, S. P., Miao, B., & Valentine, K. (2022). Corporate control contests and the asymmetric disclosure of bad news: Evidence from hostile takeover bids. *The Accounting Review*, 97(1), 123–146. <https://doi.org/10.2308/TAR-2018-0619>
- [7] Delisle, R. J., Grant, A., & Mao, R. Q. (2024). Does environmental and social performance affect pricing efficiency? Evidence from earnings conference calls. *Journal of Corporate Finance*, 86, 102585. <https://doi.org/10.1016/j.jcorpfin.2024.102585>
- [8] Liu, Y., Jin, J., Zhang, Z., & Zhao, R. (2022). Firm-level political sentiment and corporate tax avoidance. *International Review of Financial Analysis*, 84, 102358. <https://doi.org/10.1016/j.irfa.2022.102358>
- [9] Suslava, K. (2021). Stiff business headwinds and uncharted economic waters: The use of euphemisms in earnings conference calls. *Management Science*, 67(11), 7184–7213. <https://doi.org/10.1287/mnsc.2020.3826>
- [10] Call, A. C., Sharp, N. Y., & Shohfi, T. D. (2021). Which buy-side institutions participate in public earnings conference calls? *Journal of Corporate Finance*, 68, 101964. <https://doi.org/10.1016/j.jcorpfin.2021.101964>
- [11] Gow, I. D., Larcker, D. F., & Zakolyukina, A. A. (2021). Non-answers during conference calls. *Journal of Accounting Research*, 59(4), 1349–1384. <https://doi.org/10.1111/1475-679X.12371>
- [12] Li, K., Mai, F., Shen, R., & Yan, X. (2021). Measuring corporate culture using machine learning. *The Review of Financial Studies*, 34(7), 3265–3315. <https://doi.org/10.1093/rfs/hhac079>
- [13] Bassyoung, H., Abdelfattah, T., & Tao, L. (2022). Narrative disclosure tone: A review and areas for future research. *Journal of International Accounting, Auditing and Taxation*, 49, 100511. <https://doi.org/10.1016/j.intaccaudtax.2022.100511>
- [14] Bochkay, K., Brown, S. V., Leone, A. J., & Tucker, J. W. (2023). Textual analysis in accounting: What's next? *Contemporary Accounting Research*, 40(2), 765–805. <https://doi.org/10.1111/1911-3846.12825>
- [15] Ning, J., Tao, L., Mi, B., & Zhang, L. (2025). The use of managerial tone in corporate disclosure: A literature review in accounting and finance. *Journal of Accounting Literature*, 1–88. <https://doi.org/10.1108/JAL-07-2024-0150>
- [16] Todd, A., Bowden, J., & Moshfeghi, Y. (2024). Text-based sentiment analysis in finance: Synthesising the existing literature and exploring future directions. *Intelligent Systems in Accounting, Finance and Management*, 31(1). <https://doi.org/10.1002/isaf.1549>
- [17] Baik, B., Kim, A. G., Kim, D. S., & Yoon, S. (2024). Vocal delivery quality in earnings conference calls. *Journal of Accounting and Economics*, 77, 101763. <https://doi.org/10.1016/j.jacceco.2024.101763>
- [18] Banker, R. D., Ding, H., Huang, R., & Li, X. (2024). Market reaction to CEOs' dynamic hemifacial asymmetry of expressions. *Management Science*, 70(11), 7434–7455. <https://doi.org/10.1287/mnsc.2023.4922>
- [19] Liu, E., & Qin, H. (2024). Can managers' facial expressions predict future company performance and risk? Evidence from China. *Finance Research Letters*, 59, 104794. <https://doi.org/10.1016/j.frl.2023.104794>
- [20] Mayew, W. J., & Venkatachalam, M. (2012). The power of voice: Managerial affective states and future firm performance. *Journal of Finance*, 67(1), 1–43. <https://doi.org/10.1111/j.1540-6261.2011.01705.x>
- [21] Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133, 285–296. <https://doi.org/10.1016/j.jbusres.2021.04.070>
- [22] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, n71. <https://doi.org/10.1136/bmj.n71>
- [23] Yang, Y., Qin, Y., Fan, Y., & Zhang, Z. J. (2023). Unlocking the power of voice for financial risk prediction: A theory-driven deep learning design approach. *MIS Quarterly*, 47(1), 175–218. <https://doi.org/10.25300/MISQ/2022/17062>
- [24] Marzi, G., Balzano, M., Caputo, A., & Pellegrini, M. M. (2025). Guidelines for bibliometric-systematic literature reviews: 10 steps to combine analysis, synthesis and theory development. *International Journal of Management Reviews*, 27(1), 81–103. <https://doi.org/10.1111/ijmr.12381>
- [25] Raimundo, B., & Bravo, J. M. (2026). Forecasting meets portfolio theory: A bibliometric approach to decision-making under uncertainty. *Humanities and Social Sciences Communications*. Advance online publication. <https://doi.org/10.1057/s41599-026-07252-6>
- [26] Mukherjee, D., Lim, W. M., Kumar, S., & Donthu, N. (2022). Guidelines for advancing theory and practice through bibliometric research. *Journal of Business Research*, 148, 101–115. <https://doi.org/10.1016/j.jbusres.2022.04.042>
- [27] Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2), 806–841. <https://doi.org/10.1111/1911-3846.12832>
- [28] Fu, X., Wu, X., & Zhang, Z. (2021). The information role of earnings conference call tone: Evidence from stock price crash risk. *Journal of Business Ethics*, 173(3), 643–660. <https://doi.org/10.1007/s10551-019-04326-1>
- [29] Garcia, D., Hu, X., & Rohrer, M. (2023). The colour of finance words. *Journal of Financial Economics*, 147(3), 525–549. <https://doi.org/10.1016/j.jfineco.2022.11.006>
- [30] Hajek, P., & Munk, M. (2023). Speech emotion recognition and text sentiment analysis for financial disclosures. *Neural Computing and Applications*, 35(29), 21463–21477. <https://doi.org/10.1007/s00521-023-08470-8>
- [31] 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. <https://doi.org/10.1111/j.1475-679x.2010.00365.x>

- [32] Price, S. M., Doran, J. S., Peterson, D. R., & Bliss, B. A. (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking and Finance*, 36(4), 992–1011. <https://doi.org/10.1016/j.jbankfin.2011.10.013>
- [33] Hobson, J. L., Mayew, W. J., & Venkatachalam, M. (2012). Analyzing speech to detect financial misreporting. *Journal of Accounting Research*, 50(2), 349–392. <https://doi.org/10.1111/j.1475-679x.2011.00433.x>
- [34] Allee, K. D., & DeAngelis, M. D. (2015). The structure of voluntary disclosure narratives: Evidence from tone dispersion. *Journal of Accounting Research*, 53(2), 241–274. <https://doi.org/10.1111/1475-679x.12072>
- [35] Bushee, B. J., Gow, I. D., & Taylor, D. J. (2018). Linguistic complexity in firm disclosures: Obfuscation or information? *Journal of Accounting Research*, 56(1), 85–121. <https://doi.org/10.1111/1475-679X.12179>
- [36] Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1), 35–65. <https://doi.org/10.1111/j.1540-6261.2010.01625.x>
- [37] Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, 62(3), 1139–1168. <https://doi.org/10.1111/j.1540-6261.2007.01232.x>
- [38] Ewertz, J., Knickrehm, C., Nienhaus, M., & Reichmann, D. (2026). Listen closely: Measuring vocal tone in corporate disclosures. *Journal of Accounting Research*, 64(1), 229–277. <https://doi.org/10.1111/1475-679x.70015>
- [39] Yang, H., Liu, X.-Y., & Wang, C. D. (2023). *FinGPT: Open-source financial large language models*. *arXiv*. <https://arxiv.org/abs/2306.06031>
- [40] Cheng, S. Y., & Golshan, N. M. (2025). Silent suffering: Using machine learning to measure CEO depression. *Journal of Accounting Research*, 63(2), 451–510. <https://doi.org/10.1111/1475-679X.12590>
- [41] Angelo, B., Johnston, M., Singh, A., & Wan, Y. Q. (2025). Tone distance: Managerial tone divergence and market reaction to earnings announcements. *Financial Review*, 60(4), 1415–1435. <https://doi.org/10.1111/fire.70002>
- [42] Jiang, F., Lee, J., Martin, X., & Zhou, G. (2019). Manager sentiment and stock returns. *Journal of Financial Economics*, 132(1), 126–149. <https://doi.org/10.1016/j.jfineco.2018.10.001>
- [43] Bochkay, K., & Joos, P. R. (2021). Macroeconomic uncertainty and quantitative versus qualitative inputs to management earnings forecasts. *The Accounting Review*, 96(3), 59–90. <https://doi.org/10.2308/TAR-2017-0490>
- [44] Koval, R., Andrews, N., & Yan, X. (2023). Forecasting earnings surprises from conference call transcripts. In *Findings of the Association for Computational Linguistics: ACL 2023*, 8197–8209. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.520>
- [45] Comprix, J., Lopatta, K., & Tideman, S. A. (2022). The role of gender in the aggressive questioning of CEOs during earnings conference calls. *The Accounting Review*, 97(7), 79–107. <https://doi.org/10.2308/TAR-2019-1029>
- [46] De Amicis, C., Falconieri, S., & Tastan, M. (2021). Sentiment analysis and gender differences in earnings conference calls. *Journal of Corporate Finance*, 71, 101809. <https://doi.org/10.1016/j.jcorpfin.2020.101809>
- [47] Vi, L., Dang, H. X., & Vu, N. T. (2024). Impression management in the boardroom: How CEO facial trustworthiness influences turnover risk. *Journal of Behavioral and Experimental Finance*, 44, 100978. <https://doi.org/10.1016/j.jbef.2024.100978>
- [48] Maia, R. R., & Bravo, J. M. (2026). A new scale for evaluating disclosure in earnings calls on emerging markets. *Emerging Science Journal*, 10(1), 247–275. <https://doi.org/10.28991/ESJ-2026-010-01-012>
- [49] Yang, Y. J., & Tan, H. T. (2024). What you are versus what you do: The effect of noun-verb framing in earnings conference calls. *Accounting, Organizations and Society*, 113, 101573. <https://doi.org/10.1016/j.aos.2024.101573>
- [50] Jin, Z. (2024). Business aspects in focus, investor underreaction and return predictability. *Journal of Corporate Finance*, 84, 102525. <https://doi.org/10.1016/j.jcorpfin.2023.102525>
- [51] Allee, K. D., Do, C., & Do, H. N. (2025). Can analysts elicit useful information by asking unique questions in earnings conference calls? *European Accounting Review*, 34(4), 1689–1715. <https://doi.org/10.1080/09638180.2024.2410892>
- [52] Papadopoulou, T., Theoharakis, V., Jones, M. V., & Bhaumik, S. K. (2024). Analysing the macrostructure of spoken strategic communication: An application of argumentation analysis on high-technology newly public firms' earnings conference calls. *British Journal of Management*, 35(1), 68–85. <https://doi.org/10.1111/1467-8551.12670>
- [53] Kaikaus, J., Hobson, J. L., & Brunner, R. (2022). Truth or fiction: Multimodal learning applied to earnings calls. In *2022 IEEE International Conference on Big Data*, 3607–3612. IEEE. <https://doi.org/10.1109/BigData55660.2022.10020307>
- [54] Huang, J., Xiao, M., Li, D., Jiang, Z., Yang, Y., Zhang, Y., . . . , & Qian, L. (2024). *Open-FinLLMs: Open multimodal large language models for financial applications*. *arXiv*. <https://arxiv.org/abs/2408.11878>
- [55] Todd, A., Bowden, J., Cummins, M., & Su, Y. (2025). A multimodal sentiment classifier for financial decision making. *International Review of Financial Analysis*, 105, 104322. <https://doi.org/10.1016/j.irfa.2025.104322>
- [56] Li, K., Mai, F., Wong, G., Yang, C., & Zhang, T. F. (2026). Female equity analysts and corporate environmental and social performance. *Management Science*. Advance online publication. <https://doi.org/10.1287/mnsc.2024.06429>
- [57] Zhou, K., Qu, Z., Liang, J., Tao, Y., & Zhu, M. (2024). Threat or shield: Environmental administrative penalties and corporate tone management. *Finance Research Letters*, 61, 105031. <https://doi.org/10.1016/j.frl.2024.105031>
- [58] Allee, K. D., Do, C., & Sterin, M. (2021). Product market competition, disclosure framing, and casting in earnings conference calls. *Journal of Accounting and Economics*, 72(1), 101405. <https://doi.org/10.1016/j.jacceco.2021.101405>
- [59] Rennekamp, K. M., Sethuraman, M., & Steenhoven, B. A. (2022). Engagement in earnings conference calls. *Journal of Accounting and Economics*, 74(1), 101498. <https://doi.org/10.1016/j.jacceco.2022.101498>
- [60] Bai, Y., & Yao, R. (2023). Visualizing environmental management: Corporate environmental images in annual reports. *Environmental Impact Assessment Review*, 99, 107026. <https://doi.org/10.1016/j.eiar.2022.107026>

How to Cite: dos Reis Maia, R., & Bravo, J. M. (2026). More Than Words: Managerial Tone, Communication Modalities, and Multimodal Measures in Earnings Calls. *Journal of Data Science and Intelligent Systems*. <https://doi.org/10.47852/bonviewJDSIS620210679>

Appendix A: Top 15 Most-Cited Papers (2021–2026)

Table 9
Top 15 most-cited papers published during the 2021–2026 period

Author (year)	Title	Citation	Modality	Method	Outcome
Li et al. [12]	Measuring corporate culture using machine learning	513	Textual	ML	Measurement
Huang et al. [27]	FinBERT: LLM for financial sentiment extraction	362	Textual	Transformer, ML, Deep Learning (DL)	Measurement
Li et al. [56]	Female Equity Analysts and Corporate Environmental and Social Performance	86	Textual	Transformer	Measurement
Guest [2]	The information role of the media in earnings news	64	Textual	Lexicon	Market reaction
Fu et al. [28]	Information role of earnings conference call tone	56	Textual	Lexicon	Market reaction
Zhou et al. [57]	Environmental penalties and corporate tone management	50	Textual	Lexicon	Market reaction
Gow et al. [11]	Non-answers during conference calls	50	Interact.	Content	Market reaction
Hajek and Munk [30]	Speech emotion recognition and text sentiment analysis	45	Multimodal	ML, SER	Measurement
García et al. [29]	The color of finance words	45	Textual	Lexicon	Market reaction
De Amicis et al. [46]	Sentiment analysis and gender differences in calls	42	Textual	ML	Market reaction
Allee et al. [58]	Competition, disclosure framing, and casting in calls	36	Interactional	Content	Market reaction
Rennekamp et al. [59]	Engagement in earnings conference calls	25	Interactional	Content	Market reaction
Bochkay and Joos [43]	Macroeconomic uncertainty and earnings forecasts	24	Textual	Content	Analyst forecast
Bai and Yao [60]	Visualizing environmental management in annual reports	24	Visual	Content	Market reaction
Call et al. [10]	Which buy-side institutions attend earnings calls?	23	Interactional	Content	Analyst forecast

Note: Table includes modality, method, and outcome tags for each paper. This ranking highlights the most influential studies published during the 2021–2026 period within the review time frame, thereby allowing recent high-impact contributions to be represented. Citation counts are from Scopus as of March 18, 2026.

Appendix B: Screening Keywords

Table 10
Search query keywords used in WoS and Scopus

Group	Domain	Search terms
1	Corporate communication context	“impression management,” “corporate narrative,” “earnings call*,” “conference call*,” “analyst call*,” “quarterly results,” “earnings report*,” “earnings release*,” “financial disclosure,” “managerial commentary”
2	Tone and sentiment analysis	“sentiment analysis,” “opinion mining,” “polarity analysis,” “textual sentiment,” “affect analysis,” “mood analysis,” “emotion recognition,” “tone*”

Note: Search groups were combined using the AND operator. Identical Boolean query terms were applied across Web of Science (WoS) and Scopus for both target document types: journal articles (DT=Article) and conference proceedings (DT=Proceedings Paper in WoS; DOCTYPE(cp) in Scopus). Searches were restricted to English-language publications between 2000 and 2026 (PY = 2000–2026; LA=English).

Appendix C: Reproducible Search Strategy

Overview and Rationale

This appendix details the reproducible search strategy for this review. A Primary Search on March 18, 2026, retrieved 1,173 records across Scopus and Web of Science. A subsequent Robustness Search (executed July 2, 2026, applying the original March 18, 2026, temporal cut-off) expanded modality descriptors to assess sensitivity. It identified 149 additional records, with only 1 meeting full inclusion criteria. The final synthesis corpus comprises 178 peer-reviewed publications.

Search Logic and Components

All searches used a consistent two-group Boolean structure, adapted to each database's field operators:

Group 1 – Corporate Communication Context: “impression management,” “corporate narrative,” “earnings call*,” “conference call*,” “analyst call*,” “quarterly results,” “earnings report*,” “earnings release*,” “financial disclosure,” “managerial commentary.”

Group 2 – Tone and Sentiment Analysis Methods: “sentiment analysis,” “opinion mining,” “polarity analysis,” “textual sentiment,” “affect analysis,” “mood analysis,” “emotion recognition,” “tone*,” and for Robustness only: “prosody,” “acoustic feature,” “facial expression,” “multimodal fusion.”

Combined logic: (Group 1 terms) AND (Group 2 terms) + Applied Filters

Temporal Scope and Data Integrity

Primary search: Executed March 18, 2026, across Scopus (932 records) and Web of Science (241 records), totaling 1,173 records. 116 cross-database duplicates removed using DOI matching and manual title/author comparison. **Robustness search:** Conducted July 2, 2026, restricted to records published by March 18, 2026, ensuring temporal consistency. Retrieved 149 additional records; 1 met inclusion criteria. **Final corpus:** 178 papers (166 journal articles, 12 conference papers, spanning 2010–2026).

Robustness Search: Expanded Modality Terms (July 2, 2026–reference of March 18, 2026)

To assess whether the original Group 2 terms captured available evidence on multimodal sentiment analysis, a robustness search was executed on July 2, 2026, with strict reference of March 18, 2026 (see Table 11) using the same Group 1 context plus expanded Group 2 including acoustic and visual modality descriptors (prosody, acoustic feature, facial expression, multimodal fusion).

Scopus Robustness Query

Database: Scopus – **Execution:** July 2, 2026 – **Temporal restriction:** Records published by March 18, 2026 – **Additional initial records:** 127
Syntax: TITLE-ABS-KEY (“impression management” OR “corporate narrative” OR “earnings call*” OR “conference call*” OR “analyst call*” OR “quarterly results” OR “earnings report*” OR “earnings release*” OR “financial disclosure” OR “managerial commentary”) AND ALL (“sentiment analysis” OR “opinion mining” OR “polarity analysis” OR “textual sentiment” OR “affect analysis” OR “mood analysis” OR “emotion recognition” OR “tone*” OR “prosody” OR “acoustic feature” OR “facial expression” OR “multimodal fusion”) AND PUBYEAR >= 2000 AND PUBYEAR <= 2026 AND LANGUAGE (“English”) AND DOCTYPE (“ar” OR “cp”)

Web of Science Robustness Query

Database: Web of Science – **Execution:** July 2, 2026 – **Temporal restriction:** Records published by March 18, 2026 – **Additional records:** 22
Syntax: TS=((“impression management” OR “corporate narrative” OR “earnings call*” OR “conference call*” OR “analyst call*” OR “quarterly results” OR “earnings report*” OR “earnings release*” OR “financial disclosure” OR “managerial commentary”) AND (“sentiment analysis” OR “opinion mining” OR “polarity analysis” OR “textual sentiment” OR “affect analysis” OR “mood analysis” OR “emotion recognition” OR “tone*” OR “prosody” OR “acoustic feature” OR “facial expression” OR “multimodal fusion”)) AND PY = (2000-2026) AND (DT = (Article) OR DT = (Proceedings Paper)) AND LA = (English)

The robustness search executed on July 2, 2026, retrieved a total of 149 additional records, comprising 127 from Scopus and 22 from Web of Science, compared to the Primary Search. Following screening, 148 records were excluded (127 from Scopus and 21 from Web of Science), leaving a single record (1) from Web of Science that advanced and was ultimately included in the final corpus.

Finding: The expanded modality search identified one additional eligible paper ([23], DeepVoice), confirming that the original Group 2 terms captured the bulk of available evidence. Expanding to purely acoustic/visual terms added minimal scope.

PRISMA Screening Flow

Supplementary Scoping Search: IEEE Xplore

To guarantee rigorous coverage across the engineering and computational literature, we executed a supplementary scoping search in IEEE Xplore. We translated our two-group Boolean architecture into IEEE's native command syntax, restricting Group 1 corporate

Table 11
PRISMA 2020 screening flow showing combined primary and robustness
search universe (1,322 records)

PRISMA screening stage	Records	Cumulative
<i>Identification Phase</i>		
Initial database search (Scopus + WoS)	1,322	1,322
Web of Science	263	—
Scopus	1,059	—
Cross-database duplicates removed	−116	1,206
Missing metadata removed	−65	1,141
<i>Screening Phase</i>		
Title screening: excluded	−258	883
Abstract screening: excluded	−488	395
Advanced to full text	395	—
<i>Eligibility Phase</i>		
Does not address RQs: excluded	−51	344
Insufficient tone focus: excluded	−119	225
Eligible for quality assessment	225	—
<i>Inclusion Phase</i>		
Below Q1/Q2 (SJR): excluded	−47	178
Final Inclusion	178	178

context descriptors to all indexing metadata while executing Group 2 acoustic, visual, and multimodal keywords across full-text and metadata fields:

- **Database:** IEEE Xplore Digital Library – **Execution Date:** July 2, 2026
- **Temporal Restriction:** Publication years 2000–2026 (records published after March 18, 2026, were screened out to maintain temporal parity with the primary search)
- **Syntax:** ((“All Metadata”: “impression management” OR “corporate narrative” OR “earnings call” OR “earnings calls” OR “conference call” OR “conference calls” OR “analyst call” OR “analyst calls” OR “quarterly results” OR “earnings report” OR “earnings reports” OR “earnings release” OR “earnings releases” OR “financial disclosure” OR “managerial commentary”) AND (“Full Text & Metadata”: “sentiment analysis” OR “opinion mining” OR “polarity analysis” OR “textual sentiment” OR “affect analysis” OR “mood analysis” OR “emotion recognition” OR “tone” OR “tones” OR “tonality” OR “prosody” OR “acoustic feature” OR “facial expression” OR “multimodal fusion”))
- **Yield:** 42 initial records

Outcome and Justification for Protocol Omission: The search yielded 42 initial records. Following evaluation against our pre-specified quality criteria (Q1/Q2 SJR journal ranking or CORE A/A* conference standard), zero records qualified for full-text synthesis, as the retrieved items were published in unranked workshop proceedings, symposiums, or sub-tier conferences. This confirmed that peer-reviewed, high-impact multimodal earnings call research remains concentrated in mainstream accounting, finance, management, and core NLP venues already captured by Scopus and Web of Science. Consequently, IEEE Xplore was documented as an auxiliary scoping probe rather than integrated into the primary PRISMA screening flow.

Sensitivity to Corpus Definition

Table 12
Sensitivity analysis: key proportions across corpus definitions

Corpus definition	N	Textual	Visual	Transformer	Market react.
Primary (all papers)	178	65.2%	12.4%	6.2%	28.1%
Journal articles only	166	66.3%	11.8%	6.4%	28.9%
2015+ only	172	64.0%	13.0%	5.8%	27.3%
Both WoS and Scopus	154	63.6%	12.3%	6.5%	27.9%
Range	—	Δ2.7%	Δ1.2%	Δ0.7%	Δ1.6%

To assess robustness of key findings to methodological boundary decisions, three alternative corpus specifications were examined in Table 12.

As shown in Table 13, all primary findings remain robust within a narrow ± 3 percentage point margin across alternative specifications, confirming that our empirical insights are invariant to corpus boundary definitions. The only dimension exhibiting broader sensitivity was *meth_regression* (spanning 11.0%–16.9%), a variation driven by sample composition differences in baseline empirical designs across search boundaries rather than taxonomic ambiguity in our classification codebook (corroborated by excellent inter-rater agreement on the validation subset, with $\kappa = 0.918$).

Table 13
Sensitivity of key proportions to corpus definition

Corpus specification						
Tag	Baseline	Journals	Post-2014	WoS/Scopus	Range	Verdict
	(<i>n</i> = 178)	(<i>n</i> = 166)	(<i>n</i> = 172)	(<i>n</i> = 154)		
Textual modality	65.7	65.1	66.3	63.6	2.7	✓
Vocal modality	15.7	13.3	15.1	13.0	2.7	✓
Visual modality	11.8	12.7	12.2	13.0	1.2	✓
Interactional	19.7	19.3	19.2	18.2	1.5	✓
Lexicon method	9.0	9.6	9.3	9.1	0.6	✓
Deep learning	7.9	7.8	8.1	6.5	1.6	✓
Transformer	6.2	6.0	6.4	5.8	0.6	✓
Regression [†]	16.3	16.9	15.7	11.0	5.9	~
Earnings call ctx	87.6	86.7	87.2	85.7	1.9	✓
Market reaction	28.1	28.9	27.3	28.6	1.6	✓

[†]*meth_regression* sensitivity reflects corpus-dependent variation in prevalence (inter-coder $\kappa = 0.918$ on 60-paper validation sample); see Appendix E. ✓ = range ≤ 5 pp; ~ = borderline.

Appendix D: Sensitivity to Algorithmic Tag Reliability

The 60-paper validation sample assessed algorithm performance across all 31 coding tags. All 31 tags show interpretable performance with no F1 scores below 0.30. Notably strong performance appears on context tags (*ctx_earnings_call*, $F_1 = 0.950$ [0.880, 0.990]; *ctx_annual_report*, $F_1 = 0.833$ [0.700, 0.950]) and method tags (*meth_regression*, $F_1 = 0.977$ [0.930, 1.000]; *meth_classical_ml*, $F_1 = 1.000$ [1.000, 1.000]). Five tags show relatively lower (but still interpretable) F1 scores in the Fair–Moderate range ($F_1 = 0.400$ –0.556): *out_volatility* ($F_1 = 0.400$ [0.200, 0.600]; $\kappa = 0.355$), *meth_topic_model* ($F_1 = 0.500$ [0.250, 0.750]; $\kappa = 0.483$), *ctx_social_media* ($F_1 = 0.500$ [0.250, 0.750]; $\kappa = 0.483$), *actor_cfo* ($F_1 = 0.545$ [0.350, 0.720]; $\kappa = 0.510$), and *out_financial_performance* ($F_1 = 0.556$ [0.380, 0.730]; $\kappa = 0.378$).

Wide confidence intervals on rare/ambiguous categories: Wide bootstrap CIs reflect genuine ambiguity in category definitions and low prevalence in the validation sample. For example, *out_financial_performance* (κ CI [0.158, 0.598]) shows wide uncertainty because this outcome is rare and boundaries overlap between financial performance disclosure and actual company performance. *ctx_social_media* (κ CI [0.283, 0.683]) similarly reflects ambiguity in distinguishing social media earnings commentary from traditional earnings calls. These wide intervals appropriately temper overconfidence in algorithmic performance on these rare categories.

Visual modality performance and prevalence bounds: The 60-paper validation found 1 true positive for *mod_visual* among 59 papers classified as “no visual,” yielding a false-negative rate of 1.7% (95% CI [0.0%, 9.1%]). Algorithm precision for the visual modality was 0.500 [0.200, 0.800], recall = 1.000 [0.800, 1.000]. Extrapolating observed precision (0.500) to the full corpus suggests the algorithmic estimate of 12.4% visual prevalence may overestimate true prevalence by approximately 6.2 percentage points. True visual prevalence likely falls in the range 6.2%–12.4%. This indicates the algorithmic estimate is an upper bound, meaning the visual modality gap is, if anything, *larger* than reported.

Robustness of key findings: Because all 31 tags show interpretable performance ($F_1 \geq 0.30$, $\kappa \geq 0.20$), no tags are excluded from prevalence reporting. Binary tags are coded independently; pairwise correlation testing confirms near-zero correlations across categories (mean $|r| \approx 0.10$). The following core findings remain robust to algorithmic performance variation: textual modality dominance (65.7%), earnings call context prevalence (87.6%), transformer adoption (6.2%), market-reaction outcome (28.1%), and the transformer adoption trend (Cochran–Armitage $p = 0.012$).

Tags with lower agreement: For tags with moderate κ values (e.g., *out_financial_performance*, $\kappa = 0.378$; *out_volatility*, $\kappa = 0.355$), low inter-rater reliability reflects genuine ambiguity in distinguishing related category boundaries rather than algorithmic failure. We document these tags transparently in Appendix F (Table 14) and do not rely on them for confirmatory inferences.

Validation Sample Selection. To ensure balanced representation across publication cohorts (2010–2026; five temporal strata), methodological paradigms (eight classes), and target modalities (five categories), we extracted a stratified random sample of 60

manuscripts (33.7% sampling intensity) from the 178-paper corpus without replacement. Two independent evaluators coded all 60 validation papers across 31 binary dimensions ($N = 1,860$ distinct decisions per coder). Sample extraction was conducted strictly prior to and independent of downstream scoring; all 60 manuscripts were retained without post hoc exclusions or sample reweighting for inter-rater reliability (κ) estimation.

Coding Protocol and Reliability Assessment. The entire 178-paper corpus was parsed and assigned binary tags across all 31 dimensions by our computational classification pipeline. To rigorously benchmark automated classification fidelity, two evaluators independently assessed the 60-paper stratified subset under mutual and algorithmic blinding. The 31-tag coding codebook was developed iteratively, establishing explicit boundary conditions, positive and negative exemplars, and decision rules for borderline classifications prior to being permanently frozen. Coders evaluated the 60 validation papers under strict mutual blinding, with no intermediate communication, adjudication, or knowledge of each other's judgments. Unadjudicated double-coded outputs were cross-tabulated into 2×2 confusion matrices to calculate exact tag-level Cohen's κ .

Appendix E: Inter-Coder Reliability Summary

This appendix provides a summary of inter-coder reliability assessment on the 60-paper validation sample.

Overall Agreement: Cohen's κ across all 31 tags is $\kappa = 0.7116$ (Median = 0.7321), indicating substantial agreement on the Landis–Koch scale. By category: Actor ($\kappa = 0.6646$), Context ($\kappa = 0.7670$), Method ($\kappa = 0.8143$), Modality ($\kappa = 0.6678$), and Outcome ($\kappa = 0.5948$).

Key Findings: Ten tags achieved excellent agreement ($\kappa \geq 0.81$), 11 showed substantial agreement ($\kappa 0.61$ – 0.80), and 8 showed moderate agreement ($\kappa 0.41$ – 0.60). No tags fell below fair agreement ($\kappa < 0.20$), confirming that all 31 coding categories remain interpretable and suitable for systematic review. The breakdown by category appears in the main manuscript (Table: Inter-coder reliability results).

Validation Metrics: Complete validation metrics with 95% bootstrap confidence intervals (precision, recall, F1, and Cohen's κ) for all 31 tags on the 60-paper sample are detailed in Table 14.

Appendix F: Validation Methodology, Inter-Coder Agreement, and Confusion Matrices

F.1 Validation Metrics with 95% Bootstrap Confidence Intervals. Table 14 presents 95% confidence intervals (bootstrap, 1,000 iterations) for Cohen's κ , precision, recall, and F1 across all 31 coding tags. Wide confidence intervals on ambiguous tags (e.g., *out_volatility*: $\kappa = 0.355$ [0.135, 0.575]; *topic_model*: $\kappa = 0.483$ [0.283, 0.683]) reflect genuine conceptual disagreement in these category definitions; tight confidence intervals on objectively-defined tags (e.g., *ctx_earnings_call*: $\kappa = 0.689$ [0.539, 0.839], *meth_regression*: $\kappa = 0.918$ [0.838, 0.998]) indicate high reproducibility. Data source: validation sample analysis ($n = 60$ papers, 1,000 bootstrap iterations).

F.2 Complete 2×2 Confusion Matrices for All 31 Coding Tags. Table 15 displays complete 2×2 confusion matrices for all 31 coding tags, enabling transparent assessment of inter-coder agreement (Coder A – reference standard vs. Coder B) on the stratified validation sample of 60 papers. The matrices reveal which tags had clear agreement, which showed systematic disagreement patterns, and which rare tags had undefined metrics. All 31 confusion matrices are also available in machine-readable format in the supplementary CSV file `appendix_g_confusion_matrices.csv` for independent verification.

F.2a Interpretation of Agreement Patterns. Examination of the confusion matrices reveals that agreement strength correlates with definitional clarity. Categories with objective boundaries (e.g., Regression: FN = 0, $\kappa=0.918$; Classical ML: FN = 0, $\kappa=1.000$; Deep Learning: FN = 0, $\kappa=1.000$) show near-perfect agreement. Some outcome categories (Financial Performance: FN = 12, $\kappa=0.378$, Fair; Volatility: FN = 5, $\kappa=0.355$, Fair) exhibit Fair agreement, reflecting genuine conceptual boundaries in distinguishing financial performance outcomes from financial disclosure practices, and volatility outcomes from general market reactions. These Fair-agreement tags remain interpretable and operationally valid; their lower agreement reflects substantive ambiguity in the accounting and finance literature rather than framework failure. Overall, all 31 tags demonstrate reliable validation across the full interpretability spectrum (Fair through Perfect), confirming the framework's utility for systematic review.

F.3 False-Negative Analysis for Visual Modality and Transformer Adoption. This section documents manual review of papers classified as not containing the target constructs, assessing the estimated false-negative rate.

Visual Modality Analysis. Among the 60 validation papers, 59 were classified as containing no visual modality. Manual review of these 59 papers examining full-text references to video, image, or visual analysis methods identified 1 false negative (a paper with visual elements that the algorithm missed). This yields an estimated false-negative rate of 1.7% (95% CI: [0.0%, 9.1%]), confirming that the visual modality gap represents a genuine literature scarcity rather than substantial algorithmic undercounting.

Transformer Adoption Analysis. Among the 60 validation papers, 57 were classified as containing no transformer-based NLP methodology. Manual review of these 57 papers examining methodology sections and keywords identified zero false negatives (all papers genuinely lacking transformer methods). This yields an estimated false-negative rate of $\leq 0\%$ (95% CI: [0.0%, 10.4%]), confirming that transformer adoption conclusions are robust and reflect genuine patterns in the literature.

Conclusion. Neither the visual modality gap nor the transformer adoption conclusions are substantial artifacts of algorithmic misclassification. The false-negative rates, estimated conservatively at the bounds of the binomial CI, are consistent with zero actual false negatives in manual review. Data source: manual review of validation sample papers not containing the target construct.

Table 14
Validation metrics with 95% bootstrap confidence intervals for all 31 coding tags ($n = 60$ papers)

Category	Tag	κ [95% CI]	Precision [95% CI]	Recall [95% CI]	F1 [95% CI]	Agreement
<i>Modality</i>	Textual	0.700 [0.550, 0.850]	0.981 [0.900, 1.000]	0.964 [0.880, 1.000]	0.972 [0.920, 1.000]	0.950
	Vocal	0.856 [0.776, 0.936]	1.000 [1.000, 1.000]	0.778 [0.600, 0.950]	0.875 [0.750, 0.980]	0.967
	Visual	0.659 [0.509, 0.809]	0.500 [0.200, 0.800]	1.000 [0.800, 1.000]	0.667 [0.450, 0.880]	0.983
	Multimodal	0.564 [0.384, 0.744]	0.600 [0.300, 0.900]	0.600 [0.300, 0.900]	0.600 [0.350, 0.850]	0.933
	Interactional	0.560 [0.380, 0.740]	0.800 [0.600, 1.000]	0.600 [0.400, 0.800]	0.686 [0.500, 0.870]	0.817
<i>Method</i>	Classical ML	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000
	Deep Learning	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000
	Content Analysis	0.949 [0.869, 1.000]	1.000 [0.950, 1.000]	0.923 [0.800, 1.000]	0.960 [0.900, 1.000]	0.983
	Regression	0.918 [0.838, 0.998]	0.955 [0.880, 1.000]	1.000 [0.900, 1.000]	0.977 [0.930, 1.000]	0.967
	Transformer	0.869 [0.789, 0.949]	0.889 [0.750, 1.000]	0.889 [0.750, 1.000]	0.889 [0.770, 0.980]	0.967
	Acoustic/SER	0.856 [0.776, 0.936]	1.000 [1.000, 1.000]	0.778 [0.600, 0.950]	0.875 [0.750, 0.980]	0.967
	Readability	0.732 [0.582, 0.882]	0.750 [0.600, 0.900]	0.750 [0.600, 0.900]	0.750 [0.620, 0.880]	0.983
	Lexicon-based	0.521 [0.341, 0.701]	0.955 [0.850, 1.000]	0.600 [0.400, 0.800]	0.737 [0.600, 0.870]	0.750
	Topic Model	0.483 [0.283, 0.683]	0.500 [0.200, 0.800]	0.500 [0.200, 0.800]	0.500 [0.250, 0.750]	0.967
<i>Outcome</i>	Fraud/Misreporting	0.783 [0.663, 0.903]	1.000 [0.800, 1.000]	0.667 [0.400, 0.900]	0.800 [0.600, 0.950]	0.967
	Market Reaction	0.766 [0.646, 0.886]	0.923 [0.800, 1.000]	0.828 [0.650, 1.000]	0.873 [0.750, 0.970]	0.883
	Analyst Forecast	0.772 [0.652, 0.892]	0.857 [0.700, 1.000]	0.750 [0.550, 0.950]	0.800 [0.650, 0.930]	0.950
	Information Asymmetry	0.516 [0.336, 0.696]	0.667 [0.450, 0.880]	0.500 [0.300, 0.700]	0.571 [0.400, 0.750]	0.900
	Financial Performance	0.378 [0.158, 0.598]	0.714 [0.500, 0.900]	0.455 [0.250, 0.650]	0.556 [0.380, 0.730]	0.733
	Volatility	0.355 [0.135, 0.575]	0.667 [0.400, 0.900]	0.286 [0.100, 0.500]	0.400 [0.200, 0.600]	0.900

(Continued)

Table 14
(Continued)

Category	Tag	κ [95% CI]	Precision [95% CI]	Recall [95% CI]	F1 [95% CI]	Agreement
<i>Context</i>						
	ESG/Climate	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000
	Press Release	0.848 [0.728, 0.968]	1.000 [1.000, 1.000]	0.750 [0.550, 0.950]	0.857 [0.750, 0.950]	0.983
	Annual Report	0.815 [0.695, 0.935]	0.833 [0.650, 1.000]	0.833 [0.650, 1.000]	0.833 [0.700, 0.950]	0.967
	Earnings Call	0.689 [0.539, 0.839]	0.980 [0.920, 1.000]	0.923 [0.820, 1.000]	0.950 [0.880, 0.990]	0.917
	Social Media	0.483 [0.283, 0.683]	0.500 [0.200, 0.800]	0.500 [0.200, 0.800]	0.500 [0.250, 0.750]	0.967
<i>Actor</i>						
	Gender	0.733 [0.583, 0.883]	1.000 [1.000, 1.000]	0.600 [0.350, 0.850]	0.750 [0.600, 0.880]	0.967
	CEO	0.783 [0.663, 0.903]	0.867 [0.700, 1.000]	0.812 [0.620, 1.000]	0.839 [0.720, 0.930]	0.917
	Analyst	0.721 [0.571, 0.871]	0.733 [0.600, 0.870]	0.846 [0.680, 1.000]	0.786 [0.650, 0.900]	0.900
	Personality	0.707 [0.557, 0.857]	0.909 [0.750, 1.000]	0.667 [0.450, 0.850]	0.769 [0.620, 0.900]	0.900
	Top Management	0.533 [0.353, 0.713]	0.833 [0.650, 1.000]	0.735 [0.550, 0.900]	0.781 [0.620, 0.920]	0.767
	CFO	0.510 [0.330, 0.690]	1.000 [0.800, 1.000]	0.375 [0.150, 0.600]	0.545 [0.350, 0.720]	0.917

Note: Algorithm performance validated against Coder A (reference standard) on stratified validation sample. Precision = $TP/(TP + FP)$; Recall = $TP/(TP + FN)$; F1 = harmonic mean of precision and recall; Agreement = raw observed agreement [$po = (TP + TN)/60$]; κ = Cohen's kappa (inter-rater agreement between Coder A and Coder B). Confidence intervals computed from 1,000 nonparametric bootstrap resamples.

Table 15
Complete 2×2 confusion matrices for all 31 coding tags

Tag	TP	FP	FN	TN	κ
<i>Modality Tags</i>					
Textual	53	1	2	4	0.700
Vocal	7	0	2	51	0.856
Visual	1	1	0	58	0.659
Multimodal	3	2	2	53	0.564
Interactional	12	3	8	37	0.560
<i>Methodology Tags</i>					
Classical ML	8	0	0	52	1.000
Deep Learning	11	0	0	49	1.000
Content Analysis	12	0	1	47	0.949
Regression	42	2	0	16	0.918
Transformer	8	1	1	50	0.869
Acoustic/SER	7	0	2	51	0.856
Readability	3	1	1	55	0.732
Lexicon	21	1	14	24	0.521
Topic Model	1	1	1	57	0.483
<i>Outcome Tags</i>					
Fraud/Misreporting	4	0	2	54	0.783
Market Reaction	24	2	5	29	0.766
Analyst Forecast	6	1	2	51	0.772
Information Asymmetry	4	2	4	50	0.516
Financial Performance	10	4	12	34	0.378
Volatility	2	1	5	52	0.355
<i>Context Tags</i>					
ESG/Climate	8	0	0	52	1.000
Press Release	3	0	1	56	0.848
Annual Report	5	1	1	53	0.815
Earnings Call	48	1	4	7	0.689
Social Media	1	1	1	57	0.483
<i>Actor Tags</i>					
Gender	3	0	2	55	0.733
CEO	13	2	3	42	0.783
Analyst	11	4	2	43	0.721
Personality	10	1	5	44	0.707
Top Management	25	5	9	21	0.533
CFO	3	0	5	52	0.510

Note: Inter-coder agreement analysis between Coder A (reference standard) and Coder B on stratified validation sample ($n = 60$ papers). TP = true positives (both coders coded as 1); FP = false positives (Coder A coded 0, Coder B coded 1); FN = false negatives (Coder A coded 1, Coder B coded 0); TN = true negatives (both coders coded as 0); κ = Cohen's kappa inter-coder agreement statistic.