



Scripted and unscripted disclosure: Tone in Brazilian earnings calls and short-window market reactions

Rodrigo dos Reis Maia^{1b,*}, Jorge Miguel Bravo

NOVA Information Management School, Universidade NOVA de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal

ARTICLE INFO

Editor: Muhammad Naveed

Keywords:

Earnings calls
Corporate disclosure
Sentiment analysis
Abnormal returns
Emerging markets

ABSTRACT

Earnings calls provide a setting in which investors observe both scripted managerial communication and more interactive analyst–management exchanges. This paper examines whether linguistic tone in Brazilian earnings calls is associated with short-window market reactions, and whether the question-and-answer (Q&A) segment is more informative than prepared remarks. Using 1640 earnings-call transcripts from 77 Brazilian listed firms over 4Q 2016–4Q 2025, we measure tone separately for the prepared introduction and Q&A segments using Loughran–McDonald, VADER, and FinBERT, and relate these measures to cumulative abnormal returns and cumulative abnormal trading volume around call dates. The main finding is that negative tone in earnings calls is systematically associated with lower short-window abnormal returns, with negative tone in the Q&A segment emerging as the most robust textual correlate. Although formal tests indicate that the economic magnitude of negative tone in the Q&A segment is statistically indistinguishable from that in the prepared remarks, the Q&A association is estimated with substantially greater precision and cross-specification stability. By contrast, the association for tone in the prepared introduction is statistically unreliable and less consistently estimated across specifications, and the evidence for abnormal trading volume is more fragile. Additional evidence from call timing suggests that Q&A negativity is most consistently associated with abnormal returns when calls occur during trading hours, when investors can react as the discussion unfolds. The results are interpreted as reduced-form evidence that unscripted analyst–management exchanges contain information that investors regard as relevant in an emerging-market setting. The paper does not claim to identify a causal effect of tone, nor to establish the general superiority of any single sentiment model. Instead, it contributes to the disclosure literature by providing new evidence on earnings-call communication in Brazil and by showing that the internal segmentation of the call matters for understanding investor responses to qualitative corporate disclosure.

1. Introduction

Earnings calls are a central component of firms' disclosure environments. In addition to formal financial statements and earnings releases, these calls allow investors and analysts to observe how managers explain current results, discuss operating conditions, and respond to questions in real time. Because earnings calls combine prepared introductory remarks with a subsequent question-and-answer (Q&A) session, they provide a useful setting in which to study whether qualitative features of corporate communication are associated with short-window market reactions.

* Corresponding author.

E-mail addresses: 20220517@novaims.unl.pt (R. dos Reis Maia), jbravo@novaims.unl.pt (J.M. Bravo).

<https://doi.org/10.1016/j.iref.2026.105952>

Received 30 May 2026; Received in revised form 18 September 2026; Accepted 6 October 2026

Available online 15 October 2026

1059-0560/© 2026 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

A large body of literature shows that capital markets respond to corporate disclosures and that earnings calls can convey information beyond the earnings release itself. Prior studies document investor and analyst reactions during and after earnings calls, and show that linguistic tone and other communication cues may be related to stock returns, volatility, and trading activity (Bowen et al., 2002; Brockman et al., 2015; Bushee et al., 2003; Chen et al., 2023; Druz et al., 2020; Frankel et al., 1999; Keith & Stent, 2019; Matsumoto et al., 2011; Mayew & Venkatachalam, 2012). Yet three issues remain less settled. First, although earnings calls are known to be informative, many existing studies measure tone at the aggregate call level, without distinguishing prepared remarks from the Q&A exchange. This aggregation masks potentially important heterogeneity: the scripted introduction and the interactive Q&A are fundamentally different disclosure formats, and there is no reason to expect them to be equally informative to investors. Second, most existing evidence comes from U.S. settings, whereas emerging markets remain comparatively underexplored. Third, recent work increasingly uses machine-learning-based measures of tone, but stronger empirical associations should not automatically be interpreted as evidence of methodological superiority in the absence of direct validation.

This paper examines these issues in the context of Brazilian earnings calls. Using 1640 transcripts from Brazilian listed firms over 4Q 2016 to 4Q 2025, we measure tone separately for the prepared introduction and the Q&A segment, and relate these measures to cumulative abnormal returns (CAR) and cumulative abnormal trading volume (CAV) around call dates. The Brazilian setting is informative for two reasons. First, it extends the earnings-call literature beyond the U.S. market, where most existing evidence is concentrated. Second, it provides a useful institutional setting in which to examine whether qualitative features of disclosure are similarly associated with investor reactions in a large emerging market.

The paper focuses on the distinction between scripted and unscripted communication. Prepared introductory remarks are generally more structured and closer to the firm's intended disclosure message, whereas the Q&A segment may reveal additional information through spontaneous managerial responses to analyst questions. If investors view unscripted exchanges as more informative, tone in the Q&A segment should exhibit a stronger association with short-window market reactions than tone in the introductory portion of the call.

In line with expectations, the main results show that negative tone in earnings calls is associated with lower short-window abnormal returns. When comparing call segments, however, the evidence does not indicate a larger economic response to the Q&A segment: formal tests show that the estimated magnitudes of negative tone in the prepared introduction and the Q&A portion are statistically indistinguishable. Instead, an informative empirical pattern emerges in estimation precision: negative tone in the Q&A segment yields a statistically significant and consistently negative association with abnormal returns (under FinBERT), whereas tone in the prepared introduction, despite similar point estimates in some specifications, is imprecisely estimated and statistically unreliable across estimators. This asymmetry represents a novel finding in the Brazilian capital market setting, an emerging-market context grounded in specific institutional predictions that differ from those typically examined in U.S. markets. It suggests that aggregate whole-call tone measures may obscure informational content concentrated in the interactive portion of the event, highlighting the importance of segment-level analysis of earnings-call disclosures. The evidence for abnormal trading volume is weaker and less stable, consistent with volume reflecting not only information processing but also disagreement, liquidity conditions, and investor attention. The results are more consistent when tone is measured using FinBERT, suggesting that domain-adapted language models capture earnings-call language more effectively than general-purpose lexicons in this setting.

Additional, exploratory evidence from call timing is directionally consistent with the segmented-call interpretation, though it should be read as suggestive rather than definitive given the reduced power of the split sample and the endogeneity of timing choice discussed in Section 4.5. When earnings calls occur during trading hours, negative Q&A tone shows a relatively more precise and consistent association with abnormal returns. When calls take place before the market opens, the relative importance of the prepared introduction increases. To our knowledge, no prior study has examined whether call timing moderates the relative informativeness of segment-specific tone in this way, though the present evidence motivates rather than establishes this possibility. One interpretation consistent with this pattern is that the Q&A exchange functions as a live information channel whose impact depends on whether investors can respond while the discussion is unfolding, whereas prepared remarks matter more when investors process the full transcript with a delay; this interpretation is offered as a plausible reading of a descriptive pattern, not as an established mechanism.

Our contribution is fourfold. First, we provide the first segment-level analysis of earnings-call tone and short-window market reactions in Brazil, offering new evidence from an emerging-market context grounded in specific institutional predictions. Brazil differs from the U.S. in analyst coverage depth, ownership concentration, and the composition of its investor base in ways that generate testable predictions about where within the call tone is most informative. We show that the Q&A effect is more precisely and consistently estimated in lower-coverage calls, directionally consistent with the view that the unscripted exchange carries more incremental information when pre-call analyst research is limited. This distinguishes the paper from Tonin and Scherer (2021), who study unsegmented tone in Brazil, and from Manabe et al. (2022), who examine medium-to-long-term returns in Japan; neither study tests the institutional predictions examined here.

Second, we show that the internal segmentation of the call has first-order importance for measuring tone's informativeness. Although formal tests show that the economic magnitudes of negative tone are statistically indistinguishable across call segments, negative Q&A tone yields a statistically significant and consistently negative association with short-window abnormal returns, reaching conventional statistical significance under FinBERT, while tone in the prepared introduction, despite similar coefficient magnitudes in some specifications, is consistently less precise and less robust across estimators. This finding implies that the choice of *which segment* to measure matters as much as the choice of sentiment method, a distinction rarely examined in the literature.

Third, the call-timing analysis provides exploratory evidence consistent with this segmentation pattern, rather than establishing an underlying mechanism. Q&A negativity shows a relatively more precise and consistent association with abnormal returns when calls occur during trading hours, when investors can process information as the discussion unfolds; call timing is not randomly

assigned, however, and Section 4.5 discusses this limitation directly. This result is directionally consistent with the Q&A segment functioning as a live information channel rather than a scripted summary, and it has direct implications for understanding how disclosure timing shapes investor reactions.

Fourth, by comparing Loughran–McDonald, VADER, and FinBERT as competing measurement approaches in an emerging-market context, we contribute to the growing evidence on sentiment method selection in financial text analysis. The finding that FinBERT produces more precise and stable associations in this sample suggests that domain-adapted language models may offer better empirical fit than lexicon-based alternatives in this setting, while the paper deliberately refrains from claiming general model superiority in the absence of transcript-level validation.

The empirical design is intentionally reduced-form. Because earnings calls occur close to earnings releases and other firm disclosures, the analysis does not identify a causal effect of tone alone. Instead, the paper studies whether tone in different parts of the call is associated with investor reactions around the broader disclosure event. This framing is important: the results do not imply that linguistic tone independently drives market outcomes, but they do indicate that unscripted negative tone is an informative component of how investors process earnings-call disclosures.

The remainder of the paper is organised as follows. Section 2 reviews the related literature and develops the hypotheses. Section 3 describes the data, tone measurement, and empirical design. Section 4 presents the results. Section 5 concludes.

2. Related literature and hypothesis development

2.1. Earnings calls as disclosure events

Earnings conference calls occupy a distinctive position in the corporate disclosure environment. Unlike mandatory written disclosures, which follow standardised formats and are released on a fixed schedule, earnings calls unfold in real time and combine a structured managerial presentation with an unscripted exchange between management and analysts. This dual structure means that they can carry information beyond the numerical content of the accompanying earnings release (Beyer et al., 2010; Healy & Palepu, 2001), while also creating conditions in which less rehearsed managerial communication may surface.

The empirical record consistently supports the view that earnings calls convey information that markets price. Studies initiated by Frankel et al. (1999) and extended by Bowen et al. (2002) and Bushee et al. (2003) show that the initiation of conference calls is associated with narrowing information asymmetry, tighter bid–ask spreads, and a reduction in analyst forecast dispersion. Price et al. (2012) document that textual tone predicts abnormal returns and trading volume around call dates. Kimbrough (2005) finds that calls accelerate how quickly analysts incorporate earnings information into forecasts. Matsumoto et al. (2011) further show that the two structural segments of the call differ in informational content, with analyst participation adding incremental explanatory power beyond the prepared introduction.

This evidence motivates the central premise of the present study: earnings calls should not be treated as homogeneous disclosure events. Instead, their internal structure may matter for how investors interpret and respond to the information conveyed during the call.

2.2. Scripted and unscripted communication

The distinction between prepared remarks and question-and-answer (Q&A) discussions is central to this paper. Prepared remarks are relatively scripted and closely aligned with the firm’s intended communication strategy. They often summarise information already available in the earnings release and provide management with an opportunity to frame recent performance. By contrast, the Q&A segment is more interactive and less controllable, because analysts can raise concerns, request clarification, or press management on issues that may not have been emphasised voluntarily.

Prior evidence suggests that this interactive component matters for market reactions. Price et al. (2012) show that conference-call tone predicts abnormal returns and trading volume, and that the Q&A portion has incremental explanatory power. Matsumoto et al. (2011) likewise show that the two segments of the call differ in informational content, with analyst participation increasing the usefulness of the event. Brockman et al. (2015) further show that managers and analysts use systematically different tones, and that market participants react more strongly to analyst tone than to managerial tone. Together, these findings suggest that the unscripted portion of the call may reveal information that is not fully contained in the prepared narrative.

The implication is that tone in the Q&A segment may be especially informative. If analyst questions elicit less rehearsed responses, or if the tone of the exchange reflects the difficulty of the issues being raised, investors may attach greater weight to Q&A tone than to tone in the prepared introduction. This motivates the paper’s segmented-call design. Consistent with this view, Manabe et al. (2022) show, using Japanese data, that the Q&A session at earnings conference calls contains more useful information than the management presentation for medium-to-long-term returns. The present study complements this evidence by focusing on short-window market reactions (the $[0, +1]$ window) and by comparing segment-specific tone across three sentiment methods in an emerging-market setting where such evidence has been largely absent.

2.3. Tone, linguistic cues, and market reactions

A second relevant strand of literature studies the linguistic properties of corporate disclosure (Davis et al., 2015; Davis & Tama-Sweet, 2012). Research using textual analysis finds that tone, readability, complexity, and related linguistic characteristics can convey information not fully captured by accounting numbers alone (Chen et al., 2023). In the earnings-call setting, tone may reflect managerial confidence (Gow et al., 2021), uncertainty, defensiveness, or the credibility of disclosure. Hennig et al. (2023) and Price et al. (2012) document that earnings-call tone predicts both abnormal returns and trading volume. Druz et al. (2020) show that stock prices react to the negativity of managers' language, and interpret this reaction as broadly consistent with the rational use of tone-related information.

Complementing the textual evidence, Mayew and Venkatachalam (2012) show that vocal characteristics of managers and analysts during earnings calls contain information beyond what is captured in the transcript text. More broadly, research outside the earnings-call setting shows that non-verbal cues can also convey economically meaningful information about firm conditions (Aromi & Clements, 2021; Fiset et al., 2021). This evidence reinforces the broader point that communication channels beyond formal accounting numbers can be informative to capital-market participants.

Recent studies also examine linguistic features of earnings calls beyond simple positive-versus-negative tone. Bushee and Huang (2024), for example, show that managerial linguistic complexity during earnings calls contains information that analysts incorporate more efficiently than investors do. Park et al. (2025) show that financial statement comparability shapes managers' linguistic choices in earnings calls, linking the properties of formal reporting to subsequent disclosure style. Together, these studies reinforce the idea that earnings calls are not merely a venue for repeating earnings numbers; they are a structured but flexible setting in which language itself may convey economically relevant information.

Because negative language may reveal concerns about performance, uncertainty, risk, or managerial defensiveness, the first empirical prediction is that more negative tone should be associated with less favourable short-window price reactions.

Hypothesis 1. More negative tone in earnings conference calls is associated with lower short-window abnormal returns.

The segmented structure of the call then leads to a second prediction. If the Q&A segment is less scripted and more responsive to analyst pressure, the association between tone and abnormal returns should be stronger in the Q&A segment than in the prepared introduction.

Hypothesis 2. The association between tone and short-window abnormal returns is stronger in the Q&A segment than in the prepared introduction.

While economic theory predicts a stronger segment-level association, we evaluate this empirically by testing both the formal equality of coefficient magnitudes and differences in estimation precision across segments.

2.4. Measuring tone in financial text

A third strand of literature concerns how tone should be measured. The accounting and finance literature has relied heavily on dictionary-based approaches, especially (Loughran & McDonald, 2011), because finance-specific dictionaries reduce misclassification relative to generic lexicons. Hutto and Gilbert (2014) provide a more flexible rule-based alternative that captures negation, emphasis, and intensity, although it was not designed specifically for financial text.

More recently, domain-adapted language models have been proposed to improve contextual classification. Araci (2019) and Huang et al. (2023) show that FinBERT improves sentiment classification in financial text relative to standard dictionary approaches and other machine-learning methods, and argue that conventional approaches may understate the informativeness of earnings-call text. At the same time, stronger classification performance in labelled tasks does not imply that every disclosure application should be framed as an artificial-intelligence methods contribution. In this paper, alternative sentiment tools are treated as competing measurement approaches for operationalising tone, not as the primary theoretical contribution.

Beyond the three approaches used here, a methodologically distinct alternative assigns each term a continuous valence score ranging from -1 (maximally negative) to $+1$ (maximally positive), so that document-level tone is measured as the weighted average of word-level scores rather than as a proportion of binary polarity counts. Ito et al. (2018) develop such a dictionary for Japanese financial text, and Manabe et al. (2022) apply it in their analysis of Japanese earnings conference calls. The English-language literature contains analogous efforts: García et al. (2023) construct continuous word-level sentiment scores for financial text derived from stock-price reactions, and show they outperform binary bag-of-words approaches in out-of-sample stock-return prediction. Continuous-score methods offer a potentially richer representation of tone intensity and may be more sensitive to nuanced language in interactive settings such as Q&A exchanges. However, no such instrument has been validated for the Brazilian earnings-call context, and constructing one would require either a large labelled corpus or a market-calibration exercise outside the scope of the present study. Developing and applying continuous-score tone measures to Brazilian earnings calls is therefore identified as a productive avenue for future research.

Two indirect validation checks provide further evidence that FinBERT produces economically meaningful classifications on Brazilian English-language transcripts. First, FinBERT Q&A negativity is significantly higher in quarters with earnings misses (EPS below consensus) than in quarters with earnings beats. Specifically, the mean negative-sentence ratio is 0.0493 for earnings misses ($n = 850$) compared with 0.0454 for earnings beats ($n = 986$). A Welch two-sample t -test confirms that this difference

is statistically significant ($t(1771) = 2.92$, $p = 0.004$). Loughran–McDonald and VADER show no such discrimination ($p = 0.76$ and $p = 0.73$ respectively). If the FinBERT advantage reflected idiosyncratic activation on non-native English patterns rather than genuine sentiment capture, it would not systematically identify calls with more negative fundamental outcomes. Second, FinBERT Q&A negativity peaks in the second quarter of 2020 (mean 0.068, approximately 44% above the full-sample average), consistent with elevated managerial stress during the height of the COVID-19 shock, a pattern not replicated as strongly by the lexicon-based measures. These two checks are complementary to, and do not substitute for, the formal manual-annotation exercise reported in Section 4.9; together they indicate that FinBERT's classifications are not purely artefactual in this setting and, per the results of that formal exercise, do not reflect idiosyncratic behaviour specific to the non-native, translated English of this corpus.

An important caveat in comparing these approaches concerns the domain specificity of existing validation evidence. The Loughran and McDonald (2011) dictionary was developed and validated on 10-K annual filings, which are more formal and standardised than the conversational language of earnings calls. VADER (Hutto & Gilbert, 2014) was designed and benchmarked on social media text, making its application to financial transcripts a further extrapolation. FinBERT was fine-tuned on financial news and validated on the Financial PhraseBank and FiQA sentiment datasets (Araci, 2019), which consist of short analyst statements and news snippets rather than the longer, interactive discourse of earnings calls. Crucially, none of these validation benchmarks include transcript-level annotations for earnings-call segments, which means that the reported classification advantages of any one method cannot be assumed to generalise to the present context. The current study's transcripts are in English (see Appendix C), which preserves linguistic applicability, but the absence of an in-sample validation benchmark remains a limitation that the paper explicitly acknowledges throughout.

This measurement issue is increasingly relevant because textual disclosure is now processed by both human and machine readers. Cao et al. (2023) show that growing AI readership affects how firms write disclosures by encouraging language that is more machine-friendly and less negatively perceived by algorithms. dos Reis Maia and Bravo (2026) document that algorithmic and AI-based processing of earnings-call disclosures shapes investor decisions in the Brazilian capital market, suggesting that the linguistic properties of transcripts have become consequential beyond their human-readable content. This literature provides a contemporary motivation for including a finance-domain language model such as FinBERT as one of the paper's tone measures, while preserving the interpretation of the exercise as a measurement comparison rather than a test of model superiority.

2.5. Emerging-market setting

The Brazilian setting differs from the U.S. along several institutional dimensions that generate distinct predictions about the informativeness of segmented disclosure, rather than simply providing a replication exercise in a different country.

Analyst coverage. Brazilian listed firms are covered by substantially fewer sell-side analysts than comparable U.S. firms, and analyst participation in earnings calls averages fewer than six analysts per event in this sample. With limited pre-call information production, the Q&A session may be the primary venue through which investors learn management's responses to specific concerns. The association between Q&A tone and market reactions is therefore predicted to be more informative for lower-coverage firms, a prediction directly testable in this sample and examined in Section 4.6.

Concentrated ownership and dual-class shares. Brazilian public companies frequently have controlling shareholders, often family groups or state entities, and many employ dual-class share structures separating voting from economic rights. For controlling insiders, the Q&A session may convey little incremental information; for minority shareholders, institutional investors, and foreign participants, however, the unscripted exchange may be the most direct window into management's private beliefs about current conditions.

Market microstructure. The Ibovespa operates on a single continuous trading session, supplemented by opening and closing auctions. While the regular session duration is comparable to U.S. equity markets, market microstructure differs substantially due to lower liquidity, higher transaction costs, and greater concentration of trading activity. As a result, price discovery around earnings calls is more compressed, potentially amplifying the within-call contrast between scripted and unscripted segments.

Taken together, these features generate the exploratory expectation that beyond the average tone–return association, the Q&A tone relation may be more evident in calls with lower analyst engagement and in calls held during trading hours. Both patterns are examined below.

Evidence from outside the United States is growing but remains sparse. Tonin and Scherer (2021) show that tone in Brazilian calls predicts short-window returns. Manabe et al. (2022) show, using Japanese data, that the Q&A session contains more information than the management presentation for medium-to-long-term returns. Yamamoto et al. (2022) study tone strategies in global equity markets. Yang et al. (2023) find that conference tone predicts post-earnings drift in China. The present study extends this evidence by jointly testing the analyst-coverage and call-timing predictions in the Brazilian setting with explicit call segmentation and a multi-method sentiment comparison.

2.6. Tone and abnormal trading volume

Trading volume reflects several mechanisms simultaneously: information arrival, investor disagreement about interpretation, heterogeneous prior beliefs, liquidity, and portfolio-adjustment needs. Because these mechanisms can drive volume in either direction, the sign and magnitude of any tone–volume association is theoretically ambiguous. Rather than imposing a directional prediction, the paper treats abnormal trading volume as a complementary exploratory outcome. The question is not whether tone

Table 1
Sample construction and data sources.

Step	Observations
Earnings-call transcripts matched to fundamental and event-study data	1836
Less: missing abnormal return or volume estimates	(159)
Earnings calls with complete event-study data	1677
Less: missing fundamental control variables	(37)
Final regression sample	1640
Unique firms	77
Sample period (fiscal quarters)	4Q 2016–4Q 2025 (37 quarters)

Notes. The sample covers Brazilian publicly traded firms listed on B3. Fundamental data (total assets, price-to-book, leverage, volatility, EPS surprise) and transcripts are sourced from Koyfin. Abnormal returns and volumes are estimated over a two-day event window $[0, +1]$ using a Fama–French three-factor model with a 60-trading-day estimation window $[-62, -2]$. Individual earnings calls span from January 2017 to February 2026, reflecting the reporting lag between fiscal quarter-end and the earnings call. Regression tables with both firm and year fixed effects report $N = 1638$ after excluding two firm-year singleton cells (see Section 3.6).

should affect volume, but whether the Q&A-informativeness pattern documented for returns extends to trading activity. The analysis in Section 4.3 reports the evidence and interprets the results descriptively.

Taken together, the literature motivates two empirical expectations that are formulated as testable hypotheses: negative tone should be associated with less favourable short-window abnormal returns ([Hypothesis 1](#)), and this association should be stronger in the Q&A segment than in the prepared introduction ([Hypothesis 2](#)). Abnormal trading volume is analysed as a secondary outcome without a pre-specified directional prediction.

3. Data, tone measurement, and empirical design

This section describes the sample, the construction of the tone measures, and the empirical design used to relate earnings-call language to short-window market reactions. The objective is not to identify a causal effect of conference-call tone, but to test whether tone, especially in the Q&A segment, is associated with *CAR* and *CAV* around call dates. Because earnings releases, conference calls, and other firm disclosures are often closely timed, the event-study outcomes should be interpreted as reduced-form responses to the broader disclosure event rather than as clean estimates of the standalone effect of conference-call language.

3.1. Sample construction

The empirical analysis focuses on Brazilian listed firms in the Ibovespa universe ([B3, 2026](#)). The initial sample comprises earnings conference-call transcripts from 77 firms over 4Q 2016–4Q 2025, obtained from [Koyfin \(2024\)](#). Daily returns and trading volume are collected from Yahoo Finance ([Aroussi, 2023](#)), and expected returns are estimated using Brazilian Fama–French factors ([Fama & French, 1993](#)) from the Núcleo de Pesquisa em Economia Financeira/Financiadora de Estudos e Projetos (NEFIN/FINEP) ([NEFIN, 2024](#)). The merged dataset links transcript-level tone measures to event-level market outcomes using firm identifiers and call dates.

Two design decisions bound the sample. First, to avoid biases from time-varying index composition, the cross-section is fixed to the 77 Ibovespa constituents as of December 2025 and analysed over the full sample period; this static design eliminates rebalancing-induced shifts in composition but tilts the sample towards firms that remain persistently listed and liquid. The panel is unbalanced: 44 firms have transcript data from 2017, growing to 68 by 2024 as firms are added to the index or transcript coverage becomes available; among the 1677 firm-quarters with a valid event-study estimate (see [Table 1](#)), the average number of calls per firm is 21.8 (min. 1, max. 36). Critically, firms that were Ibovespa constituents between 2017 and 2024 but left the index before December 2025 are excluded, creating a form of backward-looking survivorship bias. If firms that exited the index tend to have performed worse on average, the main results may understate the true tone–return association, since the excluded firms would likely contribute observations with simultaneously high Q&A negativity and low abnormal returns. While an ideal dynamic constituent design based on quarter-by-quarter Ibovespa composition reports would fully eliminate this concern, such a reconstruction is constrained by historical transcript and sentiment data availability. Instead, Section 4.7 implements an expanded historical-constituent robustness check incorporating 14 former index constituents active during 2017–2024, mitigating rather than eliminating survivorship bias. Second, low-liquidity periods are excluded from the *CAR* and *CAV* analyses to ensure that estimated market reactions reflect information dynamics rather than trading frictions, which can be substantial in an emerging-market setting. Together, these decisions imply that the results apply primarily to persistent, liquid Brazilian listed firms rather than the broader population of Brazilian public companies.

Sample construction proceeds as follows. After merging transcripts with fundamental and event-study data, 1836 firm-quarter observations are retained. Event windows are defined using a Fama–French three-factor framework; 159 observations are subsequently excluded because of missing abnormal return or volume estimates. A further 37 observations are dropped because of missing fundamental control variables, yielding a final regression sample of 1640 firm-quarter observations across 77 firms. [Table 1](#) summarises the construction steps.

Table 2
Transcript-language provenance.

Category	Observations	%	Firms
Host call in English	542	33.0%	25
Host official translated simultaneous English calls	1098	67.0%	52
Total	1640	100%	77

Notes. Sample: 1640-observation regression sample (77 firms, 4Q2016–4Q2025). Classification is based on the authors' direct review of the earnings-call webcast for each of the 77 sample firms. Earnings call webcasts are either conducted directly in English or delivered via an official webcast with simultaneous English translation.

3.2. Transcript segmentation

Each transcript is divided into two sections: prepared remarks, which cover the managerial presentation prior to the question-and-answer exchange, and the Q&A segment, which covers the interactive portion of the call from the first analyst question to the close. This division is motivated by the economic distinction between structured disclosure and unscripted interaction, rather than by a presumption about which segment is more informative.

Segmentation proceeds in two steps. First, an algorithm scans each transcript for a set of operator transition phrases that signal the opening of the Q&A session. The phrases used as anchors include variations of: “*we will now begin the question-and-answer session*”, “*let's open the floor for questions*”, “*I will now hand the call over to the operator*”, and “[*operator instructions*]”. The anchor is used to split the transcript at the sentence boundary immediately preceding the first analyst question. Second, all 1836 transcripts are reviewed manually and corrected where the algorithmic split misidentifies the transition point, primarily in cases where introductory remarks extend beyond the operator handover or where non-standard phrasing is used.

All transcripts analysed in this study are in English, and the implications of this choice for measurement validity are addressed directly here rather than left implicit. Brazilian Ibovespa-listed firms host English-language earnings calls as their primary investor relations channel for global capital markets. We describe these conservatively as official, company-provided English-language investor communications, prepared, reviewed, and approved by each company's investor relations and executive teams before publication. The authors directly reviewed the earnings-call webcast for every one of the 77 sample firms to determine how each company's call is delivered in English. Table 2 reports the share of firm-quarters in which the call itself was conducted directly in English, as distinct from firm-quarters delivered through an official webcast with simultaneous English translation.

Of the 1640 regression-sample observations, 542 (33.0%, 25 firms) come from companies confirmed, through the authors' review, to host the call itself directly in English. The remaining 1098 observations (67.0%, 52 firms) come from an official webcast with simultaneous English translation; consistent with the conservative framing above, we describe these as official, company-provided English-language investor communication. For this group, the interpreter may exercise minor word-level delivery choice when rendering the call live, particularly during the unscripted Q&A exchange, although these interpreters are engaged directly by the company and are typically briefed in advance on firm- and industry-specific terminology and executives' word choices. As a further, sentence-level check on whether this interpretation channel measurably affects the reliability of the tone measure itself, Section 4.9 reports a manual-annotation exercise in which FinBERT's classification accuracy on this hybrid corpus, sampled unfiltered across both delivery types, is compared with an entirely native-English comparison sample. The two samples' accuracy against annotators' judgement (76.2% versus 71.6%) and inter-annotator agreement ($\kappa = 0.899$ versus 0.904) are broadly comparable, providing no clear evidence of materially poorer FinBERT performance in the Brazilian corpus. This evidence bears on measurement validity specifically and is complementary to, rather than a substitute for, the economic-identification limitation discussed in Section 4.9. The use of English-language investor relations transcripts is both a methodological choice and a contribution. Methodologically, it ensures full compatibility with standard sentiment tools: the Loughran–McDonald dictionary is built on English financial disclosures, VADER is designed for English text, and FinBERT is trained on English financial news. Using native-language inputs avoids the cross-linguistic transfer issues that would arise from applying these models to Portuguese originals. Substantively, this setting remains unexplored in the Brazilian earnings-call literature. Prior work relies exclusively on Portuguese-language materials (Tonin & Scherer, 2021). This English-language channel reflects a distinct communicative register and audience (the prepared remarks are authored by executives in both English and Portuguese), and captures an alternative dimension of firms' disclosure practices that has not been previously analysed in Brazil. The two approaches are complementary: Portuguese-based studies capture domestic communication, while ours focuses on the international investor-facing channel. According to official Bovespa data, foreign investment trading has accounted for more than 50% of total trading volume over the past five years (B3 Investimentos, 2026).

The Q&A segment is treated as a single unit; analyst questions and management responses are not separated, because the objective is to measure the overall interactive tone of the call rather than to attribute sentiment to individual participants. Text preprocessing, including uniform encoding, removal of formatting artefacts, tokenisation, and sentence-boundary detection, is implemented in Python using the Natural Language Toolkit (NLTK). Full preprocessing details are provided in Appendix C.

3.3. Tone measurement

Tone is measured using three sentiment approaches applied independently to each transcript segment. Using multiple approaches allows the analysis to examine whether the segment-informativeness finding is robust to the choice of measurement method, rather

Table 3
Tone measures.

Variable	Lexicon	Definition
Panel A: Raw frequency ratios		
neg_ratio	LM/VADER/FinBERT	Proportion of negative words to total words in the segment.
pos_ratio	LM/VADER/FinBERT	Proportion of positive words to total words in the segment.
Panel B: Negativity index (Druz et al., 2020).		
negativity	LM/VADER/FinBERT	$(N - P) / (N + P + 1)$, where N and P are the counts of negative and positive words (or sentence proportions) respectively. Bounded on $[-1, 1]$; higher values indicate more negative net tone. The +1 regularises the index when sentiment-bearing words are sparse.
Panel C: Segment subscripts		
_intro	–	Variable computed on the prepared-remarks segment of the call.
_qna	–	Variable computed on the Q&A segment of the call.

Notes. LM denotes the Loughran-McDonald (Loughran & McDonald, 2011) financial dictionary. VADER (Hutto & Gilbert, 2014) is a rule-based general-purpose sentiment lexicon. FinBERT (Huang et al., 2023) is a transformer-based model fine-tuned on financial text. For LM and VADER, word classification follows the respective dictionary/rule set applied to lowercased, punctuation-stripped tokens. For FinBERT, sentence-level probabilities are aggregated to the segment level and the majority class is assigned per sentence before computing segment-level ratios.

than being an artefact of any single tool. The three methods differ fundamentally in their unit of analysis, their treatment of context, and their domain of development, which is why their comparison is informative.

Loughran-McDonald (LM). Tone is measured at the word level using the 2011 Loughran-McDonald master dictionary (Loughran & McDonald, 2011), which was constructed from 10-K filings and classifies words into positive and negative financial sentiment categories. For each segment, the positive (negative) ratio is the count of words matching the LM positive (negative) list divided by the total number of words in that segment. Stopwords and numerals are excluded from the denominator. This word-level approach captures the density of sentiment-bearing vocabulary but does not account for word order, negation, or context.

VADER. Tone is measured at the sentence level using the Valence Aware Dictionary and sEntiment Reasoner (Hutto & Gilbert, 2014). Each sentence in the segment is scored and classified as positive, negative, or neutral based on the VADER compound score and its associated thresholds. The positive (negative) ratio for a segment is the proportion of sentences classified as positive (negative) out of the total number of sentences. This sentence-level approach preserves some contextual information within sentences (including negation and emphasis modifiers) but does not capture cross-sentence dependencies.

FinBERT. Tone is measured at the sentence level using the ProsusAI/FinBERT model checkpoint (Araci, 2019; Huang et al., 2023), a transformer-based language model fine-tuned on financial text. Each sentence is passed through the model and assigned a three-way classification: positive, negative, or neutral. The positive (negative) ratio for a segment is the share of sentences receiving a positive (negative) classification out of total sentences. Unlike the dictionary and rule-based approaches, FinBERT captures contextual and semantic dependencies within each sentence through attention mechanisms, allowing it to handle domain-specific phrasing and implicit sentiment.

For all three methods and both segments, the baseline regressions include the positive and negative ratios as separate regressors. This specification allows positive and negative language to have asymmetric associations with market outcomes, which is consistent with evidence that negative tone tends to be more informative than positive tone in financial disclosure settings (Druz et al., 2020; Loughran & McDonald, 2011).

In addition to the raw ratios, the analysis reports a negativity index following (Druz et al., 2020):

$$\text{Negativity}_{i,t} = \frac{N_{i,t} - P_{i,t}}{N_{i,t} + P_{i,t} + 1}, \quad (1)$$

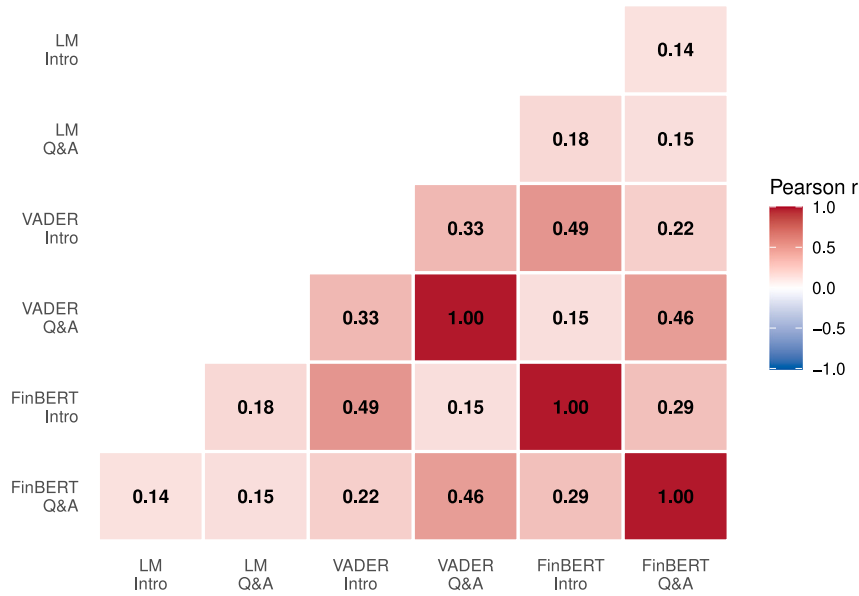
where $N_{i,t}$ and $P_{i,t}$ are the negative and positive counts (word counts for LM; sentence counts for VADER and FinBERT) for firm i in quarter t . The index is bounded in $(-1, 1)$, is scale-invariant with respect to segment length, and mitigates length-induced heteroskedasticity. The addition of one in the denominator prevents division by zero in segments with no sentiment-bearing language. A positivity index, $(P - N)/(P + N + 1)$, would be the arithmetic inverse and is therefore excluded from the regressions to avoid perfect collinearity.

Table 3 provides a complete variable definition.

Fig. 1 illustrates the pairwise Pearson correlations among the six negative-tone variables, defined by the three sentiment methods and the two call segments. Two patterns are notable. First, within-method correlations between the prepared remarks and Q&A segments are modest, ranging from 0.14 to 0.49. This is consistent with the view that the two segments carry distinct sentiment signals. Second, cross-method correlations are generally low, particularly between the LM dictionary and the other two measures, indicating that the three approaches capture related but non-identical aspects of financial language.

3.4. Event-study outcomes

The dependent variables are *CAR* and *CAV* around earnings-call dates. Both are computed over the two-day event window $[0, +1]$, where day 0 is the call date and day +1 is the following trading day. This window is short enough to focus on immediate

Figure 1 – Pairwise Correlations: Negative Tone MeasuresPearson correlations across lexicons and call segments · $n = 1,836$ observations**Fig. 1.** Pairwise Pearson correlations among negative-tone variables across sentiment methods (LM, VADER, FinBERT) and call segments (prepared remarks and Q&A). $n = 1836$ observations.

market reactions but wide enough to capture adjustment when calls occur after the close or when investors process transcripts with a one-day delay. Abnormal returns and volumes are estimated using EventStudy Tools (<https://www.eventstudytools.com>), with a minimum of 30 non-missing trading days required in the estimation window for an observation to be retained. Observations failing this threshold account for the reduction from 1836 to 1677 usable event-study observations prior to further sample restrictions.

For CAR, expected returns are estimated using a Fama–French three-factor model. Model parameters are fitted via ordinary least squares (OLS) over a 60-trading-day estimation window $[-62, -2]$ relative to the call date, using the Ibovespa as the market factor and the Brazilian Small-Minus-Big (SMB) and High-Minus-Low (HML) factors sourced from NEFIN/FINEP (NEFIN, 2024).

The choice of a 60-day estimation window is motivated by two considerations. First, because Brazilian Ibovespa firms report quarterly, the gap between consecutive earnings calls is approximately 63 trading days. A longer window, for example, 120 or 200 trading days, would routinely extend into the estimation period of the preceding quarterly call, contaminating the parameter estimates with tone and return effects from adjacent disclosure events. The 60-day window $[-62, -2]$ avoids this overlap while still providing a sufficient number of observations for reliable OLS estimation. Second, shorter estimation windows are less susceptible to structural breaks and regime changes, which are particularly relevant in Brazil given the macroeconomic volatility over the 4Q 2016–4Q 2025 sample period. A robustness check using a 120-trading-day estimation window $[-122, -2]$ with the same Fama–French three-factor specification is reported in Appendix D (column (6)); Q&A negative tone remains significant at the 1% level ($\hat{\beta} = -0.129$, $p < 0.01$), confirming that the main finding is not sensitive to window length.

The abnormal return on each event day is the residual from the Fama–French model, and CAR is the sum of daily abnormal returns over $[0, +1]$. The full model specification and parameter equations are given in Appendix B.

For CAV, daily firm-level trading volume is normalised by shares outstanding to obtain a log-turnover measure, $\ln(V_{i,t}/S_{i,t})$. A market-model regression of firm log-turnover on the log of mean Ibovespa daily turnover is estimated over the same $[-62, -2]$ window. Abnormal volume on each event day is the residual from this model, and CAV is its sum over $[0, +1]$. First-order autocorrelation in firm-level volume is accounted for in the estimation. Summary statistics for both outcome variables are reported in Table 4.

Calls are classified according to their timing relative to the trading day: before market open (10:00 Brasília Time, BRT), during trading hours, or after market close (17:00 BRT). This classification is used solely in the heterogeneity analysis of Section 4.5; the baseline event window $[0, +1]$ is kept constant across all observations to maintain comparability across timing groups.

3.5. Empirical specification

The baseline empirical model relates event-level market outcomes to segment-level tone measures:

$$Y_{i,t} = \beta_1 \text{IntroPos}_{i,t} + \beta_2 \text{IntroNeg}_{i,t} + \beta_3 \text{QnaPos}_{i,t} + \beta_4 \text{QnaNeg}_{i,t} + \gamma' \mathbf{X}_{i,t} + \alpha_i + \lambda_t + \varepsilon_{i,t}. \quad (2)$$

where $Y_{i,t}$ is either CAR or CAV for firm i in quarter t ; $\text{IntroPos}_{i,t}$ and $\text{IntroNeg}_{i,t}$ denote positive and negative tone in the prepared remarks; and $\text{QnaPos}_{i,t}$ and $\text{QnaNeg}_{i,t}$ denote positive and negative tone in the Q&A segment. The vector $\mathbf{X}_{i,t}$ contains firm-level controls, including earnings surprise, firm size, price-to-book, leverage, and return volatility. Firm fixed effects, α_i , absorb time-invariant differences across firms, while year fixed effects, λ_t , absorb common macroeconomic and market-wide conditions. The intercept is absorbed by the firm fixed effects.

Standard errors are clustered by firm (ticker) throughout the main tables, following standard practice for firm-level panel regressions (Cameron et al., 2011; Petersen, 2009). Two-way clustering by firm and call date is evaluated in the robustness section (Table A.2, column (7)). Baseline results are subjected to six robustness checks: models are estimated with and without fixed effects; a lagged CAR term is added as a momentum control (reducing the sample to 1354 observations); CAV is winsorised at the 1st and 99th percentiles; CAR is replaced with buy-and-hold abnormal returns; extreme return observations are excluded; and full-call sentiment is compared to the segmented specification. Throughout, the design tests whether tone is *associated* with market reactions, not whether it causes them; because earnings calls coincide with earnings releases, results should be read as reduced-form associations within a broader disclosure event.

3.6. Estimation and statistical inference

Sample. The regression sample comprises 1640 firm-quarter observations spanning 77 Brazilian Ibovespa-listed firms over 37 quarterly reporting periods from 4Q 2016 to 4Q 2025 (individual call dates extend into early 2026 for the final quarter, reflecting the typical one-to-three-month reporting lag between fiscal quarter-end and the earnings call). The average number of observations per firm is 21.3 (minimum 1, maximum 36), reflecting variation in listing continuity and data availability over the sample period. All regression models estimated with both firm and year fixed effects exclude two firm-year cells that are singletons in this two-way structure, reducing the effective estimation sample from 1640 to 1638 observations; this adjustment applies uniformly across the primary regression specifications (Tables 3–7) and is not repeated in each table's notes. All models are estimated using the full regression sample unless otherwise stated; the lagged-CAR specification reduces the sample from 1640 to 1354 observations. The reduction of 286 observations is related to the sample attrition from introducing the one-quarter lag stems from three constraints: 72 firm-quarters represent initial panel entries without an observable prior period; 212 observations lack the immediately preceding fiscal quarter due to non-consecutive coverage; and a further 2 observations are excluded as fixed-effects singletons, the same adjustment described above.

Estimation method. All models are estimated by ordinary least squares with the within-estimator for fixed effects, implemented using the `fixest` package in R (Bergé, 2018). Firm and year fixed effects are projected out using the Guimarães–Portugal algorithm (Guimarães & Portugal, 2010), which is numerically equivalent to the within-transformation but substantially more efficient for large fixed-effect dimensions.

Inference. Statistical significance is assessed through t -tests on individual regression coefficients under the null of no association. Three significance thresholds are used throughout: $p < 0.10$ (*), $p < 0.05$ (**), and $p < 0.01$ (***). These thresholds are indicated in all regression tables; figures display 95% confidence intervals, with intervals that exclude zero corresponding to significance at the 5% level. Standard errors are clustered by firm (ticker) across all primary specifications, following Cameron et al. (2011) and Petersen (2009). In our sample, call-date clusters exhibit severe sparsity: the median number of firms reporting on the same calendar day is two, 39% of call dates are singletons, and 75% contain three or fewer firms. Under such finite-sample imbalances, within-date dependence cannot be estimated reliably, and the asymptotic properties required for multiway cluster-robust variance estimation break down (Cameron & Miller, 2015). Moreover, common aggregate shocks are already absorbed by year fixed effects (Petersen, 2009). Nonetheless, to establish a conservative boundary, we report two-way clustered standard errors (by firm and call date) in Table A.2, column (7). Under this specification, the coefficient on Q&A negativity remains directionally consistent and economically meaningful, with statistical significance shifting from the 5% to the 10% level (standard error of 0.116 versus 0.090 in the baseline model). Given that date-level noise inflation is the expected artefact of sparse cluster structures rather than unmodelled time dependence, we treat firm-clustered standard errors as the econometrically sound baseline, while confirming that our primary Q&A inference remains robust at the 10% level under this more restrictive alternative. No correction for multiple comparisons is applied; the analysis tests a small number of pre-specified hypotheses based on the theoretical arguments in Section 2, and a Bonferroni-type correction would not materially alter the main conclusions given that the primary finding is based on a single key coefficient (QnaNeg).

Replication and robustness. This is a single observational study based on one sample of Brazilian listed firms over one historical period. The research design is, however, directly replicable using the same methodology: event-study abnormal returns, earnings call segmentation, and FinBERT-based sentiment scoring, can be applied to other emerging markets with listed-firm earnings calls or to different time periods in Brazil. Internal temporal stability is examined via a subperiod analysis (pre-COVID, COVID/turbulent, and post-2022). The subperiod coefficients are directionally consistent with the full-sample result in two of the three windows but none is individually significant, and the pattern cannot distinguish a stable association detectable only with the full nine-year sample from one concentrated in the COVID/turbulent window; this is discussed as an open question in Section 4.7 rather than treated as established temporal stability. The absence of external replication across markets remains a genuine boundary on generalisability and is explicitly acknowledged in Section 4.9. Replication across other emerging markets with active earnings call practices represents a natural avenue for future research.

Table 4
Descriptive statistics.

	N	Mean	SD	Min	P25	Median	P75	Max
CAR [0, +1] (%)	1677	−0.17	6.30	−35.20	−3.11	−0.14	2.55	108.00
CAV [0, +1]	1677	0.97	1.06	−14.19	0.40	0.93	1.52	7.48
FinBERT: Intro positive ratio	1836	0.32	0.09	0.00	0.25	0.32	0.38	0.62
FinBERT: Intro negative ratio	1836	0.08	0.04	0.00	0.05	0.07	0.10	0.35
FinBERT: Q&A positive ratio	1836	0.19	0.06	0.00	0.15	0.19	0.23	0.50
FinBERT: Q&A negative ratio	1836	0.05	0.03	0.00	0.03	0.04	0.06	0.20
LM: Intro positive ratio	1836	0.02	0.01	0.00	0.02	0.02	0.03	0.07
LM: Intro negative ratio	1836	0.02	0.01	0.00	0.01	0.02	0.02	0.07
LM: Q&A positive ratio	1836	0.02	0.01	0.00	0.02	0.02	0.03	0.07
LM: Q&A negative ratio	1836	0.04	0.01	0.00	0.03	0.04	0.04	0.14
VADER: Intro positive ratio	1836	0.20	0.04	0.09	0.18	0.21	0.23	0.36
VADER: Intro negative ratio	1836	0.04	0.02	0.00	0.03	0.04	0.05	0.12
VADER: Q&A positive ratio	1836	0.21	0.04	0.09	0.18	0.20	0.23	0.55
VADER: Q&A negative ratio	1836	0.04	0.02	0.00	0.03	0.04	0.05	0.12
EPS surprise (BRL)	1836	0.00	0.18	−4.92	−0.01	0.00	0.02	1.35
Total assets (BRL bn)	1827	174.78	456.60	0.91	15.18	32.97	65.37	3066.17
Price-to-book	1822	2.60	3.07	−25.69	1.08	1.83	3.07	43.10
Leverage (net debt/assets)	1827	0.28	0.20	0.00	0.12	0.28	0.42	0.92
Annualised volatility (%)	1812	35.75	16.23	2.66	25.18	32.07	42.33	165.40

4. Results

This section reports the empirical results on the association between earnings-call tone and short-window market reactions. The analysis focuses on two outcomes: *CAR* and *CAV* around call dates. The primary tests evaluate whether negative tone is associated with lower abnormal returns and whether this relation is stronger in the unscripted Q&A segment than in the prepared remarks. The main evidence concerns abnormal returns; abnormal trading volume is analysed as a complementary outcome because volume reflects both information processing and disagreement among investors.

4.1. Descriptive evidence

Table 4 reports descriptive statistics. The transcript data show substantial variation in call length and in the distribution of positive and negative language. The three sentiment approaches do not classify tone in identical ways: FinBERT and VADER tend to assign positive language more frequently than the Loughran–McDonald dictionary, which is more conservative in assigning polarity. This pattern is consistent with the three methods capturing related but non-identical aspects of financial language. There is also meaningful heterogeneity in tone across call segments, with the Q&A consistently showing different distributional properties from the prepared introduction; this variation is the empirical basis for the segment-informativeness tests that follow.

The sample sizes of 1836 and 1677 correspond to the initial matched sample and the completed event-study sample reported in Table 4, respectively. Variations in observation counts across individual control variables arise from specific data availability: 9 observations lack total assets ($N = 1827$), 14 lack the price-to-book ratio ($N = 1822$), and 24 lack annualised volatility data ($N = 1812$).

4.2. Baseline evidence on abnormal returns

Table 5 reports regressions of $CAR_{[0,+1]}$ on tone measures from earnings calls. Column (1) includes only sentiment ratios, column (2) adds firm-level controls, column (3) introduces year fixed effects, and column (4) includes both firm and year fixed effects with firm-clustered standard errors and is the preferred specification. Column (5) further controls for lagged *CAR*. Two-way clustering is reported in column (7) of Table A.2.

Across the main specifications, the most consistent result is that negative tone in the Q&A segment is associated with lower short-window abnormal returns. In the preferred specification (Table 5, column (4)), the Q&A negative-tone coefficient reaches conventional statistical significance ($\hat{\beta} = -0.215^{**}$, $p = 0.020$), whereas the prepared-remarks coefficient, despite being virtually identical in point magnitude ($\hat{\beta} = -0.215$, $p = 0.157$), does not. A formal Wald test of coefficient equality confirms that the two point estimates are statistically indistinguishable ($\hat{\beta}_{Intro} - \hat{\beta}_{Q\&A} = -0.0005$, $t = -0.003$, $p = 0.997$). Consequently, the empirical evidence does not support Hypothesis 2: unscripted Q&A negative tone does not exert a larger economic impact on abnormal returns than prepared-remarks tone. Rather, what emerges across specifications is an empirical pattern of greater estimation precision and cross-model stability for the Q&A segment, characterised by narrower confidence intervals and lower residual noise.

To assess economic magnitude, the preferred coefficient ($\hat{\beta} = -0.215$; Table 5, column (4)) can be evaluated at the empirical distribution of Q&A negativity in the sample. The mean FinBERT Q&A negative ratio is 4.7% of sentences, with a standard deviation of 2.9 percentage points. A one-standard-deviation increase in Q&A negativity, equivalent to moving from roughly the 50th to the 84th percentile of the within-sample distribution and corresponding to an increase in the fraction of negative sentences from approximately 4.7% to 7.6%, is associated with a reduction in two-day *CAR* of approximately $0.215 \times 0.029 = 0.006$, or **62 basis**

Table 5
Baseline regressions: CAR on prepared-remarks and Q&A tone (FinBERT).

	(1)	(2)	(3)	(4)	(5)
Intro positive ratio	−0.015 (0.021)	−0.012 (0.025)	−0.011 (0.024)	0.005 (0.039)	−0.006 (0.046)
Intro negative ratio	−0.130 (0.086)	−0.128 (0.085)	−0.169 (0.112)	−0.215 (0.151)	−0.226 (0.177)
Q&A positive ratio	−0.049 (0.046)	−0.046 (0.044)	−0.052 (0.050)	−0.042 (0.069)	−0.070 (0.082)
Q&A negative ratio	−0.059 (0.057)	−0.055 (0.061)	−0.045 (0.075)	−0.215** (0.090)	−0.227** (0.106)
Lagged CAR					−0.019 (0.012)
EPS surprise (winsorised)		0.047* (0.025)	0.045* (0.025)	0.031 (0.027)	0.033 (0.034)
Log total assets		0.000 (0.002)	0.001 (0.002)	0.004 (0.014)	0.014 (0.020)
Price-to-book (winsorised)		−0.001 (0.001)	−0.001 (0.001)	−0.002 (0.002)	−0.001 (0.003)
Leverage		0.005 (0.014)	0.006 (0.014)	0.057 (0.040)	0.071 (0.056)
Log volatility		−0.000 (0.006)	−0.000 (0.006)	0.010 (0.006)	0.011 (0.011)
Year FE	No	No	Yes	Yes	Yes
Firm FE	No	No	No	Yes	Yes
Observations	1640	1640	1640	1638	1354
R ²	0.002	0.003	0.008	0.028	0.030

Notes. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Column (4) is the baseline.

points. An interquartile-range increase (from the 25th to the 75th percentile) implies a reduction of approximately **75 basis points**. These magnitudes represent roughly 11% of the CAR interquartile range in the sample (566 basis points) and are broadly consistent with segment-level tone effects documented in Brockman et al. (2015), who report manager–analyst tone differences on the order of 50–100 basis points in U.S. settings. The effect is economically meaningful but modest relative to total short-window return variation, consistent with tone being one of several components of the broader disclosure event.

Although the formal equality test indicates that Hypothesis 2 is not supported in terms of differential economic magnitude, the greater statistical precision and cross-specification stability of the Q&A estimate represent an informative empirical pattern. One interpretation is that analyst questioning elicits less rehearsed managerial responses, thereby providing a cleaner empirical signal that investors regard as relevant for valuation. A more cautious interpretation is that Q&A tone reflects the difficulty or salience of issues raised during the call, which may themselves be correlated with underlying firm news. The empirical design cannot fully distinguish between these mechanisms. The results should therefore be interpreted as reduced-form associations rather than causal effects.

Positive-tone coefficients are less stable: some specifications show a positive association between Q&A positive tone and abnormal returns, but the pattern is inconsistent across estimators, in line with prior evidence that negative signals carry disproportionate weight in financial communication (Tetlock, 2007). The overall explanatory power of the return regressions remains modest even in the most saturated specifications, which is expected in short-window event studies where many contemporaneous factors influence returns; the appropriate reading is that negative Q&A tone is one component of the broader disclosure event that is systematically associated with market reactions, not that tone alone drives prices.

4.3. Evidence on abnormal trading volume

Table 6 reports the corresponding CAV regressions. Columns (1)–(4) mirror the CAR Table 5. Column (5) applies two-way clustering, and column (6) further controls for lagged CAV. The evidence for abnormal trading volume is weaker and less stable than for abnormal returns: some specifications suggest that negative Q&A tone is associated with higher abnormal trading activity, but sign and precision vary more across sentiment measures and estimators. This fragility is economically plausible because trading volume reflects not only the arrival of new information but also heterogeneity in investor beliefs, disagreement about interpretation, liquidity conditions, and attention effects, a richer set of determinants that tone alone cannot reliably predict. The CAV analysis is therefore treated as complementary evidence rather than an equally central test.

4.4. Comparison across tone measures

Table 7 isolates the contribution of each call segment to the CAR associations documented in Table 5. Column (I) includes prepared-remarks tone only, column (II) includes Q&A tone only, and column (III) includes both jointly. All specifications use raw positive and negative ratios with firm and year fixed effects and standard errors clustered by ticker. Comparing columns (I)–(III) shows whether each segment carries independent information or whether one segment dominates when both are included.

Table 6
Baseline regressions: CAV on prepared-remarks and Q&A tone (FinBERT).

	(1)	(2)	(3)	(4)	(5)	(6)
Intro positive ratio	1.160*** (0.300)	1.140*** (0.313)	0.853*** (0.316)	0.421 (0.351)	0.421 (0.303)	0.325 (0.355)
Intro negative ratio	−0.424 (0.801)	−0.394 (0.809)	0.025 (0.785)	−0.891 (0.697)	−0.891 (0.733)	−0.574 (0.833)
Q&A positive ratio	0.277 (0.471)	0.297 (0.462)	0.185 (0.457)	−0.452 (0.483)	−0.452 (0.524)	−0.341 (0.584)
Q&A negative ratio	0.840 (1.055)	0.882 (1.103)	0.704 (1.061)	0.532 (0.965)	0.532 (1.148)	0.468 (1.154)
Lagged CAV						0.011 (0.039)
EPS surprise (winsorised)		−0.388 (0.276)	−0.295 (0.264)	−0.243 (0.222)	−0.243 (0.215)	−0.137 (0.222)
Log total assets		0.003 (0.023)	0.005 (0.023)	0.176 (0.135)	0.176 (0.129)	0.227 (0.147)
Price-to-book (winsorised)		−0.005 (0.012)	0.011 (0.013)	−0.000 (0.020)	−0.000 (0.024)	0.012 (0.026)
Leverage		−0.073 (0.242)	−0.117 (0.228)	0.057 (0.419)	0.057 (0.380)	0.259 (0.375)
Log volatility		−0.049 (0.084)	0.077 (0.098)	−0.295*** (0.087)	−0.295*** (0.088)	−0.331*** (0.098)
Year FE	No	No	Yes	Yes	Yes	Yes
Firm FE	No	No	No	Yes	Yes	Yes
Observations	1640	1640	1640	1638	1638	1354
R ²	0.015	0.017	0.043	0.164	0.164	0.163

Notes. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 7
Segment and metric sensitivity: CAR regressions (FinBERT).

	(I) Intro only	(II) Q&A only	(III) Both
Intro positive ratio	0.002 (0.042)		0.005 (0.039)
Intro negative ratio	−0.244 (0.156)		−0.215 (0.151)
Q&A positive ratio		−0.039 (0.070)	−0.042 (0.069)
Q&A negative ratio		−0.273** (0.109)	−0.215** (0.090)
EPS surprise (winsorised)	0.032 (0.027)	0.033 (0.026)	0.031 (0.027)
Log total assets	0.004 (0.014)	0.005 (0.014)	0.004 (0.014)
Price-to-book (winsorised)	−0.002 (0.002)	−0.002 (0.002)	−0.002 (0.002)
Leverage	0.049 (0.035)	0.058 (0.041)	0.057 (0.040)
Log volatility	0.009 (0.006)	0.007 (0.006)	0.010 (0.006)
Year FE	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes
Observations	1638	1638	1638
R ²	0.027	0.026	0.028

Notes. Dep. var.: CAR [0, +1] (decimal). Firm + year FE. SE clustered by ticker. (I): prepared-remarks tone only. (II): Q&A tone only. (III): both segments jointly. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table 8 brings the three sentiment measures together in a single comparison. Because the three methods operate at different units of aggregation: LM at the word level, FinBERT at the sentence level with classified probabilities, and VADER at the sentence level with compound scores, the raw ratios are not on comparable scales and raw coefficient comparisons across methods would be uninterpretable. Table 8 therefore reports *standardised* sentiment variables: each measure is z-scored prior to estimation, so that a one-unit change equals one standard deviation within each method. The resulting coefficients represent the change in CAR associated with a one-SD increase in tone, and are directly comparable across methods. Figs. 2 and 3 visualise the raw coefficient estimates from the segment-specific tables as forest plots.

The standardised results reveal a pattern that is obscured by raw coefficients. In SD-unit terms, the three methods yield negative Q&A coefficients of similar order: FinBERT (−0.0063**), Loughran–McDonald (−0.0082) and VADER (−0.0063). Only

Table 8

Cross-method comparison: CAR and CAV regressions (FinBERT, Loughran–McDonald, VADER).

	FinBERT	LM	VADER
Panel A: Cumulative Abnormal Returns (CAR [0, +1])			
Intro positive (std)	0.000 (0.004)	−0.004 (0.005)	0.004 (0.008)
Intro negative (std)	−0.010 (0.007)	−0.015 (0.011)	−0.014 (0.009)
Q&A positive (std)	−0.003 (0.004)	−0.007 (0.005)	−0.006 (0.009)
Q&A negative (std)	−0.006** (0.003)	−0.008 (0.006)	−0.006 (0.005)
Year FE	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes
Observations	1638	1638	1638
R ²	0.028	0.034	0.032
Panel B: Cumulative Abnormal Volume (CAV [0, +1])			
Intro positive (std)	0.039 (0.032)	0.094** (0.037)	0.045 (0.040)
Intro negative (std)	−0.039 (0.031)	0.016 (0.032)	−0.046 (0.034)
Q&A positive (std)	−0.029 (0.031)	−0.046 (0.029)	−0.013 (0.026)
Q&A negative (std)	0.016 (0.028)	−0.004 (0.032)	0.028 (0.034)
Year FE	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes
Observations	1638	1638	1638
R ²	0.164	0.167	0.165

Notes. All sentiment variables are standardised (z-scored) so that a one-unit change equals one standard deviation within each method, making coefficients directly comparable across methods that operate at different units of aggregation (word level for LM; sentence level for FinBERT and VADER). Dep. var.: Panel A, CAR [0, +1]; Panel B, CAV [0, +1]. Both call segments included. Firm + year FE, full controls. SE clustered by ticker.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

FinBERT reaches conventional significance, but the directional consistency across methods, compared with the large raw-coefficient differences (−0.215, −0.621 and −0.392), confirms that the scale disparity reflects measurement units rather than differences in the informational content each method captures. This directly addresses the concern that cross-method coefficient comparisons are uninterpretable (Huang et al., 2023; Hutto & Gilbert, 2014; Loughran & McDonald, 2011).

The cross-method comparison therefore does not support the claim that FinBERT's larger raw coefficient indicates greater informativeness. What it does support is that FinBERT produces a more precisely estimated Q&A negativity effect: the standardised Q&A negative coefficient is significant at the 5% level for FinBERT ($p = 0.020$) but not for Loughran–McDonald ($p = 0.183$) or VADER ($p = 0.222$). This precision advantage is the empirical pattern to be explained, and is not itself independent evidence that FinBERT provides superior measurement: defining “better fit” by a tighter standard error and then citing that same standard error as evidence of better fit would be circular. Section 2.4 reports two checks that are independent of the CAR regression itself (discrimination between earnings-miss and earnings-beat quarters, and an elevated negativity peak during the Covid-19 shock) as separate, non-circular evidence that FinBERT's classifications track real variation in this corpus; these checks are indirect and do not amount to a formal validation exercise, and FinBERT's relative precision in the return regressions is treated as sample-specific empirical fit rather than demonstrated classification superiority. The underlying effect size is similar across methods. The cross-method comparison therefore does not establish that Q&A negativity is significant under every measurement approach.

The standardised comparison provides further direct evidence regarding Hypothesis 2. For FinBERT, the standardised intro-negativity coefficient ($\hat{\beta}_{\text{Intro, std}} = -0.010$, $p = 0.157$) is larger in absolute point estimate than the standardised Q&A-negativity coefficient ($\hat{\beta}_{\text{Q\&A, std}} = -0.006$, $p = 0.020$). A formal Wald test of equality between the two standardised coefficients, accounting for their covariance in the clustered variance-covariance matrix, fails to reject the null (difference = −0.0032, SE = 0.0063, $t = -0.513$, $p = 0.608$). Coupled with the test on raw coefficients ($p = 0.997$), this result formally confirms that Hypothesis 2 is not supported: unscripted Q&A tone does not have a larger economic association with returns. Instead, the evidence establishes an empirical pattern of greater estimation precision and stability for the unscripted segment.

Two pieces of evidence temper this concern. First, all six Q&A negative-tone estimates across the three methods and two segment specifications (Q&A-only and joint) are negative. Because these specifications are estimated on the same sample of 1638 calls and employ nested model structures, these estimates are mechanically correlated rather than statistically independent. Consequently, we interpret this uniform sign pattern strictly as descriptive evidence of directional consistency across alternative NLP methodologies, rather than as a formal joint hypothesis test. Second, the imprecision of Loughran–McDonald and VADER is consistent with the domain-specificity argument in Section 2.4: both tools were validated on corpora that differ substantially from earnings-call

Figure 2 – Negativity Index: Prepared Remarks vs. Q&A

negativity . (neg – pos) / (neg + pos + 1)

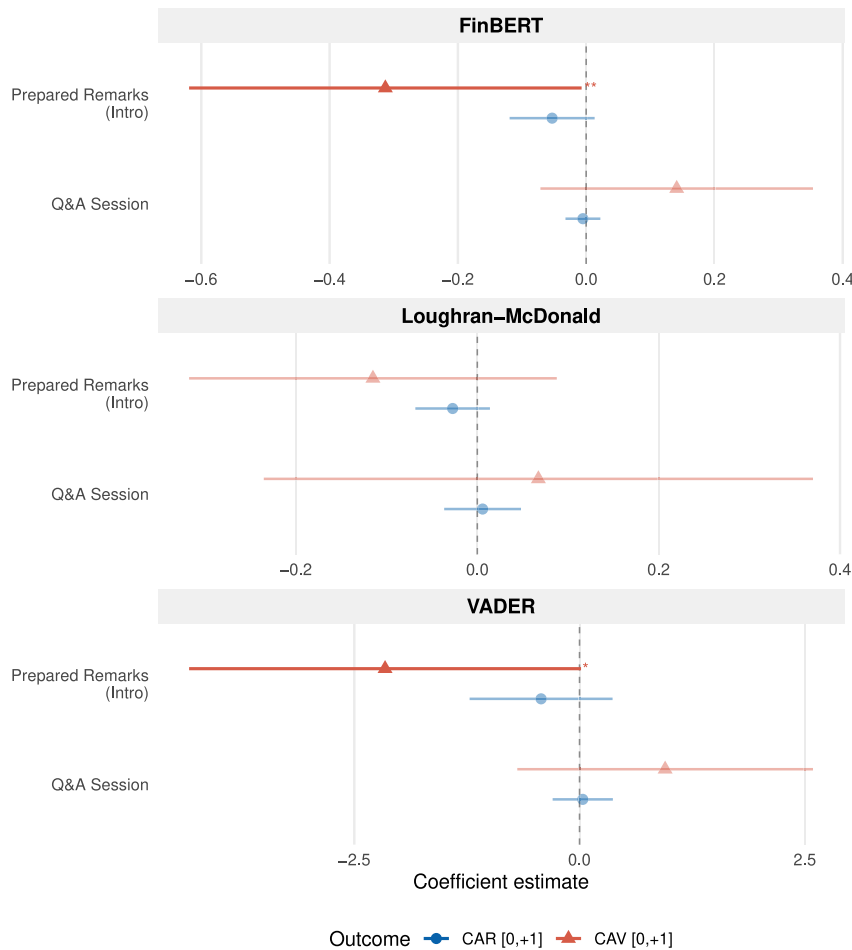
Firm + year FE · 95% CI · Clustered SE by ticker · faded = $p \geq 0.10$ 

Fig. 2. Coefficient estimates (95% CI) for the negativity index $[(N - P) / (N + P + 1)]$ across sentiment methods (FinBERT, Loughran–McDonald, VADER) and call segments (prepared remarks, Q&A). Blue circles: CAR [0, +1]; red triangles: CAV [0, +1]. All specifications include firm and year fixed effects; standard errors are clustered by ticker. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

transcripts in register and length, making them less reliable classifiers in this specific context. FinBERT’s advantage here reflects better empirical fit in this sample, as argued throughout, not a post-hoc selection of a favourable measure.

The low within- R^2 values (around 0.006 in the preferred specification) are consistent with the broader short-window event-study literature, in which disclosure tone explains a small fraction of total return variation by design (Price et al., 2012). The lagged-CAR specification (column (5) of Table 5) uses 17% fewer observations due to gaps between consecutive quarterly calls yet yields a slightly larger and still significant coefficient (-0.227^{**} , $p = 0.035$), confirming that the Q&A negativity effect is not an artefact of omitting momentum controls. Across all three methods, evidence for prepared-remarks tone is limited once firm fixed effects are included, consistent with the segmented-call interpretation.

These results should not be interpreted as evidence that FinBERT is universally superior to dictionary-based methods. As noted in Section 2.4, existing validation studies for all three methods rely on benchmarks, 10-K filings for Loughran and McDonald (2011), social media for Hutto and Gilbert (2014), and short financial news snippets for Araci (2019), that differ materially from earnings-call transcripts in register, length, and conversational structure. The transcript-level manual-annotation exercise reported in Section 4.9 validates FinBERT’s own classifications against annotators’ judgement on this corpus and on a native-English comparison corpus, but it does not extend to LM or VADER: the current design does not include a labelled benchmark or out-of-sample prediction exercise that scores all three methods against a shared set of human-annotated labels. Accordingly, the paper does not claim that FinBERT “outperforms” the alternative models in a machine-learning sense. A more appropriate interpretation is that FinBERT provides a tone measure that is more closely aligned with short-window market reactions in this particular setting, consistent with contextual language models capturing nuances in earnings-call language, especially in the less formulaic Q&A segment, that static dictionaries

Figure 3 – Sentiment Ratios: Prepared Remarks vs. Q&A

Pos./neg. ratios · Firm + year FE · 95% CI · Clustered SE by ticker
 Shaded = Q&A · faded = $p > 0.10$

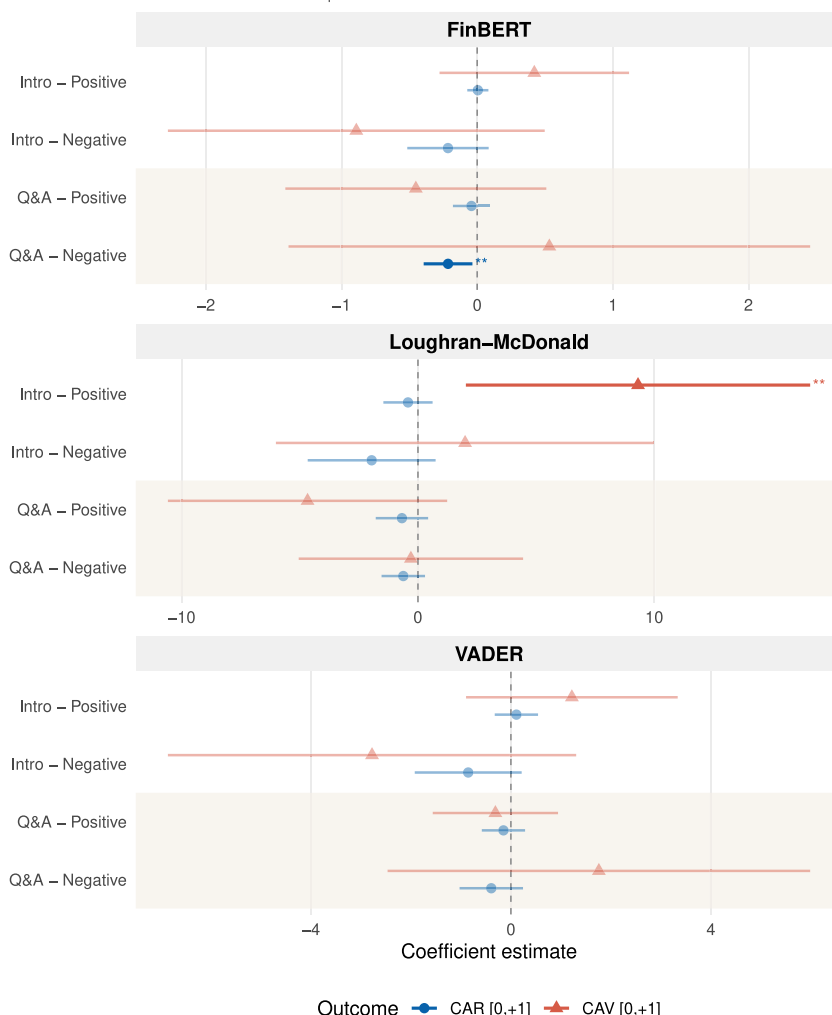


Fig. 3. Coefficient estimates (95% CI) for raw positive and negative tone ratios across sentiment methods and call segments. Blue circles: CAR [0,+1]; red triangles: CAV [0,+1]. Shaded rows correspond to Q&A segment variables. All specifications include firm and year fixed effects; standard errors are clustered by ticker. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

may miss. The contribution remains a disclosure finding with a measurement comparison, not a demonstration of general model superiority.

4.5. Heterogeneity by call timing

The timing of the earnings call relative to the trading day provides a natural source of heterogeneity. Calls held while the market is open allow investors to adjust positions as the discussion unfolds. Calls held before the market opens are processed before trading begins. This distinction may affect whether the scripted introduction or the unscripted Q&A segment carries greater weight in price formation.

To examine this possibility, the sample is split into calls that begin before market open ($N = 683$) and calls that begin during trading hours ($N = 898$). The after-market group ($N = 13$) is too small for reliable inference and is excluded. The same specification as column (III) in Table 7, using FinBERT tone measures, both call segments, and firm and year fixed effects, is estimated separately for each subsample. Results are reported in Table 9.

Three important caveats apply before reading the timing results. First, the split reduces the powered sample: relative to the Q&A negative-tone standard error in the full-sample specification ($SE = 0.090$; Table 5, column (4)), the before-market subsample ($N = 683$) implies a standard error approximately 148% larger, and the during-market subsample ($N = 898$) implies an increase

Table 9
Tone and abnormal returns by call timing — FinBERT.

	(I) Before market	(II) During market
Intro positive	−0.095 (0.066)	0.097* (0.058)
Intro negative	−0.223** (0.098)	−0.261 (0.290)
Q&A positive	0.031 (0.097)	−0.078 (0.101)
Q&A negative	−0.274 (0.223)	−0.297*** (0.104)
EPS surprise	0.013 (0.034)	0.037 (0.054)
Log total assets	−0.015 (0.010)	0.019 (0.026)
Price-to-book	−0.005* (0.003)	−0.001 (0.003)
Leverage	0.109** (0.047)	0.024 (0.066)
Log volatility	0.003 (0.010)	0.021** (0.010)
Log analysts	−0.019 (0.021)	−0.040 (0.033)
Year FE	Yes	Yes
Firm FE	Yes	Yes
Observations	683	898
Within R^2	0.023	0.014

Notes. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. Dep. var.: CAR [0, +1] (decimal). Firm + year FE. Standard errors clustered by ticker (firm-only), the baseline choice. Columns sum to 1581 of the 1640-observation regression sample: 33 observations lack call-time metadata within the matching window used to classify timing (a nearest-date bridge of up to two days, following the same procedure used to attach call-time metadata elsewhere in the sample); of the remaining 1607, 13 after-market calls are excluded as too few for reliable inference (leaving 1594, matching the sample underlying the timing-determinants logit in Section 4.5), and a further 13 observations (6 before-market, 7 during-market) are excluded as firm-year fixed-effects singletons once the sample is split by timing and each subsample is re-estimated with its own firm and year fixed effects.

of approximately 15%. The timing estimates should therefore be treated as exploratory rather than definitive. Second, call timing is not close to randomly assigned, and the before-market group is not a negligible category that can be set aside. Among the 1594 observations for which call-time metadata is available, domestic firms show no significant average preference between the two timing categories (49.3% intraday; not significantly different from an even split, $p = 0.75$), whereas ADR-listed firms lean significantly towards intraday scheduling (62.1%; $p < 0.001$ against an even split; $\chi^2 = 26.3$, $p < 0.001$ for the domestic-ADR difference). The absence of an average preference among domestic firms is not, on its own, reassuring: it is consistent either with individual firms having stable, idiosyncratic timing preferences that happen to balance out in aggregate, or with firms varying their timing choice across quarters. To distinguish these possibilities, we examined within-firm variation directly: among the 73 firms with at least three timed calls, 63 hold calls both before and during market hours across different quarters, and only 10 use a single timing category throughout. Call timing therefore varies substantially within firm, not only across firm types, which limits, without eliminating, the extent to which the timing comparison reduces to a comparison of fixed firm characteristics already absorbed by firm fixed effects. Cross-listing status and other unobserved firm characteristics correlated with scheduling choice remain a first-order concern for interpreting the timing comparison below, not a minor caveat. Third, we test directly below whether tone itself predicts timing choice, which addresses the sharpest version of this concern for the paper's main result, though it cannot rule out confounding from other, unobserved sources of timing variation.

With these caveats, the pattern across the two subsamples is as follows. For calls held during trading hours, Q&A negativity is associated with lower abnormal returns: the coefficient on Q&A negative tone is -0.297 ($p < 0.01$). Prepared-remarks tone is generally not significant in this subsample. For calls held before market open, Q&A negativity is no longer statistically significant (-0.274 , $p > 0.10$), whereas prepared-remarks negativity is significant (-0.223 , $p < 0.05$). While this contrast provides suggestive evidence consistent with the idea that real-time trading may shape the precision of segment estimates, a formal test qualifies this interpretation: a continuous interaction of Q&A negative tone with the intraday indicator, estimated on the pooled sample, yields an interaction coefficient of -0.014 ($p = 0.82$). As with the analyst-coverage and ownership splits above, the during-market versus before-market contrast is directionally consistent with the point estimates reported for the two subsamples but is not statistically distinguishable at conventional levels. However, given the power reduction and the compositional differences documented above, this interpretation is suggestive rather than definitive. Beyond the compositional evidence already documented, timing may be related to firm characteristics, investor-base composition, or disclosure strategy through several plausible channels. For instance, larger firms with broader international investor bases may tend to hold intraday calls to accommodate real-time analyst participation;

firms expecting unfavourable reactions may prefer pre-market slots to allow time for investor processing before trading begins. If such systematic patterns exist, the timing comparison could partly reflect underlying cross-firm differences rather than the mechanical effect of trading-hour availability.

To assess the severity of this concern, a logit model was estimated on the 1594 observations in Table 9 for which timing classification is available (i.e., the same before-market and during-market subsamples underlying the split above, excluding only the 13 after-market calls), with the dependent variable equal to one if the call begins during trading hours and zero if it begins before market open. The regressors include the fundamental controls from the main specification (EPS surprise, log total assets, price-to-book, leverage, log volatility), the two main tone measures (FinBERT Q&A negative ratio and FinBERT prepared-remarks negative ratio), and year fixed effects. Two results are presented. First, Q&A negative tone does not significantly predict intraday timing ($\hat{\beta} = -0.69$, $p = 0.72$), suggesting that firms are not systematically choosing intraday slots in response to anticipated negative Q&A exchanges. This directly mitigates the most important version of the endogeneity concern: the tone variable driving the main result is not correlated with the timing split. Second, prepared-remarks negative tone is positively and significantly associated with intraday timing ($\hat{\beta} = 3.97$, $p = 0.003$), suggesting that firms with more negative introductory content tend to hold calls during market hours. One interpretation is that firms expecting analyst scrutiny may prefer intraday slots to allow real-time engagement; another is that the introduction reflects information whose market-hour processing is valued by management. This correlation between prepared-remarks tone and timing choice is itself evidence that call timing is not exogenous, and it means the before-market/during-market comparison may partly reflect this compositional difference rather than a pure effect of timing on segment informativeness; it does not overturn the Q&A result, which holds within each subsample, but it is a real limitation of the timing comparison rather than a finding that can be set aside.

Firm fixed effects further absorb time-invariant firm-level timing preferences, and the fundamental controls reduce the scope for differences in observable firm characteristics to confound the comparison. The timing results are accordingly framed as descriptive evidence consistent with real-time information processing rather than as causal estimates.

Taken together, the timing results provide a suggestive descriptive pattern rather than a clean causal estimate. The Q&A segment shows a more precise and consistent association with abnormal returns in the intraday subsample, and the prepared introduction shows a more precise association in the before-market subsample. This pattern is consistent with the view that real-time trading during the call shapes the informativeness of different segments (Frankel et al., 1999). Given the power limitations, compositional differences between subsamples, and the endogeneity of timing choice documented above, these results should be read as exploratory evidence that motivates further research on disclosure timing and segment informativeness.

4.6. Heterogeneity by analyst coverage

Section 2.5 predicts that Q&A tone should be more informative in calls with lower analyst engagement, where the interactive segment reveals more incremental information to the market. This prediction is tested using two complementary specifications. First, a cross-sectional split at the median analyst count (six analysts per call). Second, a within-firm deviation split that asks whether calls with above-average analyst participation for a given firm show a different pattern from calls with below-average participation. The second design controls for all time-invariant firm-level differences by exploiting variation in analyst turnout across a firm's own earnings calls.¹

Both designs reveal a directional pattern. In the cross-sectional split, the Q&A negative tone coefficient is $\hat{\beta} = -0.271^{**}$ ($p = 0.017$; $N = 956$) for lower-engagement calls and $\hat{\beta} = -0.213$ ($p = 0.230$; $N = 672$) for higher-engagement calls, (10 observations missing from the total of $N = 1638$ observations). In the within-firm deviation split, the coefficient is $\hat{\beta} = -0.242^{**}$ ($p = 0.037$; $N = 827$) when a call draws below-average participation for that firm, and $\hat{\beta} = -0.247$ ($p = 0.146$; $N = 808$) when it draws above-average participation, (3 observations missing from the total of $N = 1638$ observations). A continuous interaction of Q&A negative tone with the within-firm analyst deviation produces an interaction coefficient of -0.094 ($p = 0.46$). The non-significant interaction suggests that analyst participation does not create a sharp moderation effect, and the split results should therefore be interpreted as directional patterns rather than statistically different regimes.

The consistent pattern across both designs is that the Q&A negativity coefficient is statistically significant precisely when analyst engagement is lower, and imprecise when it is higher. The difference lies in standard errors rather than magnitudes, both groups exhibit coefficients near -0.25 to -0.28 , suggesting that the Q&A tone–return relationship is present in both regimes but is more reliably estimated when the call draws less analyst scrutiny. This is consistent with the information-gap interpretation from Section 2.5: in calls with limited analyst engagement, the unscripted Q&A represents a cleaner channel for price discovery. When more analysts participate, their pre-call information gathering may partly pre-empt the signal content of the Q&A, producing a noisier relationship between tone and same-day returns. A second heterogeneity test uses share class as a proxy for ownership concentration and dual-class structures, both of which are prevalent in the Brazilian institutional setting. Firms listed as preferred shares (PN, ticker suffix 4 or 5) typically have controlling shareholders who retain ordinary voting shares (ON) while issuing non-voting preferred shares to the public. Controlling insiders already have superior informational access, so the Q&A session should carry less incremental information for PN-share investors than for investors in firms with more dispersed ON-share ownership.

¹ The analyst variable measures the number of sell-side analysts who participated in that specific call. This is a call-level measure rather than a firm-level I/B/E/S following count; it has a cross-sectional correlation of 0.17 with log market capitalisation and exhibits within-firm variation in 99% of firms (mean within-firm standard deviation: 1.81 analysts), making it informative both across and within firms.

Splitting the sample by share class yields a directional pattern consistent with this conjecture, although the formal interaction test qualifies this interpretation. For the 66 ON-share firms ($N = 1247$), the Q&A negative tone coefficient is $\hat{\beta} = -0.230^{**}$ ($p = 0.025$). For the 11 PN-share firms ($N = 184$), the coefficient is near zero ($\hat{\beta} = +0.002$, $p = 0.987$) (209 observations missing from the total $N = 1640$, corresponding to unit shares). The contrast is directionally consistent with the concentrated-ownership conjecture: when controlling shareholders hold preferential informational access through their ON-share position, the unscripted Q&A may carry less incremental information for the PN-share market. The PN-share subsample is small and these firms differ on other dimensions (sector, size), so this result should be read as an exploratory pattern rather than definitive evidence. A formal test highlights this need for caution: a continuous interaction of Q&A negative tone with the PN-share indicator, estimated on the pooled sample, yields an interaction coefficient of 0.211 ($p = 0.215$). As with the analyst-coverage split, the descriptive contrast between subsamples is therefore not statistically distinguishable at conventional levels, and the ON/PN comparison should be read as a directional pattern consistent with the concentrated-ownership channel rather than evidence of a statistically distinct regime. Examining heterogeneity by directly measured ownership stakes and foreign institutional holdings remains a direction for future research.

A third heterogeneity test uses cross-listing status identified from the J.P. Morgan ADR programme database. ADR listing is used here as a proxy for *foreign investor accessibility and international capital market integration* rather than for direct foreign ownership: firms with active ADR programmes have formally made themselves accessible to international institutional capital and typically attract greater foreign investor participation, but the binary indicator does not capture the degree of foreign ownership. Of the 77 sample firms, 28 have listed ADRs on U.S. markets (primarily Level I and Level II). The Q&A negative tone coefficient is $\hat{\beta} = -0.277$ ($p = 0.232$; $N = 756$) for ADR-listed firms and $\hat{\beta} = -0.139$ ($p = 0.177$; $N = 882$) for domestically listed firms only. Neither estimate is statistically significant, and the split roughly halves the sample. A formal test of the difference between these two independently estimated coefficients, using their reported standard errors, does not reject equality ($z = -0.54$, $p = 0.586$). As with the analyst-coverage and ownership splits above, the ADR contrast should therefore be read as directional only, consistent with the hypothesis that internationally integrated firms attract more foreign institutional participation, rather than as evidence of a statistically distinct regime.

Two exploratory proxies help shed light on foreign investor accessibility without requiring direct ownership data. ADR status captures the formal channel through which firms open themselves to international capital; market capitalisation captures size-driven visibility through which foreign funds allocate capital via passive index inclusion (e.g., MSCI Emerging Markets) and liquidity thresholds. The large-cap versus small-cap robustness check in Section 4.7 shows negative Q&A tone coefficients in both groups (-0.197 for small-cap, -0.224 for large-cap), with the small-cap estimate reaching conventional significance ($p = 0.045$) and the large-cap estimate imprecisely estimated ($p = 0.241$), consistent with greater information asymmetry for lower-coverage firms rather than a pure market-cap-driven visibility effect. Combining these two descriptive tests provides suggestive insights into the international-visibility channel, though direct ownership data would be required for definitive tests. Obtaining direct measures of foreign institutional ownership from CVM periodic filings would allow a more precise test and is identified as a direction for future research.

4.7. Robustness tests

The baseline results are subjected to eight robustness checks, reported in Table A.2 in Appendix D. Column (1) replaces CAR with buy-and-hold abnormal returns (BHAR) over the same $[0, +1]$ window. Column (2) replaces the prepared-remarks/Q&A split with aggregate full-call FinBERT sentiment, testing whether segmentation is necessary to detect the association. Column (3) excludes the six observations with $|CAR| > 3\sigma$. Column (4) adds the log of total transcript sentences as a control for call length. Column (5) adds the number of sell-side analysts participating in the call. Column (6) re-estimates the baseline using a 120-trading-day estimation window $[-122, -2]$. Column (7) repeats the baseline with two-way clustered SE (ticker $\times$ call date) as a conservative clustering robustness check. Columns (8) and (9) provide an expanded historical-constituent robustness check that mitigates, rather than eliminates, survivorship bias, re-estimating a tone-only specification on the current 77-firm sample and an expanded 91-firm historical sample respectively.

The Q&A negative ratio remains negative and statistically significant in columns (3) and (5), with coefficients of -0.149 ($p < 0.05$) and -0.221 ($p < 0.01$) respectively, the last corresponding to the 120-trading-day estimation window. In column (1), the coefficient is negative (-0.438) but imprecisely estimated under BHAR, consistent with the noisier nature of buy-and-hold aggregation over short windows. Column (2) does not include a Q&A-specific variable by construction; the aggregate full-call negative ratio is negative (-0.402) but not statistically significant ($p = 0.14$), consistent with the Q&A signal being diluted when pooled with prepared remarks. Column (7) confirms that switching to two-way clustering preserves the sign and significance ($p < 0.10$). Columns (8) and (9) provide a robustness check for survivorship bias. Column (8) re-estimates a tone-only specification (firm and year FE, no fundamental controls) on the current 77-firm sample ($\hat{\beta} = -0.200^{**}$, $p = 0.022$, $N = 1675$); column (9) expands the sample by adding 14 historical Ibovespa constituents active during 2017–2024 but absent from the December 2025 index ($\hat{\beta} = -0.252^{***}$, $p = 0.002$, $N = 1857$). The Q&A negativity estimate strengthens in the broader sample, consistent with survivorship selection making the main-sample estimate conservative rather than upward-biased, although this 14-firm expansion is an approximation to, rather than a full reconstruction of, quarter-by-quarter Ibovespa membership, and the direction of the bias should be read accordingly as suggestive rather than definitively established. Overall, the robustness tests support the conclusion that negative Q&A tone maintains directional consistency across specifications, reaching conventional statistical significance in the baseline panel regressions and window variations. The association remains statistically significant under two-way clustering ($p < 0.10$), though it is estimated with greater noise when abnormal returns are measured using buy-and-hold returns.

A further robustness check addresses the concern that EPS surprise alone may be an insufficient control for the earnings announcement content. Earnings releases convey segment results, guidance revisions, and other quantitative information beyond EPS; if Q&A tone is mechanically correlated with this richer information set, the estimated tone–return association may partly reflect earnings news rather than independent call content. Adding the percentage earnings surprise, defined as (actual EPS minus consensus estimate) divided by the absolute value of the consensus, as a second earnings-news control alongside the winsorised EPS surprise yields a Q&A negative ratio coefficient of $\hat{\beta} = -0.252$ ($p = 0.064$; $N = 1459$). The coefficient's magnitude is unchanged from the baseline ($\hat{\beta} = -0.215$), but statistical significance falls from the 5% level ($p = 0.020$) to the 10% level ($p = 0.064$) on a sample approximately 11% smaller ($N = 1459$ versus $N = 1638$), since percentage earnings surprise is unavailable for 179 firm-quarters. This confirms that adding a richer measure of the magnitude of the earnings surprise does not attenuate the Q&A tone–return association, though the reduction in significance and sample size should be stated plainly rather than described as leaving the result “essentially unchanged”.

The segmented-call design offers a partial, though limited, internal check on earnings-announcement confounding, rather than a source of identification. The prepared introduction is a scripted summary that closely mirrors the earnings press release, covering the same numerical results and forward outlook in a structured format. In the specification that includes both segments jointly, the prepared-remarks negative coefficient ($\hat{\beta} = -0.215$, $p = 0.157$) is statistically indistinguishable from zero, while the Q&A negative coefficient ($\hat{\beta} = -0.215$, $p = 0.020$) is significant; however, a formal Wald test of equality between the two coefficients does not reject the null ($p = 0.997$; Section 4.2). The difference in statistical precision between the two segments therefore cannot, on its own, be interpreted as evidence that Q&A tone carries incremental information beyond what is captured by the structured introduction. The within-call contrast is suggestive of a segmented-information channel but does not provide identification: the design cannot distinguish whether Q&A tone conveys information beyond the earnings press release or simply estimates a similar underlying association more precisely.

Subperiod analyses address the concern that the sample period encompasses multiple macroeconomic and institutional regimes: Brazil's post-recession recovery, the Bolsonaro presidency, the Covid-19 shock, and the Lula administration, which may drive disclosure tone and market reactions through channels unrelated to firm-specific information. The preferred specification is estimated separately for three subperiods: 4Q 2016–4Q 2019 (pre-Covid; $N = 410$), 1Q 2020–4Q 2022 (Covid and turbulent macro period; $N = 548$), and 1Q 2023–4Q 2025 fiscal quarters (recent; $N = 673$), totalling 1631 observations (a net loss of 7 observations relative to the pooled sample of 1638 due to subperiod-specific fixed-effect singletons).

The Q&A negative tone coefficients from these three subperiod regressions are reported inline. The pre-Covid estimate is near zero ($\hat{\beta} = 0.009$, $p = 0.971$), the Covid/turbulent estimate is negative and larger in magnitude ($\hat{\beta} = -0.351$, $p = 0.238$), and the recent-period estimate is negative but small ($\hat{\beta} = -0.088$, $p = 0.518$). None of the subperiod estimates is individually significant, which reflects reduced statistical power: each subperiod contains fewer than 700 observations, representing less than half the full-sample size of 1638. Importantly, no subperiod estimate is statistically distinguishable from the full-sample coefficient of -0.215 .

Two interpretations are consistent with these patterns. First, the result may reflect a stable but modest relationship that is only detectable with the full nine-year sample. Second, the negative tone association may be amplified during periods of elevated macro uncertainty, when Q&A exchanges are more likely to surface firm-specific information that is not already reflected in prior prices. The second interpretation is economically plausible, information gaps between management and investors may widen during stress periods, but the subperiod samples are too small to discriminate between these explanations. The full-sample result should therefore be interpreted as an average over heterogeneous periods, and subperiod stability is acknowledged as an open question that larger future samples could address.

A further check confirms that the result is not an artefact of sample composition along the size or cross-listing dimension. Splitting at median firm size yields negative Q&A tone coefficients in both groups: large-cap firms ($\hat{\beta} = -0.224$, $p = 0.241$, $N = 817$) and small-cap firms ($\hat{\beta} = -0.197^{**}$, $p = 0.045$, $N = 818$). The small-cap estimate reaches 5% significance while the large-cap estimate is directionally consistent but imprecisely estimated, suggesting the Q&A signal is more precise and consistent where information asymmetry is greater. The ADR versus non-ADR split (Section 4.6) produces a directionally consistent pattern, although the two subsamples are statistically indistinguishable. Since all sample firms produce official English-language IR communications, approved by management and reflecting the company's intended investor-facing message, the relevant question is not translation quality but whether smaller Ibovespa firms produce equally information-rich Q&A exchanges as larger ones. The broadly consistent negative coefficients across size and listing splits are directionally consistent with this interpretation.

A key point of interpretation is that the empirical design does not isolate the conference call from the earnings release with complete precision. When the earnings release and the call occur close together, short-window returns may reflect both. The evidence is therefore best understood as showing that tone, especially negative tone in unscripted discussion, is associated with the broader disclosure event and with how investors react to it.

4.8. Discussion

The results support a segmented-information view of earnings calls. The prepared introduction and the Q&A discussion are not interchangeable components of the event. Across the main specifications, tone in the unscripted portion of the call is more reliably and precisely associated with short-window returns than tone in the prepared introduction, particularly when that tone is negative, even though a formal test does not detect a difference in the underlying magnitude of the two associations (Section 4.2). This pattern is consistent with the idea that analyst–management interaction reveals information that is less visible in the firm's scripted narrative.

The findings also point to an asymmetry in market interpretation. Negative language appears more informative than positive language, especially in the Q&A segment. This is consistent with the idea that market participants attach greater weight to signs of uncertainty, concern, or defensive communication than to broadly favourable statements.

An exploratory analysis sheds light on whether the tone–return association depends on the underlying earnings news. Splitting the sample into earnings-miss quarters (EPS surprise < 0; $N = 746$) and earnings-beat quarters (EPS surprise ≥ 0 ; $N = 888$) and re-estimating the preferred specification separately for each group reveals a directional difference: the Q&A negative tone coefficient is larger in magnitude for beats ($\beta = -0.491$, $p = 0.17$) than for misses ($\beta \approx +0.096$), though neither estimate is statistically significant at conventional levels. An interaction model is directionally consistent with this pattern ($\hat{\beta}_{\text{interaction}} = +0.586$, $p = 0.32$) but cannot reject the null of equal slopes. One interpretation is consistent with the main finding: when earnings beat expectations, negative Q&A tone represents incremental adverse information not already captured by the earnings surprise, and investors react accordingly; when earnings miss, Q&A negativity may be partly anticipated and therefore less surprising to the market. However, the results are imprecise, the subsamples have roughly half the statistical power of the full sample, and should be read as suggestive rather than definitive. Examining this heterogeneity more rigorously with larger samples is identified as a productive avenue for future research.

At the same time, the results should not be overstated. Most specifications explain only a small fraction of the variation in abnormal returns and trading volume. That is typical of short-window event studies, but it also requires disciplined interpretation. The appropriate conclusion is not that tone determines market outcomes, but that tone, especially unscripted negative tone, is one informative element of the earnings-call disclosure event.

4.9. Limitations

Several limitations of the analysis bear directly on how the findings should be interpreted.

The most important concerns identification. Earnings calls are held on or close to the same day as earnings releases, which convey segment results, guidance revisions, and other quantitative information beyond EPS that may correlate mechanically with call tone. The EPS surprise control addresses the headline numerical news, and an additional robustness check using the percentage earnings surprise confirms the main result is not driven by this specific limitation, though at reduced statistical significance and on a smaller sample (Section 4.7). Although the within-call segmentation isolates unscripted from scripted discourse, differential statistical significance across segments cannot be interpreted as a source of identification. Because a formal Wald test does not reject equality of the two coefficients (Section 4.2), the difference in estimation precision does not, on its own, establish that Q&A tone carries incremental information beyond the introduction. This design feature is accordingly best read as narrowing the object of study rather than as a source of identification: Q&A tone may still partially reflect segment-level surprises or guidance revisions that differentially affect the interactive portion of the call. Full identification would require text-based measures of the concurrent earnings press release, which are outside the scope of this study. The conclusions are deliberately framed as reduced-form associations throughout.

A second limitation concerns generalisability. The sample is restricted to Ibovespa constituents, firms that produce official English-language IR materials as standard practice, and the results therefore apply to the medium or large-cap, internationally visible segment of the market. The size-split robustness check (Section 4.7) shows near-identical Q&A negative tone coefficients for large-cap (-0.224 , $p = 0.241$) and small-cap (-0.197^{**} , $p = 0.045$) Ibovespa firms, suggesting the finding is not an artefact of the largest index members and is more precisely estimated where analyst coverage is lower; but whether it extends to the broader Brazilian public-company universe remains an open question, as smaller firms outside the index face lower stock liquidity and limited analyst coverage. Whether the segment-informativeness finding also replicates in other emerging markets, or in developed markets with different institutional structures, is a further open question. Finally, the sample period encompasses episodes of macroeconomic stress, including the Covid-19 shock and subsequent monetary tightening, that may have altered the informativeness of earnings-call communication in ways the empirical design does not explicitly model.

A third limitation concerns the endogeneity of call timing. The heterogeneity analysis in Section 4.5 conditions on whether calls occur during trading hours or before market open, but this classification is not randomly assigned. Firms may systematically select call timing based on characteristics such as investor-base composition, expected market reactions, or disclosure strategy. Although firm fixed effects absorb stable firm-level timing preferences and the regressions control for fundamental variables (EPS surprise, size, leverage, volatility), these controls do not fully resolve the endogeneity of timing choice. Although the determinants analysis reported in Section 4.5 finds that Q&A negative tone does not predict timing choice, this does not fully resolve the endogeneity concern; a rigorous treatment would require quasi-experimental variation in call timing or a structural model of the timing decision.

Fourth, regarding sentiment-classification validity, FinBERT is the only tone measure reaching statistical significance for Q&A negativity (Table 8), which we interpret as sample-specific empirical fit rather than intrinsic classification superiority (Section 2.4). Beyond our indirect checks, we formally validated FinBERT via double-blind manual coding of 200 randomly sampled sentences from the Brazilian corpus (100 prepared remarks, 100 Q&A; one per call) using a three-way codebook. Inter-annotator reliability was high (Cohen's $\kappa = 0.899$; 94.5% agreement), and FinBERT matched the annotators' consensus on 76.2% of the 189 agreed-upon sentences (recall: 88.0% neutral, 63.3% positive, 60.0% negative). This comparison provides no clear evidence of materially poorer FinBERT performance in the Brazilian corpus, though without a formal non-inferiority framework and given the small number of negative instances, it should not be taken as proof of identical classifier behaviour. Furthermore, the mild under-classification of negative relative to neutral sentiment is consistent across both corpora, representing a general genre-level property that would attenuate rather than inflate the estimated association between Q&A negativity and abnormal returns. The negative category is the exception: each sample contains only 10 and 4 annotator-coded negative sentences respectively, so the negative-recall estimates above are imprecise and should not be read as establishing equivalent negative-sentiment performance across corpora.

5. Conclusion

This paper examines whether linguistic tone in Brazilian earnings conference calls is associated with short-window market reactions and whether the Q&A segment appears more informative than prepared managerial remarks. Using earnings-call transcripts from listed Brazilian firms over 4Q 2016–4Q 2025, we measure tone separately for the prepared introduction and Q&A segments, and relate these measures to *CAR* and *CAV* around call dates.

The main finding is that negative tone in the Q&A segment shows a consistently negative association with short-window abnormal returns across methods and specifications, with the FinBERT measure reaching conventional statistical significance and Loughran–McDonald and VADER providing directionally consistent but less precisely estimated evidence. While formal tests show that the economic magnitude of negative tone in the prepared remarks is statistically indistinguishable from that in the Q&A, its association is less precisely estimated and statistically unreliable across estimators, while the evidence for abnormal trading volume is more fragile across specifications. These results are consistent with a segmented-information view of earnings calls: the scripted and unscripted portions of the call represent distinct communication environments. Although the data fail to support the prediction that Q&A tone has a larger economic magnitude (*Hypothesis 2*), the unscripted segment provides an empirical pattern of substantially greater estimation precision and cross-specification stability.

Additional, exploratory evidence from call timing is directionally consistent with this interpretation, though it should be interpreted cautiously given that the interaction test fails to detect a statistically distinct regime ($p = 0.82$). Negative Q&A tone is more precisely and consistently associated with abnormal returns when calls occur during trading hours, when investors can react while the discussion unfolds. For calls held before the market opens, estimation precision shifts towards the prepared introduction. This pattern is consistent with the exploratory conjecture that the Q&A segment is processed as a live channel during trading hours, whereas prepared remarks carry relatively more weight when the call occurs outside trading hours.

The paper contributes to the literature in four ways. First, it provides the first segment-level, multi-method analysis of earnings-call tone and short-window market reactions in Brazil, extending (Tonin & Scherer, 2021) with a larger sample and explicit call segmentation. Second, it shows that the internal structure of the call matters when evaluating qualitative corporate disclosure. Although formal equality tests indicate that the economic magnitude of the association is statistically indistinguishable between segments, negative tone in the unscripted Q&A segment is estimated with substantially greater precision and cross-model stability than tone in the prepared remarks. Third, the call-timing analysis provides suggestive evidence consistent with the idea that segment informativeness may vary with whether the call occurs during trading hours, although the continuous interaction test fails to establish a statistically distinct regime, motivating further work on disclosure timing and market responsiveness. Fourth, it compares alternative sentiment measures as competing ways of operationalising tone; FinBERT produces more precise and stable associations in this sample, which is interpreted as better empirical fit rather than demonstrated classification superiority.

As with any event-study design, some caveats apply. Separating prepared remarks from the Q&A portion helps narrow the locus of the association, though the concurrent timing of calls and earnings releases means the design cannot fully isolate conference-call tone from other announcement content. Sentiment measures account for only a modest share of short-window return variation, consistent with the many channels through which earnings information reaches investors. The findings are best understood as reduced-form associations between unscripted negative tone and short-window investor reactions.

CRedit authorship contribution statement

Rodrigo dos Reis Maia: Writing – original draft, Methodology, Investigation, Data curation, Conceptualization. **Jorge Miguel Bravo:** Writing – review & editing, Validation, Funding acquisition, Formal analysis.

Funding statement

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)).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Variable definitions

See [Table A.1](#).

Table A.1

Variable definitions.

Variable	Label	Construction	Source
Panel A: Outcome variables			
car_value	CAR [0, +1]	Sum of market-model abnormal returns over the two-day window [0, +1], centred on the earnings calls date. Estimation window: [-62, -2] trading days.	Event study (Fama–French three-factor model, IBOVESPA index)
bhar_value	BHAR [0, +1]	Buy-and-hold abnormal return: product of gross returns minus product of benchmark gross returns over the two-day window [0, +1].	Event study (Fama–French three-factor model, IBOVESPA index)
cav_value	CAV [0, +1]	Cumulative abnormal volume: sum of abnormal log-volume over two-day window [0, +1], where abnormal volume is actual minus expected from a pre-event mean model.	Constant Mean model, IBOVESPA index
Panel B: Sentiment Variables (FinBERT, LM, VADER, Baier)			
All tone measures follow the framework defined in Table 2. Raw ratios give the segment-level share of negative or positive words/sentences. The negativity index equals $(N - P)/(N + P + 1)$, bounded on [-1,1], where N and P are raw counts (sentence proportions for VADER). Subscript _intro = prepared-remarks segment; _qna = Q&A segment. See Appendix C for preprocessing details.			
finbert_{intro qna}_neg_to_total	FinBERT Neg. Ratio	Share of sentences classified as negative by FinBERT in the given segment.	Transcripts; FinBERT (Huang et al., 2023)
finbert_{intro qna}_pos_to_total	FinBERT Pos. Ratio	Share of sentences classified as positive by FinBERT in the given segment.	Transcripts; FinBERT
finbert_{intro qna}_negativity	FinBERT Negativity	Negativity index (Druz et al., 2020); FinBERT.	FinBERT
lm_{intro qna}_neg_to_total	LM Neg. Ratio	Share of words matched to the LM negative dictionary.	Transcripts; LM dictionary (Loughran & McDonald, 2011)
lm_{intro qna}_pos_to_total	LM Pos. Ratio	Share of words matched to the LM positive dictionary.	Transcripts; LM dictionary
lm_{intro qna}_negativity	LM Negativity	Negativity index applied to raw LM word counts.	Transcripts; LM dictionary
vader_{intro qna}_neg_to_total	VADER Neg. Ratio	Share of sentences classified as negative by VADER.	Transcripts; VADER (Hutto & Gilbert, 2014)
vader_{intro qna}_pos_to_total	VADER Pos. Ratio	Share of sentences classified as positive by VADER.	Transcripts; VADER
vader_{intro qna}_negativity	VADER Negativity	Negativity index applied to VADER	Transcripts; VADER
Panel C: Control variables			
eps_surprise	EPS Surprise	Actual EPS minus consensus analyst EPS estimate (BRL). Consensus defined as the mean of all estimates available in the 30 days prior to the announcement.	Koyfin
total_assets	Total Assets	Book value of total assets at the most recent fiscal quarter end (BRL millions).	Koyfin
pb	Price-to-Book	Market capitalisation divided by book value of equity at quarter end.	Koyfin
leverage	Leverage	Net debt (total debt minus cash and equivalents) divided by total assets.	Koyfin
volatility	Volatility	Annualised historical volatility computed from daily returns over the 63 trading days preceding the earnings calls (%).	Koyfin
marketcap	Market Capitalisation	Market value of equity at quarter end (BRL millions).	Koyfin
Panel D: Transformed variables			
log_total_assets	ln(Total Assets)	Natural logarithm of total assets (BRL millions). Applied to reduce right skew.	Derived from total_assets
log_volatility	ln(Volatility)	Natural logarithm of annualised volatility (%). Applied to reduce right skew.	Derived from volatility
log_marketcap	ln(Market Cap)	Natural logarithm of market capitalisation (BRL millions).	Derived from marketcap

(continued on next page)

Table A.1 (continued).

log1p_gross_debt	$\ln(1 + \text{Gross Debt})$	Log-plus-one transform of gross debt to accommodate zero values.	Derived from gross_debt
eps_surprise_w	EPS Surprise (winsorised)	EPS surprise winsorised at the 1st and 99th percentiles to limit influence of extreme analyst forecast errors.	Derived from eps_surprise
pb_w	Price-to-Book (winsorised)	Price-to-book ratio winsorised at the 1st and 99th percentiles.	Derived from pb
net_debt_w	Net Debt (winsorised)	Net debt winsorised at the 1st and 99th percentiles; retains sign (negative values indicate net cash positions).	Derived from gross_debt, marketcap

Appendix B. Event-study implementation

For each firm-quarter observation, a Fama–French three-factor model is estimated over the 60-trading-day pre-event estimation window $[-62, -2]$, where day 0 is the earnings call date. The model takes the form:

$$AR_{i,t} = R_{i,t} - (\hat{\alpha}_i + \hat{\beta}_{1i} RM_t + \hat{\beta}_{2i} SMB_t + \hat{\beta}_{3i} HML_t), \quad (\text{B.1})$$

where $R_{i,t}$ denotes the daily log-return of firm i on day t ; RM_t is the Brazilian market excess return; SMB_t and HML_t are the Brazilian size and value factors, respectively, sourced from NEFIN/FINEP (NEFIN, 2024); and $\hat{\alpha}_i$, $\hat{\beta}_{1i}$, $\hat{\beta}_{2i}$, $\hat{\beta}_{3i}$ are OLS estimates from the estimation window. The CAR for each observation is the sum of daily abnormal returns, $AR_{i,t}$, over the $[0, +1]$ event window.

Abnormal trading volume is constructed in log-turnover space to remove level heterogeneity across firms. For each trading day, firm i 's log-turnover is the log of shares traded divided by shares outstanding, $\ln(V_{i,t}/S_{i,t})$. Expected log-turnover is estimated over the same $[-62, -2]$ window using a market model that regresses firm log-turnover on the log of the Ibovespa's mean daily turnover, $\ln(\bar{V}_t^M)$:

$$AV_{i,t} = \ln\left(\frac{V_{i,t}}{S_{i,t}}\right) - (\hat{\alpha}_i + \hat{\beta}_i \ln(\bar{V}_t^M)). \quad (\text{B.2})$$

The estimation accounts for first-order autocorrelation in firm-level volume series. The CAV for each observation is the sum of daily abnormal volume over the same $[0, +1]$ window.

Appendix C. Text preprocessing

Transcripts are preprocessed using Python with NLTK. The preprocessing steps include uniform encoding, removal of formatting artefacts, tokenisation, and sentence-boundary detection. Segmentation into prepared remarks and Q&A is performed algorithmically using transition expressions, such as “let's begin the question-and-answer session”, and verified manually for all transcripts. Preprocessed files are stored in a flat directory structure and matched to event identifiers using firm ticker and call date.

Appendix D. Robustness checks

Table A.2 reports eight robustness checks on the baseline CAR regression (Table 5, column (4)). Column (1) replaces CAR with buy-and-hold abnormal returns (BHAR). Column (2) uses aggregate full-call FinBERT sentiment instead of the prepared-remarks/Q&A split. Column (3) excludes the six observations with $|\text{CAR}| > 3\sigma$. Column (4) adds the log of total transcript sentences as a length control. Column (5) adds the number of sell-side analysts participating in the call. Column (6) re-estimates the baseline using a 120-trading-day estimation window $[-122, -2]$.

Column (7) repeats the baseline with two-way clustered SE (ticker $\times$ call date); all directional conclusions are preserved. Regarding two-way clustering, all main specifications cluster standard errors by firm (ticker), following Petersen (2009), who shows that firm clustering is the appropriate default when residuals contain a persistent firm effect. Two-way clustering by firm and call date (Cameron et al., 2011) is the natural conservative extension when residuals may also be correlated across firms sharing the same reporting date. However, Cameron and Miller (2015) caution that cluster-robust inference relies on having a sufficient number of observations *within* each cluster to estimate within-cluster dependence reliably. In the present sample, 39% of call dates have a single firm and 75% have three or fewer firms (median: two), so the date-level clusters are too sparse to estimate within-date dependence with precision. Moreover, year fixed effects already absorb the common time-varying shocks that date clustering is designed to address. Column (7) confirms that all directional and significance-ranking conclusions are preserved under two-way clustering, supporting the use of the more efficient firm-only specification in the main tables.

The survivorship robustness check (columns (8)–(9) of Table A.2) adds 14 historical Ibovespa constituents active during 2017–2024 that are not among the December 2025 index constituents. This robustness check mitigates, rather than eliminates, survivorship bias, as full quarterly Ibovespa reconstruction remains future work. Auditing B3 rebalancing records (4Q16–4Q25) reveals 21 unique index departures, excluding corporate share-class consolidations. Our sample includes 14 of these firms (66.7% of genuine exits), restricted solely by the availability of pre-existing FinBERT sentiment scores from an identical processing pipeline. Rather than a selective subset, these 14 firms represent a substantial, data-constrained sample of historical constituents. FinBERT

Table A.2

Robustness checks: CAR regressions (FinBERT).

	(1) BHAR	(2) Full	(3) No outliers	(4) Length	(5) Analysts	(6) 120-day	(7) Two-way SE	(8) 77 firms	(9) 91 firms
Intro pos.	−0.097 (0.059)		0.042** (0.017)	−0.010 (0.051)	−0.005 (0.047)	0.028 (0.020)	0.005 (0.045)	0.005 (0.038)	0.015 (0.038)
Intro neg.	−0.292 (0.261)		−0.017 (0.035)	−0.248 (0.179)	−0.241 (0.182)	−0.065* (0.033)	−0.215 (0.162)	−0.192 (0.151)	−0.136 (0.128)
Q&A pos.	−0.072 (0.080)		0.029 (0.025)	−0.056 (0.063)	−0.045 (0.058)	0.063*** (0.016)	−0.042 (0.058)	−0.042 (0.064)	−0.034 (0.058)
Q&A neg.	−0.438 (0.354)		−0.149** (0.057)	−0.209* (0.113)	−0.221* (0.113)	−0.129*** (0.046)	−0.215* (0.116)	−0.200** (0.085)	−0.252*** (0.078)
Full pos.		0.006 (0.077)							
Full neg.		−0.402 (0.266)							
Length				−0.025* (0.014)					
Analysts					−0.005 (0.004)				
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1638	1638	1632	1638	1638	1775	1638	1675	1857
R ²	0.064	0.027	0.090	0.030	0.032	0.100	0.028	0.025	0.031

Notes. Baseline specification is Table 5, col. (4) (firm and year FE, full controls, SE clustered by ticker, $N = 1638$). Any column with $N \neq 1638$ is noted below. (1): BHAR [0, +1] as alternative return measure. (2): aggregate full-call FinBERT sentiment replacing the intro/Q&A split. (3): excludes $|CAR| > 3\sigma$ ($N = 1632$; 6 obs. removed). (4): adds log total transcript sentences as a length control. (5): adds number of sell-side analysts on the call. (6): 120-trading-day estimation window [−122, −2] using the Fama–French three-factor model with NEFIN risk factors and the Ibovespa as market proxy; CARs computed from daily prices downloaded via Yahoo Finance ($N = 1775$, larger than the 1638-observation baseline because the longer estimation window is more lenient on the minimum-non-missing-trading-days requirement, so more firm-quarters yield a valid abnormal return before the standard sample filters are applied). (7): same specification as the baseline but with two-way clustered SE by ticker and call date; date clusters are sparse (median 2 firms per date, 39% singletons) so the direction and significance ranking ($p < 0.10$ in this study) of all results are preserved. (8)–(9): survivorship robustness; tone-only specification (firm and year FE, no fundamental controls, SE clustered by firm) since balance-sheet data are unavailable for historical constituents. Col. (8): current 77-firm sample ($N = 1675$, using the tone-only base of 1677 observations rather than 1640, since fundamental-control completeness is not required in this specification). Col. (9): adds 14 historical Ibovespa constituents absent from the December 2025 index (91 firms total, $N = 1857$); this is an expanded historical-constituent robustness check that mitigates, rather than eliminates, survivorship bias (Section 3.1), not a full reconstruction of quarterly Ibovespa membership. Cols (1)–(7) use SE clustered by ticker. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

sentiment scores for both call segments were available from a prior study using the same pipeline. A tone-only specification is used (firm and year FE, SE clustered by firm, no fundamental controls) to compare both sets. Column (8) applies this specification to the current 77-firm sample ($\hat{\beta} = -0.200^{**}$, $p = 0.022$, $N = 1675$); column (9) expands to 91 firms ($\hat{\beta} = -0.252^{***}$, $p = 0.002$, $N = 1857$). The estimate strengthens in the broader sample, consistent with survivorship selection making the main-sample result conservative rather than upward-biased, though this directional inference rests on only 14 additional firms rather than a full dynamic reconstruction and should be read as suggestive.

Table A.2 also reports a robustness check using CAR computed from a 120-trading-day estimation window [−122, −2] with the same Fama–French three-factor model and NEFIN risk factors used in the baseline (column (6)). Daily returns were downloaded from Yahoo Finance and the event study was re-estimated from scratch; the market excess return was constructed as the Ibovespa daily return minus the NEFIN risk-free rate. The Q&A negative ratio remains negative and statistically significant at the 1% level ($\hat{\beta} = -0.129$, $p < 0.01$) under this alternative window, so the main finding is not sensitive to the choice of estimation window. The Q&A positive coefficient is also statistically significant here ($\hat{\beta} = 0.063$, $p < 0.01$), but Section 4.2 already treats positive-tone coefficients as unstable across specifications, and this single significant estimate in one robustness column does not overturn that assessment; it is reported for completeness rather than as confirmatory evidence for a symmetric segmented-call effect.

Appendix E. Alternative sentiment measures

E.1. Loughran–McDonald

Tables A.3 and A.4 replicate the main CAR and CAV specifications using Loughran–McDonald (2011) word counts. Each table presents six specifications, all including firm and year fixed effects, with standard errors clustered by ticker. Columns (I)–(III) use the fractions of positive and negative words separately for the prepared remarks and the Q&A session. Column (I) includes only the prepared-remarks ratios, column (II) includes only the Q&A ratios, and column (III) includes both. Columns (IV)–(VI) replace the separate word-ratio measures with the negativity index, defined as $(neg - pos)/(neg + pos + 1)$, for each transcript segment: column (IV) includes prepared-remarks negativity only, column (V) includes Q&A negativity only, and column (VI) includes both simultaneously.

Table A.3
Tone and CAR — Loughran–McDonald.

	(I)	(II)	(III)	(IV)	(V)	(VI)
Intro positive	−0.594 (0.625)		−0.422 (0.523)			
Intro negative	−2.083 (1.430)		−1.959 (1.358)			
Q&A positive		−0.764 (0.644)	−0.678 (0.557)			
Q&A negative		−0.755 (0.555)	−0.621 (0.462)			
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Intro negativity				−0.026 (0.018)		−0.027 (0.021)
Q&A negativity					−0.001 (0.017)	0.006 (0.021)
EPS surprise	0.038 (0.026)	0.040 (0.025)	0.042 (0.026)	0.033 (0.026)	0.035 (0.026)	0.033 (0.026)
Log total assets	0.006 (0.014)	0.005 (0.014)	0.005 (0.014)	0.005 (0.014)	0.005 (0.014)	0.006 (0.014)
Price-to-book	−0.002 (0.002)	−0.002 (0.002)	−0.003 (0.002)	−0.002 (0.002)	−0.002 (0.002)	−0.002 (0.002)
Leverage	0.048 (0.034)	0.044 (0.033)	0.045 (0.034)	0.049 (0.036)	0.049 (0.035)	0.050 (0.037)
Log volatility	0.009 (0.007)	0.008 (0.006)	0.011 (0.007)	0.008 (0.006)	0.006 (0.006)	0.008 (0.006)
Observations	1638	1638	1638	1638	1638	1638
Within R ²	0.009	0.005	0.012	0.003	0.001	0.003

Notes. Dep. var.: CAR [0, +1] (decimal). Firm + year FE. Clustered SE by ticker. (I)–(III): pos./neg. ratios. (IV)–(VI): negativity = $(neg - pos)/(neg + pos + 1)$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.4
Tone and CAV — Loughran–McDonald.

	(I)	(II)	(III)	(IV)	(V)	(VI)
Intro positive	8.001** (3.738)		9.311** (3.655)			
Intro negative	2.008 (3.842)		1.994 (4.017)			
Q&A positive		−2.453 (3.052)	−4.671 (2.965)			
Q&A negative		−0.073 (2.350)	−0.299 (2.383)			
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Intro negativity				−0.103 (0.092)		−0.115 (0.102)
Q&A negativity					0.037 (0.142)	0.067 (0.152)
EPS surprise	−0.256 (0.219)	−0.224 (0.216)	−0.243 (0.217)	−0.240 (0.220)	−0.231 (0.218)	−0.238 (0.220)
Log total assets	0.184 (0.134)	0.183 (0.134)	0.187 (0.136)	0.181 (0.133)	0.183 (0.134)	0.184 (0.134)
Price-to-book	0.004 (0.020)	0.003 (0.020)	0.003 (0.020)	0.002 (0.020)	0.003 (0.020)	0.002 (0.020)
Leverage	0.055 (0.405)	0.038 (0.397)	0.052 (0.396)	0.042 (0.399)	0.043 (0.399)	0.047 (0.393)
Log volatility	−0.297*** (0.089)	−0.303*** (0.086)	−0.298*** (0.087)	−0.292*** (0.088)	−0.304*** (0.085)	−0.295*** (0.086)
Observations	1638	1638	1638	1638	1638	1638
Within R ²	0.015	0.011	0.016	0.011	0.010	0.011

Notes. Dep. var.: CAV [0, +1] (cumul. abnormal log-volume). Firm + year FE. Clustered SE by ticker. (I)–(III): pos./neg. ratios. (IV)–(VI): negativity = $(neg - pos)/(neg + pos + 1)$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

E.2. VADER

Tables A.5 and A.6 repeat the same exercise using VADER, the Valence Aware Dictionary and sEntiment Reasoner, with compound scores aggregated at the sentence level. The column structure mirrors that of Tables A.3 and A.4: columns (I)–(III) use the

Table A.5
Tone and CAR — VADER.

	(I)	(II)	(III)	(IV)	(V)	(VI)
Intro positive	0.072 (0.177)		0.110 (0.217)			
Intro negative	−0.940 (0.574)		−0.854 (0.537)			
Q&A positive		−0.136 (0.193)	−0.149 (0.217)			
Q&A negative		−0.576 (0.404)	−0.392 (0.318)			
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Intro negativity				−0.418 (0.361)		−0.427 (0.398)
Q&A negativity					−0.032 (0.110)	0.036 (0.168)
EPS surprise	0.029 (0.028)	0.037 (0.027)	0.031 (0.028)	0.030 (0.029)	0.035 (0.026)	0.030 (0.029)
Log total assets	0.001 (0.013)	0.004 (0.013)	0.000 (0.012)	0.004 (0.013)	0.006 (0.014)	0.004 (0.013)
Price-to-book	−0.003 (0.002)	−0.002 (0.002)	−0.003 (0.002)	−0.002 (0.002)	−0.002 (0.002)	−0.002 (0.002)
Leverage	0.052 (0.039)	0.048 (0.037)	0.052 (0.041)	0.050 (0.037)	0.048 (0.035)	0.050 (0.038)
Log volatility	0.012 (0.007)	0.009 (0.007)	0.013 (0.008)	0.009 (0.007)	0.006 (0.006)	0.008 (0.006)
Observations	1638	1638	1638	1638	1638	1638
Within R ²	0.009	0.004	0.010	0.006	0.001	0.006

Notes. Dep. var.: CAR [0, +1] (decimal). Firm + year FE. Clustered SE by ticker. (I)–(III): pos./neg. ratios. (IV)–(VI): negativity = $(neg - pos)/(neg + pos + 1)$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.6
Tone and CAV — VADER.

	(I)	(II)	(III)	(IV)	(V)	(VI)
Intro positive	1.173 (1.062)		1.220 (1.062)			
Intro negative	−2.273 (2.085)		−2.778 (2.049)			
Q&A positive		−0.140 (0.623)	−0.310 (0.629)			
Q&A negative		1.222 (2.212)	1.760 (2.122)			
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes	Yes
Intro negativity				−1.923* (1.021)		−2.159* (1.092)
Q&A negativity					0.605 (0.752)	0.951 (0.824)
EPS surprise	−0.258 (0.227)	−0.233 (0.216)	−0.260 (0.224)	−0.257 (0.227)	−0.230 (0.217)	−0.256 (0.227)
Log total assets	0.172 (0.133)	0.181 (0.132)	0.169 (0.132)	0.174 (0.134)	0.179 (0.132)	0.169 (0.134)
Price-to-book	−0.000 (0.020)	0.004 (0.020)	0.000 (0.020)	0.000 (0.020)	0.004 (0.020)	0.001 (0.020)
Leverage	0.047 (0.398)	0.048 (0.408)	0.062 (0.401)	0.044 (0.402)	0.047 (0.401)	0.054 (0.398)
Log volatility	−0.286*** (0.088)	−0.309*** (0.088)	−0.294*** (0.089)	−0.289*** (0.088)	−0.306*** (0.087)	−0.294*** (0.087)
Observations	1638	1638	1638	1638	1638	1638
Within R ²	0.013	0.011	0.014	0.013	0.011	0.014

Notes. Dep. var.: CAV [0, +1] (cumul. abnormal log-volume). Firm + year FE. Clustered SE by ticker. (I)–(III): pos./neg. ratios. (IV)–(VI): negativity = $(neg - pos)/(neg + pos + 1)$. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

positive and negative sentence-ratio measures, whereas columns (IV)–(VI) use the negativity index, $(neg - pos)/(neg + pos + 1)$. As in the Loughran–McDonald tables, column (IV) includes only prepared-remarks negativity, column (V) includes only Q&A negativity, and column (VI) includes both.

Data availability

Data will be made available on request.

References

- Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. <http://dx.doi.org/10.48550/arXiv.1908.10063>, arXiv preprint. unreviewed preprint; see Huang et al. (2023) for peer-reviewed corroboration.
- Aromi, J. D., & Clements, A. (2021). Facial expressions and the business cycle. *Economic Modelling*, 102, Article 105563. <http://dx.doi.org/10.1016/j.econmod.2021.105563>.
- Aroussi, R. (2023). Yfinance: Yahoo! finance market data downloader. URL: <https://github.com/ranaroussi/yfinance>. python package, version 0.2.
- B3 (2026). *Methodology of IBOVESPA index calculation: Technical report*, B3 – Brasil, Bolsa, Balcão: URL: https://www.b3.com.br/pt_br/market-data-e-indices/indices/indices-amplos/ibovespa.htm. (Accessed 20 June 2026).
- B3 Investimentos (2026). Brasil é o ‘novo ouro’? investidor estrangeiro chega a patamar recorde de participação na B3. URL: <https://borainvestir.b3.com.br/noticias/mercado/brasil-e-o-novo-ouro-investidor-estrangeiro-chega-a-patamar-recorde-de-participacao-na-b3/>. (Accessed 20 June 2026).
- Bergé, L. (2018). *Efficient estimation of maximum likelihood models with multiple fixed-effects: the r package fENmlm*. CREA Discussion Paper 13, CREA, University of Luxembourg, URL: <https://github.com/lrberge/fixest>.
- Beyer, A., Cohen, D. A., Lys, T. Z., & Walther, B. R. (2010). The financial reporting environment: Review of the recent literature. *Journal of Accounting and Economics*, 50, 296–343. <http://dx.doi.org/10.1016/j.jacceco.2010.10.003>.
- Bowen, R. M., Davis, A. K., & Matsumoto, D. A. (2002). Do conference calls affect analysts forecasts? *The Accounting Review*, 77, 285–316. <http://dx.doi.org/10.2308/accr.2002.77.2.285>.
- Brockman, P., Li, X., & Price, S. M. (2015). Differences in conference call tones: Managers vs. analysts. *Financial Analysts Journal*, 71, 24–42. <http://dx.doi.org/10.2469/faj.v71.n4.1>.
- Bushee, B. J., & Huang, Y. (2024). Do analysts and investors efficiently respond to managerial linguistic complexity during conference calls? *The Accounting Review*, 99, 143–168. <http://dx.doi.org/10.2308/TAR-2019-0358>.
- Bushee, B. J., Matsumoto, D. A., & Miller, G. S. (2003). Open versus closed conference calls: The determinants and effects of broadening access to disclosure. *Journal of Accounting and Economics*, 34, 149–180. [http://dx.doi.org/10.1016/S0165-4101\(02\)00073-3](http://dx.doi.org/10.1016/S0165-4101(02)00073-3).
- Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics*, 29, 238–249. <http://dx.doi.org/10.1198/jbes.2010.07136>.
- Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50, 317–372. <http://dx.doi.org/10.3368/jhr.50.2.317>.
- Cao, S., Jiang, W., Yang, B., & Zhang, A. L. (2023). How to talk when a machine is listening: Corporate disclosure in the age of AI. *Review of Financial Studies*, 36, 3603–3642. <http://dx.doi.org/10.1093/rfs/hhad021>.
- Chen, Y., Han, D., & Zhou, X. (2023). Mining the emotional information in the audio of earnings conference calls: A deep learning approach for sentiment analysis of securities analysts' follow-up behavior. *International Review of Financial Analysis*, 88, Article 102704. <http://dx.doi.org/10.1016/j.irfa.2023.102704>.
- Davis, A. K., Ge, W., Matsumoto, D., & Zhang, J. L. (2015). The effect of manager-specific optimism on the tone of earnings conference calls. *Review of Accounting Studies*, 20, 639–673. <http://dx.doi.org/10.1007/s11142-014-9309-4>.
- Davis, A. K., & Tama-Sweet, I. (2012). Managers' use of language across alternative disclosure outlets: Earnings press releases versus md&a. *Contemporary Accounting Research*, 29, 804–837. <http://dx.doi.org/10.1111/j.1911-3846.2011.01125.x>.
- Druz, M., Petzev, I., Wagner, A. F., & Zeckhauser, R. J. (2020). When managers change their tone, analysts and investors change their tune. *Financial Analysts Journal*, 76, 47–69. <http://dx.doi.org/10.1080/0015198X.2019.1707592>.
- Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, 33, 3–56. [http://dx.doi.org/10.1016/0304-405X\(93\)90023-5](http://dx.doi.org/10.1016/0304-405X(93)90023-5).
- Fiset, J., Oldford, E., & Chu, S. (2021). Market signaling capacity of written and visual charismatic leadership tactics. *Journal of Behavioral and Experimental Finance*, 29, Article 100465. <http://dx.doi.org/10.1016/j.jbef.2021.100465>.
- Frankel, R., Johnson, M., & Skinner, D. J. (1999). An empirical examination of conference calls as a voluntary disclosure medium. *Journal of Accounting Research*, 37, 133–150. <http://dx.doi.org/10.2307/2491400>.
- García, D., Hu, X., & Rohrer, M. (2023). The colour of finance words. *Journal of Financial Economics*, 147, 525–549. <http://dx.doi.org/10.1016/j.jfineco.2022.11.006>.
- Gow, I. D., Larcker, D. F., & Zakolyukina, A. A. (2021). Non-answers during conference calls. *Journal of Accounting Research*, 59, 1349–1384. <http://dx.doi.org/10.1111/1475-679X.12371>.
- Guimarães, P., & Portugal, P. (2010). A simple feasible procedure to fit models with high-dimensional fixed effects. *Stata Journal*, 10, 628–649. <http://dx.doi.org/10.1177/1536867X1101000406>.
- Healy, P. M., & Palepu, K. G. (2001). Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature. *Journal of Accounting and Economics*, 31, 405–440. [http://dx.doi.org/10.1016/S0165-4101\(01\)00018-0](http://dx.doi.org/10.1016/S0165-4101(01)00018-0).
- Hennig, J. C., Firk, S., & Wolff, M. (2023). Credibility signals from soft information: Evidence from investor reactions to tone in earnings conference calls. *European Accounting Review*, 33, 1319–1348. <http://dx.doi.org/10.1080/09638180.2023.2244009>.
- Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40, 806–841. <http://dx.doi.org/10.1111/1911-3846.12832>.
- Hutto, C. J., & Gilbert, E. E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the 8th international AAAI conference on weblogs and social media ICWSM*, (pp. 216–225). <http://dx.doi.org/10.1609/icwsml.v8i1.14550>.
- Ito, T., Sakaji, H., Tsubouchi, K., Izumi, K., & Yamashita, T. (2018). Text-visualizing neural network model: Understanding online financial textual data. In *Advances in knowledge discovery and data mining: 22nd Pacific-Asia conference (PAKDD 2018), proceedings, part III* (pp. 247–259). Springer, http://dx.doi.org/10.1007/978-3-319-93040-4_20.
- Keith, K. A., & Stent, A. (2019). Modeling financial analysts' decision making via the pragmatics and semantics of earnings calls. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 493–503). Florence, Italy: Association for Computational Linguistics, <http://dx.doi.org/10.18653/v1/P19-1047>.
- Kimbrough, M. D. (2005). The effect of conference calls on analyst and market underreaction to earnings announcements. *The Accounting Review*, 80, 189–219. <http://dx.doi.org/10.2308/accr.2005.80.1.189>.
- Koyfin (2024). Koyfin financial data platform. URL: <https://app.koyfin.com>. (Accessed 30 April 2026).
- Loughran, T., & McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *Journal of Finance*, 66, 35–65. <http://dx.doi.org/10.1111/j.1540-6261.2010.01625.x>.

- Manabe, T., Kuroki, Y., Sashida, S., & Nakagawa, K. (2022). *Analysis between sentiment of financial results briefing text data and stock return: JSAI Technical Report, Special Interest Group on Financial Informatics*, (pp. 47–53). http://dx.doi.org/10.11517/jsaisigtwo.2022.fin-029_47, in Japanese. Published in the JSAI SIG on Financial Informatics.
- Matsumoto, D., Pronk, M., & Roelofsen, E. (2011). What makes conference calls useful? the information content of managers' presentations and analysts' discussion sessions. *The Accounting Review*, 86, 1383–1414. <http://dx.doi.org/10.2308/accr-10034>.
- Mayew, W. J., & Venkatachalam, M. (2012). The power of voice: Managerial affective states and future firm performance. *Journal of Finance*, 67, 1–43. <http://dx.doi.org/10.1111/j.1540-6261.2011.01705.x>.
- NEFIN (2024). *Brazilian fama-french risk factors*. Núcleo de Pesquisa em Economia Financeira, Universidade de São Paulo, URL: <https://nefin.com.br/data/risk-factors>. (Accessed 30 April 2026).
- Park, J. E., Wang, Y., Wei, S., & Zhang, J. I. (2025). Financial statement comparability and managers' linguistic choices in conference calls. *European Accounting Review*, 34, 2125–2152. <http://dx.doi.org/10.1080/09638180.2024.2424833>.
- Petersen, M. A. (2009). Estimating standard errors in finance panel data sets: Comparing approaches. *The Review of Financial Studies*, 22, 435–480. <http://dx.doi.org/10.1093/rfs/hhn053>.
- 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 & Finance*, 36, 992–1011. <http://dx.doi.org/10.1016/j.jbankfin.2011.10.013>.
- dos Reis Maia, R., & Bravo, J. M. (2026). A new scale for evaluating disclosure in earnings calls on emerging markets. *Emerging Science Journal*, 10, 247–275. <http://dx.doi.org/10.28991/ESJ-2026-010-01-012>.
- Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance*, 62, 1139–1168. <http://dx.doi.org/10.1111/j.1540-6261.2007.01232.x>.
- Tonin, J. M. d. F., & Scherer, L. M. (2021). Market reaction to the tones of earnings conference calls. *RAE – Revista de Administração de Empresas*, 62, 1–18. <http://dx.doi.org/10.1590/S0034-759020220107>.
- Yamamoto, R., Kawadai, N., Kurita, M., & Takahashi, H. (2022). Managements' tone strategies by earnings call transcripts in the global markets. *Journal of Asset Management*, 23, 246–255. <http://dx.doi.org/10.1057/s41260-022-00256-2>.
- Yang, B. C., Liu, Y. F., & Su, Y. P. (2023). Earnings communication conferences and post-earnings-announcement drift. *Accounting & Finance*, 63, 4679–4709. <http://dx.doi.org/10.1111/acfi.12956>.

Rodrigo Maia is a Ph.D. Candidate in Data Science with a specialisation in Finance at NOVA Information Management School (NOVA IMS), NOVA University Lisbon, Portugal. His research focuses on the application of data science, machine learning, and quantitative methods to finance, investor relations, and capital markets. He has published research in leading international peer-reviewed journals in finance and related fields.

Prior to his doctoral studies, he accumulated more than 25 years of professional experience in investor relations, corporate finance, and capital markets. He participated in debt and equity issuances exceeding USD 10 billion and was recognised by Institutional Investor magazine as the Best Investor Relations Professional in the Latin American Metals & Mining sector in 2020 and 2021. His professional experience includes leadership positions at Gerdau, Braskem, and CSN. His research interests include sustainable finance, corporate disclosure, investor relations, and financial analytics.

Jorge Miguel Bravo is Associate Professor of Economics and Finance at NOVA Information Management School (NOVA IMS), NOVA University Lisbon, and Invited Professor at Université Paris-Dauphine PSL (France) and NOVA School of Science and Technology (NOVA FCT). He holds a Ph.D. and a B.Sc. in Economics from the University of Évora and an M.Sc. in Monetary and Financial Economics from ISEG, Technical University of Lisbon.

He is Director of the Postgraduate Programs in Financial Markets and Risks and Data Science for Finance, as well as the M.Sc. Program in Statistics and Information Management. He is an Integrated Member of MagIC (NOVA IMS Research Center), BRU-ISCTE, and CEFAGE-UE, and a member of the Scientific Experts Forum of the BBVA Pensions Institute (Spain).

Professor Bravo was recognised as one of the World's Top 2% Scientists in 2021, according to Stanford University's global ranking. His research has been published in leading journals, including the *Journal of Banking and Finance*, *Insurance: Mathematics and Economics*, *Expert Systems with Applications*, *Applied Soft Computing*, *Socio-Economic Planning Sciences*, and the *Journal of Pension Economics and Finance*, among others.

His research and consulting work focuses on longevity risk management, pension systems, ageing and social protection, retirement and savings behaviour, and credit risk management. He has served as a consultant to institutions such as the European Union (SOCIEUX+), Statistics Portugal, the Portuguese Ministry of Finance, the Ministry of Labour and Social Security, as well as insurance companies and pension fund associations. He was also an external member of the Interministerial Commission for the Reform of the Portuguese Social Security System.