Corporate Earnings Calls and Analyst Beliefs

Giuseppe Matera
EPFL and Swiss Finance Institute

June 1, 2026

Abstract

I study how linguistic features of corporate disclosure shape earnings expectations. I develop a methodology that uses large language models to generate counterfactual versions of earnings call transcripts that differ in language while holding underlying quantitative content fixed, creating a form of counterfactual variation that is impossible to obtain experimentally. I first show that textual features from earnings calls improve out-of-sample predictions of analyst forecasts and realized earnings above and beyond a rich set of quantitative fundamentals. I then intensify each transcript along six linguistic features—Forward Guidance, Jargon, Macro Focus, Uncertainty, Sentiment, and Confidence—and trace how these shifts move predicted analyst expectations. The results show that analysts move their forecasts in the right direction but weight linguistic dimensions unevenly: they place excessive weight on optimistic and macro-focused language and do not fully incorporate information embedded in language emphasizing risks and uncertainty. Linguistic shifts also change which fundamentals drive predictions: optimistic language amplifies the role of recent earnings surprises, while risk-laden language diminishes it, evidence that presentation shapes not just the level of analyst beliefs but the information they attend to.

*Giuseppe Matera is at EPFL and Swiss Finance Institute. Email: giuseppe.matera@epfl.ch. I am very grateful to my advisors, Andreas Fuster and Semyon Malamud, for precious advice and invaluable support. I also thank Swaminathan Balasubramaniam (discussant), Federico Baldi-Lanfranchi, Daniele Bianchi, Svetlana Bryzgalova, Francesco Celentano, Pierre Collin-Dufresne, Andrea Della Vecchia, Luise Eisfeld, Ruediger Fahlenbrach, Emanuele Luzzi (discussant), Roxana Mihet, Luca Pagliuca, Zhenkai Ran (discussant), Alberto Rossi, Lukas Schmid, Norman Schürhoff, and Xiao Yin for helpful comments, and seminar and conference participants at the 8th QMUL Economics and Finance Workshop for PhD and Post-doctoral Students, EPFL-Unil Brownbag Spring 2026, EPFL-Unil PhD Workshop Fall 2025, Liechtenstein Workshop on AI in Finance 2026, and SFI PhD Workshop 2025. I gratefully acknowledge support from the Swiss National Supercomputing Centre (CSCS) through project ID lp85 on the Daint-Alps system.

1 Introduction

How do economic agents form beliefs? In financial markets, much of what they respond to is hard information: earnings announcements, balance sheets, stock returns, and other quantitative signals about firm performance. Yet this information rarely arrives unaccompanied. Alongside the numbers they report, managers choose how confidently to discuss performance, how much uncertainty to express, how specific to be about the future, and which economic conditions to emphasize. These linguistic dimensions may affect the outcome of the communication, but they are difficult to quantify and even harder to separate from the underlying facts being communicated. The natural audience for this communication is equity analysts, who are tasked with distilling corporate disclosures into forward-looking assessments of firm value. This paper studies whether the linguistic content of corporate earnings calls shapes analyst earnings expectations beyond the hard information available in financial statements and market data.

Earnings calls are a natural setting for this question because they combine hard information with executives' own language. Around the same event, firms release accounting results, financial statements, and other quantitative disclosures, while executives discuss performance, risks, future prospects, and economic conditions. If analyst judgments are sensitive to how managers present results—not just what results are reported—then corporate communication becomes a channel through which their beliefs can be systematically influenced. This makes earnings calls especially useful for asking whether linguistic content shapes beliefs beyond the numerical information available to market participants. The setting is also empirically attractive because analysts issue and revise their forecasts intensively in the days following these calls, creating a concentrated window in which belief updating can be observed.

First, I show that textual features extracted from earnings calls improve the out-of-sample prediction of both analyst consensus forecasts and realized earnings. These gains are incremental to a comprehensive set of hard-information controls, including more than 300 quantitative controls and the most recent financial statements. Adding textual embeddings from earnings call transcripts to a gradient boosting model trained on these fundamentals reduces mean squared prediction error by 7-13%, depending on the outcome and forecast horizon. The improvements are statistically significant under the Clark and West (2007) test for

nested forecast comparisons, indicating that the predictive contribution of text is not driven by noise from estimating a larger model. A model that uses only textual features—with all numerical content masked—still delivers roughly half to two-thirds of the explanatory power of the fundamentals-only benchmark, suggesting that the linguistic content of earnings calls is informative on its own.

Second, I develop a novel text-morphing methodology to identify which dimensions of managerial communication drive these effects and whether analyst expectations are consistent with what those dimensions predict about realized earnings. The ideal experiment would rerun the same earnings call while varying only one aspect of executives' remarks—an experiment that is infeasible in practice. I simulate it by using a large language model to generate counterfactual versions of each earnings call that systematically intensify one of six linguistic dimensions at a time—Forward Guidance, Jargon, Macro Focus, Uncertainty, Sentiment, and Confidence—while preserving all quantitative content. These dimensions capture, respectively, the specificity of forward-looking statements, the use of specialized corporate language, the extent to which firm performance is tied to broader macroeconomic conditions, the emphasis on risks and unknowns, the positivity of tone, and the assertiveness of managerial claims. To ensure the credibility of these counterfactuals, I implement an automated validation procedure in which a second LLM instance verifies that each morphed text preserves the original quantitative content while exhibiting a clear shift along the intended dimension; variants that fail this check are discarded. Feeding these counterfactual transcripts into the predictive model trained on fundamentals and text and comparing predictions to those based on the original text yields a predicted treatment effect (PTE) for each dimension, measuring how a shift in language moves the model-implied forecast of analyst expectations and of realized earnings. The predictive model is trained exclusively on original transcripts; morphed texts enter only at the evaluation stage.

Comparing PTEs across the two outcomes reveals that analysts generally move their forecasts in the direction future earnings justify, but weight different dimensions of managerial communication systematically more or less than the information warrants. The results are consistent with analysts overweighting sentiment: morphing a transcript toward a more optimistic tone raises the predicted change in analyst expectations by approximately 35 basis points but is associated with only a 26 basis point increase in predicted realized earnings. The pattern for Macro Focus is even starker: it raises expected earnings by 6 basis points despite being associ-

ated with lower future realized earnings (-19 basis points). Conversely, the results are consistent with analysts underweighting uncertainty: language emphasizing risks and challenges predicts a 41 basis point decline in future realized earnings but moves analyst expectations by only 9 basis points. Analysts react in line with realized earnings to Jargon, while the gaps for confidence and forward guidance are also consistent with underweighting. Beyond these level effects, the morphing exercise reveals that uncertainty is the dominant driver of forecast disagreement among analysts, generating dispersion effects several times larger than any other dimension.

To benchmark these magnitudes, I consider that an interquartile shift in standardized unexpected earnings (SUE), a widely used measure of fundamental news, moves the predicted change in analyst expectations by roughly 18 basis points in the same model. The sentiment PTE of 35 basis points is thus roughly twice the effect of a typical earnings surprise, while the uncertainty PTE on realized earnings (-41 basis points) is more than twice as large. These are not marginal effects: shifts in the linguistic dimensions that carry no new quantitative information are associated with forecast movements comparable to, and in some cases exceeding, those associated with the most important piece of fundamental news in the earnings announcement.

The framework also sheds light on the mechanisms behind these patterns. Using SHAP values, which decompose each prediction into per-feature contributions, I compare how much weight the model places on each fundamental under the original versus the morphed transcript—holding the firm fixed, so that any change reflects language alone. The decomposition shows that linguistic shifts do not simply move the level of analyst expectations; they change the structure of how fundamentals are processed. Sentiment morphs increase the model's reliance on recent earnings surprises, leading analysts to treat recent good performance as more diagnostic of quality than it is, while uncertainty morphs reduce it, leading them to discount the latest results. The same balance sheet is effectively read differently depending on the language in which it is presented.

A central challenge is extracting meaningful linguistic representations from long, complex corporate disclosures without contaminating them with information about future outcomes. To represent the linguistic content of earnings calls, I compute textual embeddings using FinBERT, a transformer-based encoder pretrained on financial text with training data extending only to 2017 (Araci, 2019).

This choice is deliberately conservative. Recent work has raised legitimate concerns about look-ahead bias when modern large language models are applied to financial data, since these models may have been trained on the very outcomes the researcher is trying to predict (Glasserman and Lin, 2023; Sarkar and Vafa, 2024). By relying on an older, domain-specific encoder that I do not fine-tune on earnings call data, I ensure that the embeddings reflect general financial language understanding rather than memorized associations with specific firms or events in my sample. The results should therefore be read as a lower bound: more expressive modern encoders would likely capture additional linguistic structure and yield stronger findings. To further isolate language from quantitative content, I mask all numerical values, dates, and ticker-like strings before computing embeddings, forcing the model to rely on semantics, tone, and discourse structure alone.

This paper makes three contributions. First, it provides evidence that the linguistic content of corporate earnings calls carries incremental information about both analyst expectations and future realized earnings, beyond what a comprehensive set of quantitative fundamentals can capture. Managerial communication does not merely echo the hard data; it carries independent predictive content that equity analysts incorporate into their beliefs. Second, the paper introduces a text-morphing methodology that allows researchers to construct credible counterfactual corporate disclosures and to measure the predicted effect of specific linguistic dimensions on outcomes of interest. This advances beyond existing approaches based on sentiment dictionaries or aggregate tone scores by making it possible to intervene directly on the text along interpretable dimensions while holding quantitative content fixed, and it generalizes to other domains of corporate or institutional communication, such as central bank statements, regulatory filings, or sustainability disclosures. Third, the paper documents systematic patterns consistent with over- and underweighting of specific dimensions of managerial communication, and shows that linguistic variation changes not just the level of analyst expectations but the structure of how fundamentals are processed. Together, these findings offer a granular view of how corporate communication shapes belief formation in financial markets, and of where the gap between what analysts attend to and what actually predicts future earnings is largest.

Related Literature

This paper connects the literature on subjective belief formation with the emerging research that leverages large language models (LLMs) for social science inference. By using earnings calls as the textual input, the analysis constructs experimental environments that cannot be implemented through traditional designs.

Subjective beliefs and analyst expectations. A substantial body of research shows that equity analysts form expectations under cognitive constraints and frictions, and examines the determinants and consequences of these limitations (Diether et al., 2002; Livnat and Mendenhall, 2006; Bouchaud et al., 2019; Bordalo et al., 2019, 2024, 2025; Bastianello et al., 2024; Ben-David and Chinco, 2024; Décaire and Guenzel, 2023; Décaire et al., 2024). A related strand combines machine learning and textual analysis with finance to measure how beliefs form and update (Binsbergen et al., 2023; Charles and Sui, 2025; Charles and Kendall, 2025; Bybee, 2025; Giglio et al., 2022; Bastianello, 2022; Gormsen and Huber, 2025; Bianchi et al., 2024; Gabaix et al., 2025; Chen and Huang, 2024; Hailiang Chen and Peng, 2025; Lopez-Lira and Tang, 2023; Lv, 2024; Sarkar, 2025; Stolborg and Greenwood, 2025).[1]

In this paper, I take a complementary approach and examine analyst expectation formation using corporate earnings calls, which allows me to measure how the information environment of the call shapes analyst beliefs. A growing literature uses sell-side equity analyst reports as a primary data source to study related questions.[2]

LLMs as tools for empirical social science. Recent work shows that large language models can reproduce classic behavioral patterns, simulate experimental decision-making environments, and act as transparent proxies for human judgment (Horton, 2023; Hansen et al., 2024; Manning et al., 2024; Bhagwat et al., 2025; Ludwig et al., 2025; Kazinnik and Sinclair, 2025). They can also implement structured textual interventions, emulate survey-based expectation-formation exercises,

[1] A parallel literature studies the expectations of households and investors rather than analysts, typically using survey data; see, e.g., Greenwood and Shleifer (2014); Giglio et al. (2021); Dahlquist and Ibert (2024); Nagel and Xu (2022); De La O and Myers (2021, 2024); De La O et al. (2025); Ma et al. (2024). For more extensive reviews, see Kothari et al. (2016) and Giglio et al. (2025).

[2] For example, De Rosa (2024), Ke (2024), and Bastianello et al. (2024) combine these reports with I/B/E/S forecast data to study dimensions of analyst information processing, including attention and the structure of their mental models.

and capture heterogeneity in how different *in-silico* "agents" interpret the same information. I build on this work by using LLMs not only to measure textual content but to generate targeted counterfactual variation in corporate disclosures, enabling a structured analysis of how specific linguistic dimensions move analyst beliefs and relate to firm outcomes.

Financial disclosures and textual analysis. This paper also builds on research examining the informational role of corporate communication. A growing body of work analyzes how financial disclosures shape market participants' beliefs and decisions. Studies using corporate earnings calls—including Dzieliński et al. (2017), Hassan et al. (2019), Cohen et al. (2020), Li et al. (2021), Mamaysky et al. (2023), Meursault et al. (2023), Sautner et al. (2023), Cohen and Nguyen (2024), and Siano (2025)—investigate the linguistic and tonal features of managers' communication and their impact on investor reactions. Complementary research on 10-K filings, such as Loughran and Mcdonald (2011), Hoberg and Phillips (2010), Hoberg and Maksimovic (2015), and Hoberg and Phillips (2016), explores how disclosure content and structure transmit firm-specific information. Together, these studies highlight the central role of corporate communication in shaping how information is interpreted in financial markets.[3]

Contribution. By integrating these strands, this paper proposes a framework in which LLMs construct counterfactual versions of earnings calls to test how specific linguistic changes affect analyst belief updating. In doing so, the paper contributes to the literature on expectation formation, to the study of corporate disclosure, and to the methodological literature on LLMs as instruments for social-scientific experimentation.

Outline. Section 2 describes the data. Section 3 defines the target variables and describes the construction of textual embeddings. Section 4 presents the predictive framework and documents the incremental predictive contribution of earnings call text over quantitative fundamentals. Section 5 develops the text-morphing methodology and presents the treatment effects of counterfactual earnings calls. Section 6 concludes.

[3] See Hoberg and Manela (2025) for a more extensive review.

2 Data

Annual earnings announcements are scheduled events at which firms disclose fundamental information to the market, most notably the earnings figures that give the event its name. On these occasions, publicly traded firms typically hold an earnings call, during which senior management reviews recent financial results and provides forward-looking guidance to investors and equity analysts. It has also become common for firms to file their annual 10-K with the SEC within two days of the earnings release (Arif et al., 2019).

In forming their expectations, equity analysts draw on a wide range of data sources, build detailed models, and interact directly with firms. I assume that the financial statements contained in these 10-Ks enter analyst information set and shape their views on future firm performance. To isolate the predictive contribution of earnings call content beyond such hard information, I also control for a broad set of standard stock-level characteristics from Jensen et al. (2023), hereafter the JKP dataset.

2.1 Corporate Earnings Calls

I obtain earnings conference call transcripts from Capital IQ Transcripts, which provides verbatim records of corporate and institutional events. I retain all English-language transcripts for U.S. public companies announcing annual earnings. I use the full history of available calls, but coverage before 2008 is sparse and of lower quality, relying largely on other sources such as SeekingAlpha, so the sample begins in 2008. The sample covers the period 2008–2023 and comprises 35,203 annual earnings call transcripts for 5,370 companies, covering all forecast horizons considered in this study (one-, two-, and three-year-ahead forecasts). Section A of the appendix describes the dataset in detail.

An earnings call typically has two parts: prepared remarks by executives, followed by a question-and-answer (Q&A) session in which investors and analysts comment on, question, and at times challenge management. Because the Capital IQ data are organized at the speaker level, I can separate the two segments and restrict attention to the prepared management remarks, whose effect on analyst expectations is the focus of this paper.

Figure 1 shows that the number of annual earnings calls is stable over the sam-

ple at roughly 2,000 per year, as is their structure, with management remarks averaging about 3,000 words. The only notable deviations occur around the 2008 financial crisis and the onset of the Covid-19 pandemic, when these remarks were on average about 10% longer.

2.2 Analyst Forecasts

I measure analyst expectations using analyst-level earnings-per-share (EPS) forecasts from the Institutional Brokers' Estimate System (IBES) Detail Unadjusted History database, which records the forecasts submitted by individual sell-side analysts. I restrict the sample to U.S. companies and to one-, two-, and three-year-ahead forecasts.[4] The Unadjusted History file reports analyst–firm–date forecasts and realizations without retroactive adjustment for corporate actions such as stock splits.[5] Because such actions can introduce mechanical discrepancies between forecasts and realizations (Diether et al., 2002; Bouchaud et al., 2019), I restate all forecast and realized EPS values to the prevailing share count. I take cumulative share adjustment factors (`CFACSHR`) from the CRSP daily stock file (`dsf`) and merge them with IBES through the WRDS IBES–CRSP linking table (`ibcrsphist`), retaining only observations with a valid link at the forecast announcement date and, when a forecast falls outside the CRSP trading calendar, using the adjustment factor from the nearest prior trading date. I then apply standard data-quality filters, including corrections for temporal inconsistencies across the relevant announcement dates.[6]

The IBES forecasts and realizations enter the notation below as split-adjusted raw EPS, not yet as earnings yields. I convert them into earnings-to-price units only when constructing the target variables in Section 3, using the price prevailing at the time of the earnings call to avoid look-ahead bias. Scaling by price absorbs heterogeneity in firm size and price levels and makes comparisons across firms and over time more meaningful.

Table 1 reports the number of firms, brokers, and analysts in the final sample by horizon. Broker coverage is uneven across horizons: most brokers issue near-term forecasts and far fewer project three years out, so the number of observations

[4] All industries are included in the baseline sample, since the measures used in this paper are normalized using each firm's own past accounting figures as a benchmark.

[5] Evidence suggests that concerns about other retroactive changes and ID reshuffling are relatively minor in U.S. data, so the dataset can serve as a reliable proxy for market expectations. See Law (2023) for a comprehensive discussion of these limitations and their empirical relevance.

[6] See this WRDS support article for examples.

declines as the horizon lengthens.

2.3 Matching Forecasts to Earnings Calls

I link analyst forecasts to conference calls by matching Capital IQ transcripts to the IBES Unadjusted Detail file using the WRDS linking tables. To capture how the information released on the call feeds into analyst expectations, I keep all forecasts issued within 15 days of the call date. This window is motivated by the timing in Figure 2: forecast issuance spikes sharply in the days immediately after the call, so a 15-day window captures the large majority of event-related forecasts while limiting contamination from subsequent, unrelated news. I then form the consensus as the median of these forecasts rather than using the vendor-supplied consensus, which may pool forecasts issued outside this event window and would therefore blur the link between a given earnings call and the expectations it induces.

I exclude earnings call transcripts associated with more than one company, since these create ambiguous links between text and firm-specific outcomes and could induce spurious predictability.

2.4 Accounting and Market-Based Information

To measure the incremental predictive power of textual features, I first assemble a benchmark of accounting and market-based information that captures both realized fundamentals and market perceptions.[7]

2.4.1 Jensen–Kelly–Pedersen (JKP) stock characteristics

My goal is to gauge the predictive value of earnings call transcripts over and above the hard information available to analysts from financial statements and the stock market. Establishing this benchmark exactly is impossible, since I cannot observe what analysts actually know or think. I therefore proxy for their information set with what is arguably the richest publicly available characterization of a firm's quantitative state: the 153 market- and accounting-based characteristics of the JKP dataset (Jensen et al., 2023), together with the three primary financial statements and a freshly computed earnings surprise. This proxy is necessarily incomplete,

[7] In the machine learning literature, the outcome and predictors are typically referred to as the "target" and "features." I adopt this nomenclature throughout the paper.

and I do not claim it spans everything analysts know. The relevant question is weaker and testable, namely whether text predicts analyst expectations *beyond* this benchmark. If it does, then either text carries information absent from a very broad set of fundamentals, or analysts respond to soft information per se. Both of these are the object of this paper.

The JKP dataset includes 153 characteristics, which, following the taxonomy of Jensen et al. (2023), I group into 13 clusters spanning the major themes of the asset-pricing literature, such as profitability, investment, momentum, and value. These clusters provide a structured representation of the large cross-section of characteristics used in asset-pricing studies.

I merge these characteristics so that every conditioning variable reflects the most recent vintage available prior to the call. Because the JKP characteristics are constructed at monthly frequency, this is the snapshot from the month before the announcement, and I add their one-year changes. The one exception is SUE, the earnings surprise tied directly to the announcement, which I recompute as of the call date (see below). Two points follow from this timing. First, it keeps the conditioning set strictly point-in-time and rules out leakage from information released on or after the announcement. Second, it is conservative: the month-before snapshot omits any hard information that reaches analysts between that date and the call, so the benchmark is, if anything, weaker than the true analyst information set on the call date. Any such information that text appears to predict would in fact be hard information missing from my controls, which biases against finding an incremental role for language. The estimates below should therefore be read as a lower bound on the true incremental contribution of text.

Standardized Unexpected Earnings (SUE). SUE is a measure of earnings surprise, defined as the deviation of reported EPS from the analyst consensus prevailing before the release. The JKP data already include a SUE measure, but because I merge the dataset using information from the month before the call, that measure refers to the prior quarter rather than the quarter of interest. I therefore recompute SUE directly from IBES. Following Livnat and Mendenhall (2006), I use only estimates issued within 90 days before the call date, retaining each analyst's latest forecast to avoid stale information, and normalize the resulting surprise by the stock price at the report date.

2.4.2 Financial Statements

Financial statements are a central input to analyst work, since they reveal trends in accounting figures and provide the basis for constructing financial ratios. I include the three main statements filed in the 10-K: the income statement, balance sheet, and statement of cash flows.[8] I match these to the dataset described above by company identifier and fiscal year-end, and, as with the JKP characteristics, I add their one-year changes.

Because these variables differ in scale, distribution, and time variation, I rank-standardize each one within date to the $[-0.5, 0.5]$ interval before merging. This places all variables on a common, bounded scale that is robust to outliers and to time-varying dispersion.

3 Variable Definitions and Textual Embeddings

3.1 Analyst Expectation Formation around Earnings Calls

Analysts issue forecasts far more intensively in the days following an earnings call. Figure 2 shows a sharp spike in forecast issuance immediately after the call. Prior work attributes this surge to analysts processing the newly released financial statements and earnings figures. I argue that the way management presents this information, and the qualitative guidance it provides during the call, matters over and above the quantitative disclosures themselves. Figure 3 summarizes the timing of the earnings announcement, earnings call, analyst forecast revision, and subsequent earnings realization.

I track how analysts revise their beliefs along three dimensions. First, I measure the median expected change in forecasts, a robust summary of how expectations shift. Second, I assess whether the consensus overshoots or undershoots the earnings ultimately realized, which signals over- or under-reaction. Third, I measure the cross-sectional dispersion of forecasts to gauge disagreement among analysts. Together, these capture both the central movement in beliefs and the degree of consensus around it.

To formalize these measures, let $\xi_{i,t}$ denote firm i's raw EPS for the period just ended, realized and released at t. After the earnings call, analyst a issues forecast

[8] To retrieve the balance sheet, income statement, and statement of cash flows, I use this WRDS query.

$\widehat{\xi}_{a,i,t,h}$ for firm i's earnings to be released at horizon h. These individual forecasts aggregate into a consensus

$$F_{i,t}\left(\xi_{i,t+h}\right) := \text{median}_{a \in \mathcal{A}_{i,t,h}} \widehat{\xi}_{a,i,t,h},$$

where $\mathcal{A}_{i,t,h}$ is the set of analysts issuing forecasts for that firm-event-horizon. I reserve F for analyst forecasts throughout the paper. The objects $\xi_{i,t}$, $\widehat{\xi}_{a,i,t,h}$, and $F_{i,t}(\xi_{i,t+h})$ are raw EPS quantities; they become earnings-to-price measures only after division by the earnings-call-date price $P_{i,t}$. Equation 1 defines **Analyst Expected Change in Earnings**, the price-scaled difference between this consensus and the most recently realized value:

$$\underbrace{EC_{i,t,h}}_{\text{Expected Change in Earnings}} = \frac{F_{i,t}\left(\xi_{i,t+h}\right) - \xi_{i,t}}{P_{i,t}}. \tag{1}$$

Expectations vary across analysts even when all observe the same earnings call. This dispersion can reflect heterogeneous priors, risk tolerances, and models, and, crucially, differences in how the same language is interpreted. To study whether particular linguistic features are associated with more or less disagreement, I measure the cross-sectional dispersion of raw forecast changes, scaled by the stock price at the earnings call date for comparability with the other variables. I define **Forecast Disagreement** as

$$\underbrace{DISP_{i,t,h}}_{\text{Forecast Disagreement}} = \frac{\text{St. Dev.}_{a \in \mathcal{A}_{i,t,h}}\left(\widehat{\xi}_{a,i,t,h} - \xi_{i,t}\right)}{P_{i,t}}. \tag{2}$$

Finally, to assess how far features of corporate communication predict future fundamentals, I model the realized change in earnings. This quantity also anchors the *in-silico* counterfactual developed in Section 5, since it lets me compare how language moves analyst expectations against how it relates to what firms actually realize. I define **Realized Change in Earnings** as

$$\underbrace{RC_{i,t,h}}_{\text{Realized Change in Earnings}} = \frac{\xi_{i,t+h} - \xi_{i,t}}{P_{i,t}}. \tag{3}$$

All three target variables are trimmed at the 5th and 95th percentiles. Analyst

forecast data contain extreme observations that arise mainly from stale forecasts and data-recording errors rather than from genuine economic variation, and these can exert disproportionate influence on the estimates. Summary statistics for the trimmed variables appear in Table 2. The number of observations is smaller for Forecast Disagreement than for the other targets, since dispersion is defined only for firms covered by more than one analyst. I collect these targets as $Y^o_{i,t,h}$, where $o \in \mathcal{O} = \{EC, DISP, RC\}$ indexes, respectively, Analyst Expected Change in Earnings, Forecast Disagreement, and Realized Change in Earnings.

3.2 Representation in Textual Spaces: Earnings Call Embeddings

Text is inherently high-dimensional and context-dependent. To obtain compact numeric representations that preserve semantics and tone, I use textual embeddings.

Textual embeddings are continuous, fixed-length vector representations of text that place words, sentences, or documents in a geometric space where semantic similarity corresponds to proximity. The geometry of this space captures nuances such as synonymy, tone, and topical relatedness, providing a compact and meaningful summary of the original text (Sarkar, 2025).

State-of-the-art embeddings are produced by transformer encoders trained on large corpora with self-supervised objectives; they yield contextual representations in which the same word can map to different vectors depending on its surrounding context.

Representing long earnings call transcripts as a single vector presents two challenges: text length feasibility and look-ahead bias. Earnings calls are long and often exceed transformer input limits, which can make downstream tasks sensitive to how long texts are chunked and aggregated. To mitigate this, I focus on the management's prepared remarks, the most structured and information-dense segment of each transcript.

To represent each earnings call transcript as a single vector, I compute textual embeddings on non-overlapping segments of the transcript that fit the encoder's context window. The resulting segment-level vectors must be combined to form a document-level representation. There is no consensus on a uniquely superior pooling rule for long texts, and—consistently with prior practice—reasonable alternatives perform similarly in this setting. I therefore employ an approach that preserves as much information as possible, using the Hugging Face `transformers`

library: within each segment, I average the token representations from the last transformer layer across all tokens, then aggregate the segment vectors to obtain a single embedding per transcript. Results are qualitatively similar under common alternatives such as mean, max, or `[CLS]`-based pooling.

Throughout the paper, $\text{text}_{i,t}$ denotes the processed transcript for firm i at event time t, while $e^T_{i,t}$ denotes its transcript-level textual embedding.

A central threat to validity is that managers often mention metrics and accounting figures during the call, making soft information indistinguishable from quantitative cues in the textual embedding. To address this, I mask all numerals—accounting figures, guidance ranges, percentages, dates, quarter and year references, and ticker-like digit strings—using the `[MASK]` token provided by the Hugging Face tokenizer. This forces the model to rely on semantics, tone, and discourse structure rather than magnitudes or calendar markers (e.g., terms like "Q4 2022" or "+35%"), and prevents subtle forward references (such as "updated 2027 outlook") from acting as proxies for realized outcomes.

A further concern is that text models can "peek into the future" if *content leakage* lets them read explicit future cues (dates, magnitudes), or *temporal leakage* lets them learn from documents that occur after the prediction target (Sarkar and Vafa, 2024). Either failure mode would inflate performance and weaken the credibility of the counterfactual exercise. To address this, I use ProsusAI/FinBERT (Araci, 2019), a finance-specific pretrained encoder whose training data extend only to 2017, which I do not fine-tune on my earnings call dataset. This model was trained on general-domain and financial news corpora, not on earnings-call transcripts, and its upstream training data predate my evaluation windows, attenuating the risk of look-ahead bias.

4 Methodology

This section describes the predictive framework used to assess whether textual features from earnings calls carry incremental information about analyst expectations and realized earnings beyond quantitative fundamentals.

To address these questions, I train and evaluate two model specifications. The first (i) is a stock characteristics-only model, which uses the JKP dataset together with the most recent financial statements (in levels and one-year changes), the prevailing abnormal stock return at the earnings call date, and the most recent

Standardized Unexpected Earnings. The second (ii) augments this baseline with textual embeddings extracted from earnings call transcripts. The two specifications comprise 446 and 1,214 features, respectively. The 446 hard-information features start from the 153 JKP characteristics measured in levels and one-year changes (306 variables), then add financial-statement variables, abnormal returns, and SUE; the text-augmented specification adds a 768-dimensional FinBERT transcript embedding, yielding 1,214 total features.

I employ a Gradient Boosting regressor, a machine-learning model which accommodates non-linearities and interactions while offering competitive computational performance. The analysis is conducted separately by forecast horizon, in order to assess whether the contribution of transcript-based textual features is concentrated in the near term or persists at longer horizons.

For each forecast horizon and target variable, the most recent 30% of observations are held out as an out-of-sample test set, while the remaining 70% are used for model fitting and hyper-parameter tuning. To avoid overfitting and to ensure robustness of the selected hyper-parameters, I implement K-fold cross-validation within the training set.

As discussed above, all numerical values and dates are comprehensively masked, the text features contain no forward-looking cues, eliminating content-based look-ahead. In addition, the test data for these predictive exercises are well after the last time stamp of the FinBERT training sample, ensuring that the embeddings are applied strictly out-of-sample.

Out-of-sample performance is evaluated using mean squared error (MSE) and R-squared; for readability, I omit the "out-of-sample" label in what follows.

4.1 The Informational Content of Managerial Communication

To begin, I investigate the extent to which standard numeric firm characteristics can predict analyst expected earnings change and realized earnings change, and whether textual features extracted from earnings calls carry predictive power for the same outcomes. To do so, I set up a simple comparative framework.

Let $\{(Y^o_{i,t,h}, x^S_{i,t}, e^T_{i,t})\}_{i,t}$ be the time-ordered sample for outcome $o \in \mathcal{O}$ and forecast horizon h, where $x^S_{i,t} \in \mathbb{R}^{p_S}$ are numeric stock characteristics, accounting variables, and other hard-information controls, and $e^T_{i,t} \in \mathbb{R}^{p_T}$ are textual embeddings of earnings call transcripts. I define $x_{i,t} := (x^S_{i,t}, e^T_{i,t})$ and adopt a single generic

modeling framework,

$$Y^{o}_{i,t,h} \;=\; \widehat{g}^{(p)}_{o,h}(z^{(p)}_{i,t}) + \varepsilon^{o}_{i,t,h},$$

where $\widehat{g}^{(p)}_{o,h}$ denotes the trained model for outcome o, horizon h, and predictor set p. For notational simplicity, I suppress the forecast horizon h when the model is estimated separately by horizon. I instantiate the framework by varying only the predictor set:

(S) $z^{(S)}_{i,t} = x^{S}_{i,t}$.

(T) $z^{(T)}_{i,t} = e^{T}_{i,t}$.

(ST) $z^{(ST)}_{i,t} = (x^{S}_{i,t}, e^{T}_{i,t})$.

Figure 4 reports the out-of-sample R-squared for the (T) and (S) specifications across forecast horizons (one-, two-, and three-year-ahead) and three outcome variables: *Expected Change in Earnings*, *Forecast Disagreement*, and *Realized Change in Earnings*.

For Expected Change and Realized Change, predictive power declines with the forecast horizon under both specifications. This pattern is consistent with the well-documented finding that earnings become increasingly difficult to forecast at longer horizons, as the signal-to-noise ratio deteriorates when predicting more distant outcomes (see, e.g., Dessaint et al., 2024).

In contrast, the R-squared for Forecast Disagreement remains roughly stable across horizons for both models. This could be explained by disagreement largely reflecting structural differences in how analysts interpret the same information, rather than the precision of any individual forecast. Since these interpretive differences do not vanish at longer horizons, the predictability of disagreement remains largely unaffected by the forecast distance.

Unsurprisingly, the (S) specification achieves higher R-squared levels than the (T) specification across all panels, since fundamentals directly capture the quantitative drivers of earnings dynamics. However, the text-only model—in which all numerical information within the transcript is masked—delivers a sizable share of the fundamentals' explanatory power, ranging from roughly one-half to two-thirds depending on the outcome and horizon. This suggests that managerial language encodes economically meaningful forward-looking information that may substantially overlap with, but is not directly derived from, the numerical data.

Empirical Test of Analyst Attention to Managerial Communication. A natural question is whether fundamentals simply subsume the information conveyed by text, or whether the textual component adds predictive power above and beyond quantitative variables. If the former is the case, the (ST) specification should yield no statistically significant improvement in out-of-sample R-squared relative to the (S) specification.

To answer this question, I apply the MSPE-adjusted test for nested models developed by Clark and West (2007). This approach is particularly suitable in settings where models are nested and traditional measures such as out-of-sample R-squared may be noisy or difficult to interpret in small samples.

The test compares the squared forecast errors of the competing models but adjusts for the mechanical disadvantage faced by the larger, nested specification: additional estimated parameters can inflate prediction error even when the extra information is genuinely useful. The adjustment removes this noise-induced penalty and evaluates whether the average difference in accuracy is positive, with the resulting test statistic defined in Eq. 4.

$$CW = \frac{\bar{\tilde{d}}}{\sqrt{\widehat{\mathrm{Var}}(\tilde{d}_{i,t})/N}}, \qquad \bar{\tilde{d}} = \frac{1}{N} \sum_{(i,t)\in\mathcal{O}_{\text{test}}} \left[(e_{i,t}^{(r)})^2 - (e_{i,t}^{(u)})^2 + \left(\hat{y}_{i,t}^{(r)} - \hat{y}_{i,t}^{(u)} \right)^2 \right] \quad (4)$$

where $e_{i,t}^{(r)} = y_{i,t} - \hat{y}_{i,t}^{(r)}$ and $e_{i,t}^{(u)} = y_{i,t} - \hat{y}_{i,t}^{(u)}$ are the forecast errors of the restricted and unrestricted models, $N = |\mathcal{O}_{\text{test}}|$ is the number of test observations, and $\widehat{\mathrm{Var}}$ is estimated with HAC standard errors.

I report Clark–West statistics for nested forecast comparisons in Table 3. The statistics are computed from the adjusted loss differential using heteroskedasticity- and autocorrelation-robust (Newey–West) standard errors appropriate for h-step-ahead forecasts. The alternative is one-sided (the augmented model is better). Across targets and horizons, the Clark and West t-statistics are positive and large, indicating that adding transcript-based textual features reduces forecast MSE relative to the benchmark. The implied mean-squared-error reductions are economically meaningful, typically amounting to about 7–13%.

4.2 Measuring the Impact of Fundamental News

The machine learning models used to predict the outcomes of interest are highly non-linear, and it is therefore not straightforward to describe how different variables impact model predictions. Here, I focus on one variable that existing work has shown to strongly matter for stock market and analyst responses to earnings releases, namely the Standardized Unexpected Earnings (SUE). The SUE captures the surprise component of announced earnings relative to the prior analyst consensus and is among the most robust predictors of post-earnings-announcement drift and forecast revisions (Livnat and Mendenhall, 2006; Bouchaud et al., 2019). By examining how model predictions vary with SUE, I establish a quantitative benchmark for the magnitude of "fundamental news" effects, against which I later compare the predicted treatment effects of linguistic morphing. Figure 5 reports partial dependence plots (PDPs) of the one-, two-, and three-year-ahead targets on SUE. PDPs trace the marginal effect of a single predictor on the model's output, averaging over the empirical distribution of all other covariates. They thus provide a transparent view of how predicted outcomes vary with earnings surprises.

The patterns reveal familiar asymmetries in analyst reactions: (a) realized changes in earnings exhibit an asymmetric, step-like increase after large negative surprises, plateauing and then slightly tapering for very positive surprises, consistent with mean reversion; (b) expected forecast changes decline with SUE—negative surprises are followed by upward revisions, while positive surprises are tempered by smaller or even negative adjustments; and (c) forecast disagreement is U-shaped, lowest near zero surprise and higher at the extremes, suggesting greater dispersion when news is unusually good or bad.

I then compute the interquartile shift (from the 25th to the 75th percentile), a "typical" move in SUE, as a benchmark against which, in later sections, I compare the effects of language morphing with large language models. This will allow me to assess whether soft information in earnings calls can move forecasts in ways comparable to fundamental earnings news.

Discussion. The results provide clear evidence that equity analysts pay attention to the textual information in earnings calls and that this information has incremental predictive value beyond traditional quantitative indicators. Across specifications, augmenting benchmark models based on accounting variables and stock

characteristics with textual features from earnings calls improves predictive performance. While out-of-sample R^2 remains a noisy metric, the Clark and West (Clark and West, 2007) test consistently indicates that the incremental forecasting value of the linguistic component is genuine rather than spurious.

These findings imply that analysts do not rely exclusively on hard data when forming expectations; they also process qualitative and contextual dimensions of managerial communication. From an informational-efficiency perspective, this is potentially beneficial: even models predicting realized earnings improve when textual features are included, suggesting that the linguistic dimensions emphasized in earnings calls embed information about fundamentals that is not fully captured by standard predictors.

At the same time, the nature of analyst attention raises important open questions. It remains to be established whether analysts focus on the same linguistic dimensions that predict realized earnings—thereby incorporating value-relevant information in a manner consistent with rational processing—or whether attention is drawn to different, potentially more salient but less informative aspects of corporate communication. Disentangling these possibilities is essential to determine whether linguistic dimensions primarily operate as carriers of genuine information or as channels for behavioral bias in financial markets.

5 Counterfactual Earnings Calls with LLMs

The ideal experimental setting for isolating the effect of language on analyst perception would resemble a controlled environment in which a group of equity analysts resets their priors and issues new forecasts in response to alternative linguistic formulations of the same earnings call. Such a laboratory-style manipulation is infeasible in practice. To address this challenge, I adopt a novel in-silico approach that leverages recent advances in natural language processing (NLP). While early NLP techniques enabled the transformation of textual data into vector representations suitable for econometric modeling, they lacked the capacity to simulate variations in soft information or tonal emphasis. The emergence of large language models (LLMs) overcomes this limitation by enabling the systematic generation of semantically equivalent, but stylistically distinct, versions of executives' earnings call presentations. This technological development allows for the construction of a counterfactual framework in which the informational content remains fixed, while

linguistic dimensions are varied exogenously—thus permitting a thorough analysis of language effects on analyst behavior.

To do so, I generate synthetic variants of management speech using large language models (LLMs) implemented through the Hugging Face Transformers library. Specifically, I employ Meta's LLaMA 3 70B-Instruct model, a chat-optimized transformer fine-tuned for instruction-following and coherent text generation in conversational contexts.

For each document, I generate the output paragraph by paragraph in sequential forward passes, ensuring internal consistency and preserving global structure. To control the realism and interpretability of the generated texts, I apply prompt engineering techniques that guide the model to preserve numerical content (e.g., financial figures) while varying qualitative language. Specifically, I structure prompts in a way that keeps key values unchanged and induces linguistic transformations within a narrow range of word count variation, thereby enabling controlled comparisons between original and modified statements. Section B in the appendix reports the language-generation settings, and Section C documents the prompts.

This setup allows for systematic manipulation of the linguistic dimensions and tone of the earnings call content, holding constant the quantitative core of the message. Look-ahead bias does not affect my results because I train the predictive model ex ante, apply modifications through prompting to the LLM, and evaluate performance exclusively on out-of-sample morphed embeddings.

To ensure the credibility of the generated counterfactuals, I implement an automated LLM-as-a-judge validation procedure, described in Appendix D, that classifies each morphed transcript as acceptable or inadequate; variants that fail the quality check are discarded.

5.1 Linguistic Dimensions and Counterfactual Earnings Calls

To isolate the effect of language on analyst forecasts, I generate counterfactual versions of earnings call transcripts in which the underlying quantitative content is held constant, but the linguistic dimensions are systematically modified.

The goal of this exercise is to assess how much of the analyst reaction can be attributed to style and delivery, rather than to substance. Specifically, I generate versions of the same earnings call that vary according to a number of linguistic characteristics, and evaluate how the predictive model trained on historical data

responds to each variant. The citations below motivate each linguistic feature by linking it to prior evidence on managerial communication, disclosure, and analyst or market responses. The linguistic characteristics I use are:

- **Forward Guidance**: Reflects how specific and forward-looking the company's statements are. Stronger guidance is elicited by prompting clear, actionable expectations, directional indicators, and strategic intent, avoiding vague or generic phrasing.

  (Call et al., 2024; Bozanic et al., 2018; Anilowski et al., 2007)

- **Jargon**: Indicates the degree of technical or finance-specific language used. Higher technical jargon is introduced through prompts that incorporate institutional and domain-specific terminology without changing the underlying facts.

  (Li, 2008; Miller, 2010)

- **Macro Focus**: Captures the emphasis placed on broader macroeconomic and industry-wide conditions. The model is guided to highlight external forces, such as inflation, regulation, or global trends, and to link them explicitly to the firm's performance.

  (Song and Stern, 2025; Hassan et al., 2019)

- **Uncertainty**: Reflects the visibility and emphasis of risk-related language. To increase risk salience, prompts guide the model to explicitly highlight uncertainties, market challenges, and operational risks in a measured and professional tone.

  (Kravet and Muslu, 2013; Lyle et al., 2023)

- **Sentiment**: Denotes the overall emotional tone of the message. The text is modified to sound more positive and optimistic by emphasizing momentum, wins, and enthusiasm for the company's future.

  (Price et al., 2012; Huang et al., 2014)

- **Confidence**: Refers to how assertive and in-control the speaker sounds. This is adjusted by prompting the model to adopt the tone of confident executives, using decisive language, avoiding hedging, and conveying authority over strategy and performance.

(Mayew and Venkatachalam, 2012; De la Parra and Gallemore, 2024)

Because the generated earnings calls preserve quantitative content while varying targeted linguistic dimensions, changes in the model's prediction can be interpreted as model-implied effects of soft information. This setup quantifies how targeted linguistic features move analyst perceptions and whether those shifts align with or deviate from future fundamentals.

I illustrate the morphing exercise in Figure 6, where I project both the original and morphed transcripts into a two-dimensional semantic space. Each panel corresponds to a distinct linguistic attribute: Forward Guidance, Jargon, Macro Focus, Uncertainty, Sentiment, and Confidence, with the original baseline in the bottom panel. The heatmaps reveal that morphing pushes transcripts toward the extremes of these attributes, generating clearly differentiated clusters. This demonstrates that the procedure succeeds in producing systematically varied earnings calls, thereby enabling the in-silico experimental design.

Table 4 further illustrates how the original text of an earnings call can be systematically morphed along distinct linguistic dimensions. Each row presents a representative passage from an original managerial statement alongside a transformed version that amplifies a specific linguistic feature. The six dimensions, i.e., Forward Guidance, Jargon, Macro Focus, Uncertainty, Sentiment, and Confidence, capture complementary aspects of managerial communication that may influence analyst interpretation of firm fundamentals. These examples serve to clarify the nature of the textual transformations applied in the empirical analysis and to make the corresponding treatments intuitively interpretable.

5.2 Predicted Treatment Effects (PTEs) induced by Linguistic Dimensions

This section converts the LLM-generated linguistic morphs into predicted treatment effects (PTEs). For each earnings call, I compare the model's prediction under the morphed transcript with the prediction under the original transcript, holding the same stock characteristics and accounting controls fixed. Formally, for the earnings call transcript from company i at time t, each linguistic dimension $d \in \mathcal{D} = \{\text{Forward Guidance}, \text{Jargon}, \text{Macro Focus}, \text{Uncertainty}, \text{Sentiment}, \text{Confidence}\}$ gen-

erates the counterfactual transcript

$$\widetilde{\text{text}}_{i,t}^{d} = \mathcal{M}_d(\text{text}_{i,t}), \tag{5}$$

where $\mathcal{M}_d(\cdot)$ is the LLM-based transformation associated with dimension d. I embed the morphed transcript as $\widetilde{e}_{i,t}^{T,d}$ and compute the *Predicted Treatment Effect (PTE)* for outcome o and horizon h as

$$\text{PTE}_{i,t,h}^{o,d} = \widehat{g}_{o,h}^{(ST)}\left(x_{i,t}^{S}, \widetilde{e}_{i,t}^{T,d}\right) - \widehat{g}_{o,h}^{(ST)}\left(x_{i,t}^{S}, e_{i,t}^{T}\right), \tag{6}$$

where $x_{i,t}^{S}$ denotes the hard-information controls, $e_{i,t}^{T}$ is the embedding of the original transcript, and $\widehat{g}_{o,h}^{(ST)}$ is the trained gradient boosting model for outcome o and horizon h using both hard information and textual embeddings. Since only the transcript is altered, the PTE isolates the model-implied effect of the targeted linguistic dimension.

Figures 7 and 8 present the average predicted treatment effects (PTEs), expressed in basis points, for each linguistic morph applied to earnings calls, with observations pooled across all forecast horizons. These effects quantify how shifts in communication style—independent of any change in firm fundamentals—alter analyst forecasts and realized outcomes. To benchmark the magnitude of these effects, I also report the model-implied response to an interquartile change in fundamental earnings news, defined as the difference in predicted outcomes when standardized unexpected earnings (SUE) move from the 25th to the 75th percentile of its distribution, as computed in prior sections.

Figure 7 reports the average predicted treatment effects on analyst expected earnings change and realized earnings change across linguistic dimensions. For *Sentiment*, morphing a transcript toward a more optimistic tone raises predicted analyst expectations by approximately 35 basis points while being associated with only a 26 basis point increase in predicted realized earnings—a gap of 9 basis points consistent with overweighting of positive tone. The pattern for *Macro Focus* is even starker: it raises predicted analyst expectations by 6 basis points despite being associated with a 19 basis point *decline* in predicted realized earnings. By contrast, the results for *Uncertainty* are consistent with systematic underweighting: risk-laden language is associated with a 41 basis point decline in predicted realized earnings but moves analyst expectations by only 9 basis points in the same

direction. Analysts react more in line with realized earnings for *Jargon*, while the gaps for *Confidence* and *Forward Guidance* are also consistent with underweighting. To benchmark these magnitudes, the model-implied response to an interquartile shift in SUE—the most important piece of fundamental news—is approximately 18 basis points for analyst expectations, implying that the reaction to sentiment (35 basis points) is roughly twice the typical fundamental news effect, and the uncertainty gap between realized and expected outcomes (32 basis points) exceeds it by a similar margin.

Figure 8 isolates the predicted treatment effects of each linguistic dimension on *forecast disagreement*, capturing how language shapes the dispersion of analyst expectations rather than their central tendency. Here, "Uncertainty" again stands out as the dominant driver, generating dispersion effects several times larger than any other dimension. This finding aligns with the theoretical expectation that ambiguous communication amplifies interpretive heterogeneity among analysts, leading to greater disagreement. However, confident, sentiment-rich, and globally oriented discourse also increase forecast dispersion, albeit to a lesser extent. This pattern suggests that even positive or outward-looking linguistic dimensions can introduce divergent interpretations among analysts, possibly because such language allows multiple readings of firms' strategic intent or the macroeconomic implications of managerial optimism. Together, Figures 7 and 8 underscore that managerial linguistic dimensions not only shift the level of analyst forecasts but also shape the distribution of beliefs across forecasters, revealing a dual channel through which corporate communication affects market expectations.

5.3 Determinants of Predicted Treatment Effects

To better understand when and for which firms soft information exerts a stronger influence on analyst expectations, Table 7 relates the PTEs described in Section 5.2 to market- and firm-level characteristics. The dependent variable is the PTE on analyst expected earnings changes induced by each linguistic morph, while the explanatory variables capture prevailing market uncertainty and cross-sectional firm attributes.

The estimates reveal heterogeneity at multiple levels. At the market level, higher aggregate uncertainty is associated with PTEs that are larger in absolute value: the uncertainty PTE becomes more negative, while the sentiment, confidence, and for-

ward guidance PTEs become more positive, indicating that analysts become more responsive to soft information in riskier environments. At the firm level, characteristics related to risk, valuation, and maturity, such as earnings volatility, book-to-market, size, and firm age, shape these sensitivities differently across morphs. Overall, the results suggest that soft information interacts with the information environment and firm fundamentals rather than moving expectations uniformly.

5.4 How Language reshapes Attention Allocation

The previous sections show that morphing managerial language moves predicted analyst expectations even when quantitative information is held fixed. This section asks whether these effects are purely additive, or whether soft information also changes the model's use of hard information.

If linguistic features operate as soft information, they should affect not only forecast levels but also the weight placed on fundamentals. The same earnings surprise, leverage ratio, or profitability measure may receive different predictive weight depending on whether management presents the results with optimism, uncertainty, confidence, or macroeconomic emphasis.

This hypothesis connects naturally to limited attention models (Hirshleifer and Teoh, 2003; Peng and Xiong, 2006), in which agents selectively allocate cognitive resources across available signals, and to research on presentation effects (Kahneman and Tversky, 1984), which shows that identical objective information can be processed differently depending on how it is presented.

The empirical strategy exploits the counterfactual framework developed in Section 5. For each earnings call, I compare how the trained machine-learning model assigns importance to hard-information variables under the original transcript versus morphed versions of the same transcript. Because the firm, date, and fundamentals are fixed, changes in feature contributions reveal whether the linguistic intervention alters the model-implied role of fundamentals in predicting analyst expectations.

5.4.1 Measuring Attention Allocation

To measure these changes, I adopt the framework of Shapley values (Lundberg and Lee, 2017). Intuitively, SHAP values assign credit to each input variable for a given prediction: they indicate how much each feature pushes the prediction

above or below a baseline value. In machine learning applications, SHAP values decompose a prediction into additive feature-level contributions. For the gradient boosting models estimated above, I compute TreeSHAP values (Lundberg et al., 2020), which efficiently calculate exact Shapley values for tree-based models.

For a given observation (i, t) with textual embedding $e^T_{i,t}$, hard-information controls $x^S_{i,t}$, and prediction $\hat{y}_{i,t}$, the SHAP decomposition is

$$\hat{y}_{i,t} = \phi_0 + \sum_{j \in \mathcal{T}} \phi_j(x_{i,t}) + \sum_{k \in \mathcal{S}} \phi_k(x_{i,t}), \tag{7}$$

where $\mathcal{T}$ indexes textual embedding features, $\mathcal{S}$ indexes hard-information features, ϕ_0 is the baseline prediction, $\phi_j(x_{i,t})$ is the contribution of textual feature j, and $\phi_k(x_{i,t})$ is the contribution of hard-information variable k.

A natural concern is that morphing the text may simply increase the total importance of textual features, mechanically reducing the aggregate contribution of hard-information variables. This would show that text matters more, but not that attention is reallocated *within* fundamentals. To isolate redistribution across fundamentals, I normalize the absolute SHAP value of each fundamental by the total absolute SHAP contribution across all fundamentals:

$$\text{RelativeSHAP}_{i,t,k} = \frac{|\phi_k(x_{i,t})|}{\sum_{\ell \in \mathcal{S}} |\phi_\ell(x_{i,t})|}. \tag{8}$$

This measure captures fundamental k's share of total fundamental-based importance, with $\sum_{k \in \mathcal{S}} \text{RelativeSHAP}_{i,t,k} = 1$. For each linguistic dimension d and hard-information variable k, I then compute

$$\Delta\text{RelativeSHAP}^{(d)}_{i,t,k} = \text{RelativeSHAP}^{(d)}_{i,t,k} - \text{RelativeSHAP}^{(0)}_{i,t,k}, \tag{9}$$

where superscript (0) denotes the original transcript and superscript (d) denotes the transcript morphed along linguistic dimension d. Both terms are computed for the same firm-date observation and differ only in the transcript shown to the model.

The model being explained is trained to predict analyst consensus forecasts, not realized earnings. The SHAP values should therefore be interpreted as a decomposition of the model's prediction rule for analyst expectations, rather than as direct evidence of analyst cognitive processes. The identification of language effects comes from the counterfactual morphing design; SHAP is used to interpret

the channels through which the trained model maps linguistically altered disclosures into predicted analyst beliefs.

Before studying reallocation within hard information, Table 5 shows that all six linguistic morphs increase the model's reliance on textual features relative to fundamentals. This confirms that the morphing procedure generates meaningful variation. The increase is largest for uncertainty and smallest for jargon, indicating that the amount of qualitative information added by each morph differs substantially. This also motivates the normalization above: without relative SHAP values, changes could simply reflect a greater overall role for text rather than a shift in how fundamentals are weighted.

Table 6 illustrates the mechanism using Standardized Unexpected Earnings (SUE), a salient measure of recent earnings news. The results reveal three patterns. First, forward-looking language reduces the relative importance of SUE: forward guidance and uncertainty morphs lower SUE's share of fundamental importance by 1.7 and 1.6 percentage points, respectively. This suggests that when management emphasizes future prospects or risks, the model relies less on backward-looking earnings news to predict analyst expectations. Second, sentiment strongly increases the importance of SUE, by 2.4 percentage points ($t = 11.3$). Optimistic language therefore appears to make recent earnings surprises more diagnostic in the model's prediction rule, consistent with representativeness-based reasoning. This provides a mechanism for the overweighting of sentiment documented above: optimistic soft information amplifies the role of recent performance in predicted analyst belief updating. Third, macro-focused language increases the importance of SUE by 1.6 percentage points while reducing attention to firm-specific growth measures, suggesting that macro-focused soft information makes realized outcomes more salient and firm-specific forward-looking investments less central.

Overall, the evidence indicates that linguistic variation does not merely shift predicted analyst expectations additively. It also changes the model-implied mapping from fundamentals to expectations. In this sense, language reshapes attention allocation: the same quantitative information receives different predictive weight depending on how managers present the earnings call.

6 Conclusions

Corporate communication is often viewed as a soft complement to the hard data in financial statements. This paper shows that the distinction between substance and presentation is consequential: how managers present results–the tone, emphasis, and linguistic choices in their earnings call remarks–systematically shapes analyst expectations in ways that diverge from what the same linguistic features predict about future realized earnings. The results are consistent with analysts overweighting optimistic soft information and underweighting risk-laden language, and these patterns are not merely level shifts–they alter the structure of how quantitative fundamentals are processed and weighted.

These findings carry practical implications. For corporate managers, the results suggest that rhetorical choices in earnings calls are not cheap talk: they measurably move analyst forecasts, and the text-morphing framework provides a tool for experimenting with alternative linguistic dimensions in silico before committing to a communication strategy. For analysts, the results highlight a specific vulnerability–the tendency to discount complex risk language while amplifying clear optimistic signals–that could be addressed through more disciplined information processing protocols. For regulators concerned with the quality of price discovery, the evidence that soft information affects not just the level but the dispersion of analyst beliefs suggests that the linguistic dimension of corporate disclosure deserves attention alongside its quantitative content.

The methodology developed here is not specific to earnings calls. The same framework–generating targeted linguistic counterfactuals with large language models, validating them automatically, and measuring their predicted effects through a model trained on observed outcomes–can be applied to any setting where language accompanies quantitative disclosure. Central bank press conferences, IPO prospectuses, regulatory filings, and sustainability reports all combine hard information with interpretive soft information, and the question of whether audiences respond to the right dimensions of that soft information is as relevant in those domains as it is here. More broadly, the paper illustrates how advances in language modeling can be harnessed not just to measure textual content, as a growing literature already does, but to construct the counterfactual variation that empirical research in finance has traditionally lacked.

7 Figures

FIGURE 1: **Earnings Calls: Number per Year and Average Length**

This figure illustrates the combined dataset of corporate earnings calls and associated analyst forecasts. The blue bars (left axis) represent the total number of earnings calls held each calendar year, while the red line (right axis) indicates the average number of words in the management remarks section per call. The sample includes only earnings calls followed by at least one individual analyst forecast, across all forecast horizons. Each earnings call typically consists of two parts: the prepared management remarks and the question&answer session with analysts and investors. Only earnings call transcripts containing non-empty management remarks are retained in the dataset.

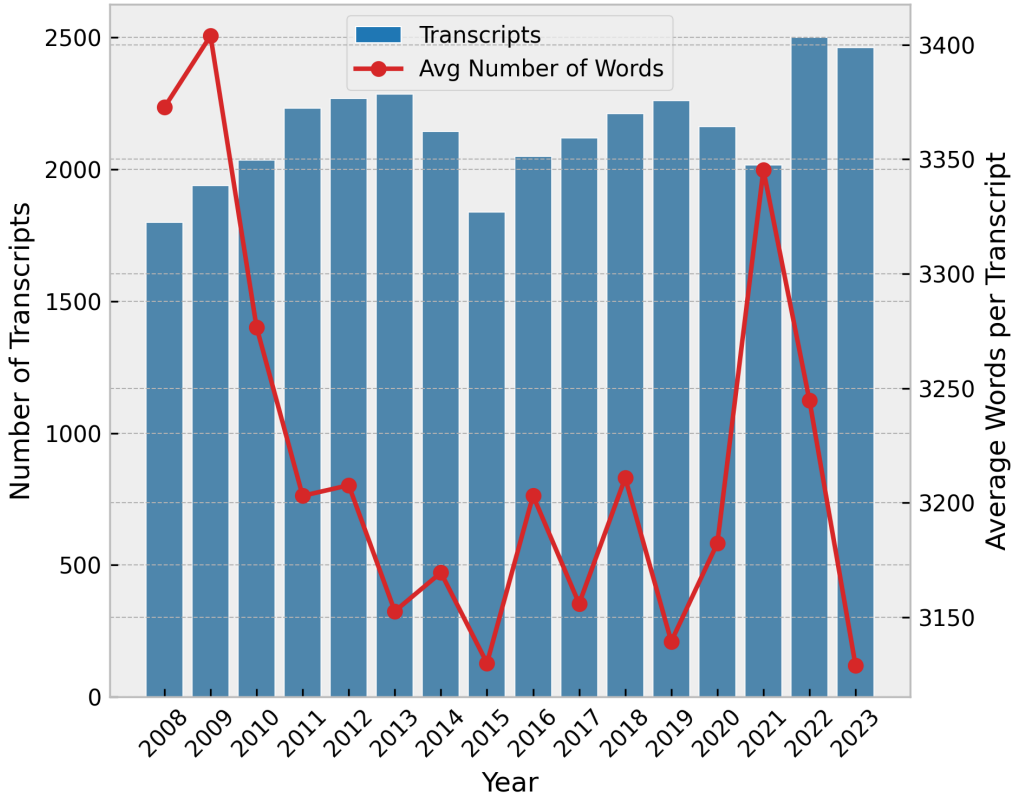


FIGURE 2: **Timing of Analyst Forecasts Around Earnings Calls**

This figure depicts the distribution of analyst forecast releases in the weeks surrounding earnings calls. It reports the weekly proportion of individual analyst forecasts with annual periodicity, across all forecast horizons, relative to the total number of forecasts issued within the seven-week window around the call. Week 0 corresponds to the week beginning on the day of the earnings call. Since earnings calls occur quarterly, forecasts issued more than seven weeks before or after a call are associated with the preceding or subsequent event. The sample covers the period 2008–2023.

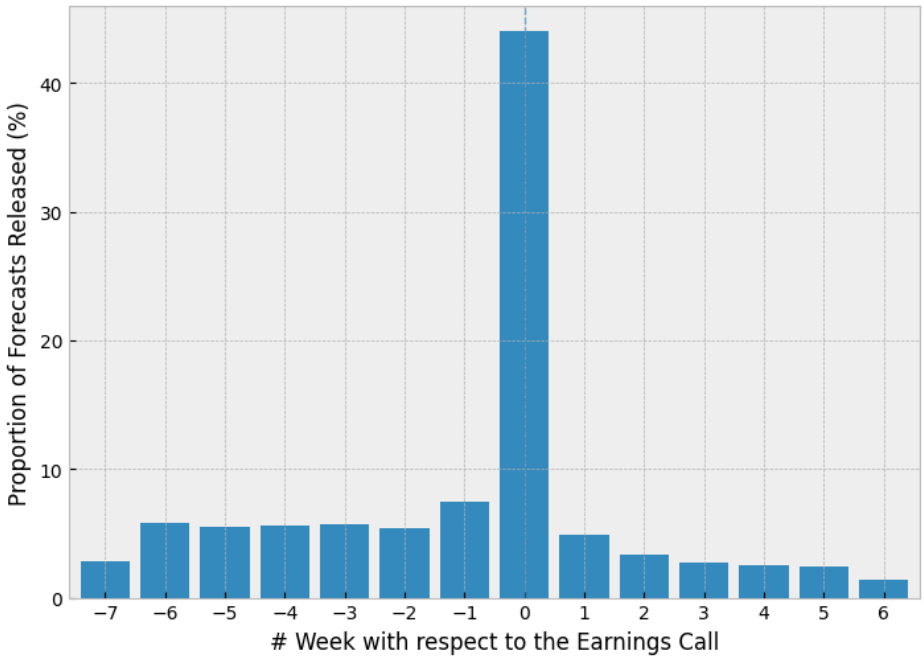


FIGURE 3: **Timeline of the Earnings Announcement Cycle**

This figure illustrates the standard sequence of events in the corporate earnings announcement cycle. At time t, the firm reports realized earnings for the period just ended ($\xi_{i,t}$) and files its Form 10-K with the SEC, which includes the income statement, balance sheet, and statement of cash flows. Around the same time, management holds an earnings conference call with investors and equity analysts to discuss recent performance and provide forward-looking guidance. Analysts then revise their earnings forecasts for the horizon-h period ($\xi_{i,t+h}$) in light of the new information. At time $t+h$, where h denotes the forecast horizon, the firm reports the realized earnings for that forecasted period ($\xi_{i,t+h}$).

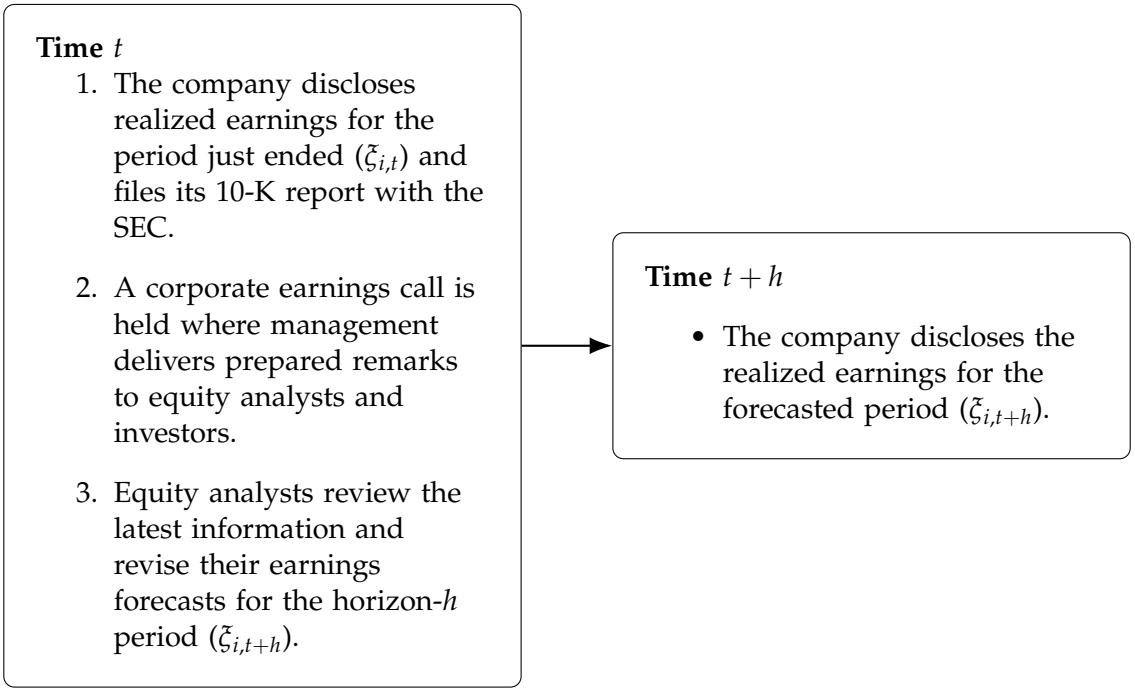


FIGURE 4: **Textual vs. Fundamental Predictive Power Across Horizons**

This chart reports the out-of-sample explanatory power, expressed in R-squared, of two models across different forecast horizons (one-, two-, and three-year-ahead) for three outcomes: (i) *Analysts' Expected Change in Earnings* (analysts' consensus after the earnings call minus the value just realized, scaled by price), (ii) *Forecast Disagreement* (the standard deviation of analyst-level forecast changes, scaled by price), and (iii) *Realized Change in Earnings* (the value to be realized in one, two, or three years minus the value just realized, scaled by price). The first model (Text) relies solely on textual features extracted from earnings calls, while the second (Fundamentals) uses standard accounting and financial variables.

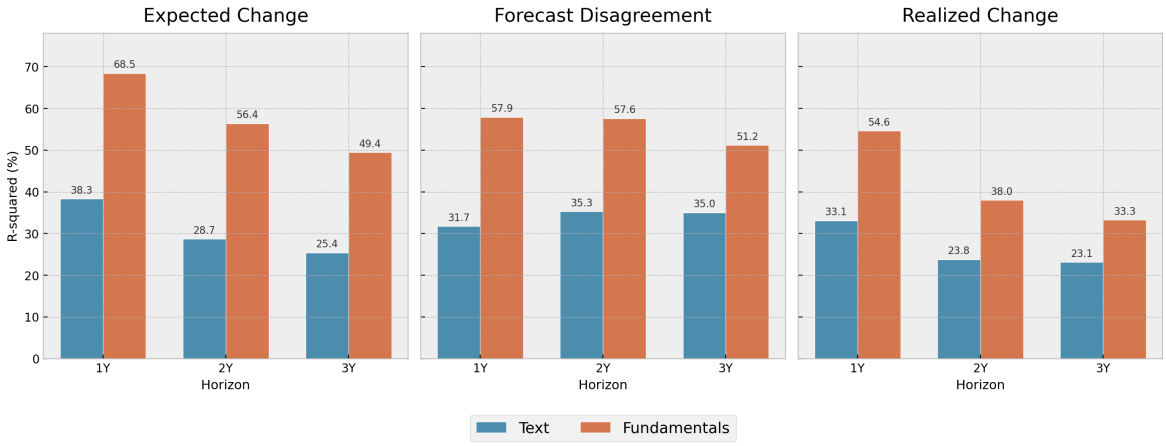


FIGURE 5: **Predicted Effects of Fundamental Information**

This figure shows partial dependence plots (PDPs) of predicted one-, two-, and three-year-ahead outcomes with respect to Fundamental Information, proxied by Standardized Unexpected Earnings (SUE). PDPs show how the model's predictions vary with a given feature while averaging over the empirical distribution of all other features. Panels show, respectively: (top left) *Analyst Expected Earnings Change* (analyst consensus after the earnings call minus the value just realized, scaled by price), (right) *Forecast Disagreement* (the standard deviation of analyst-level forecast changes, scaled by price), and (bottom left) *Realized Change in Earnings* (the value to be realized in one, two, or three years minus the value just realized, scaled by price). The line traces the model's average prediction as SUE varies (holding other features at their empirical distribution).

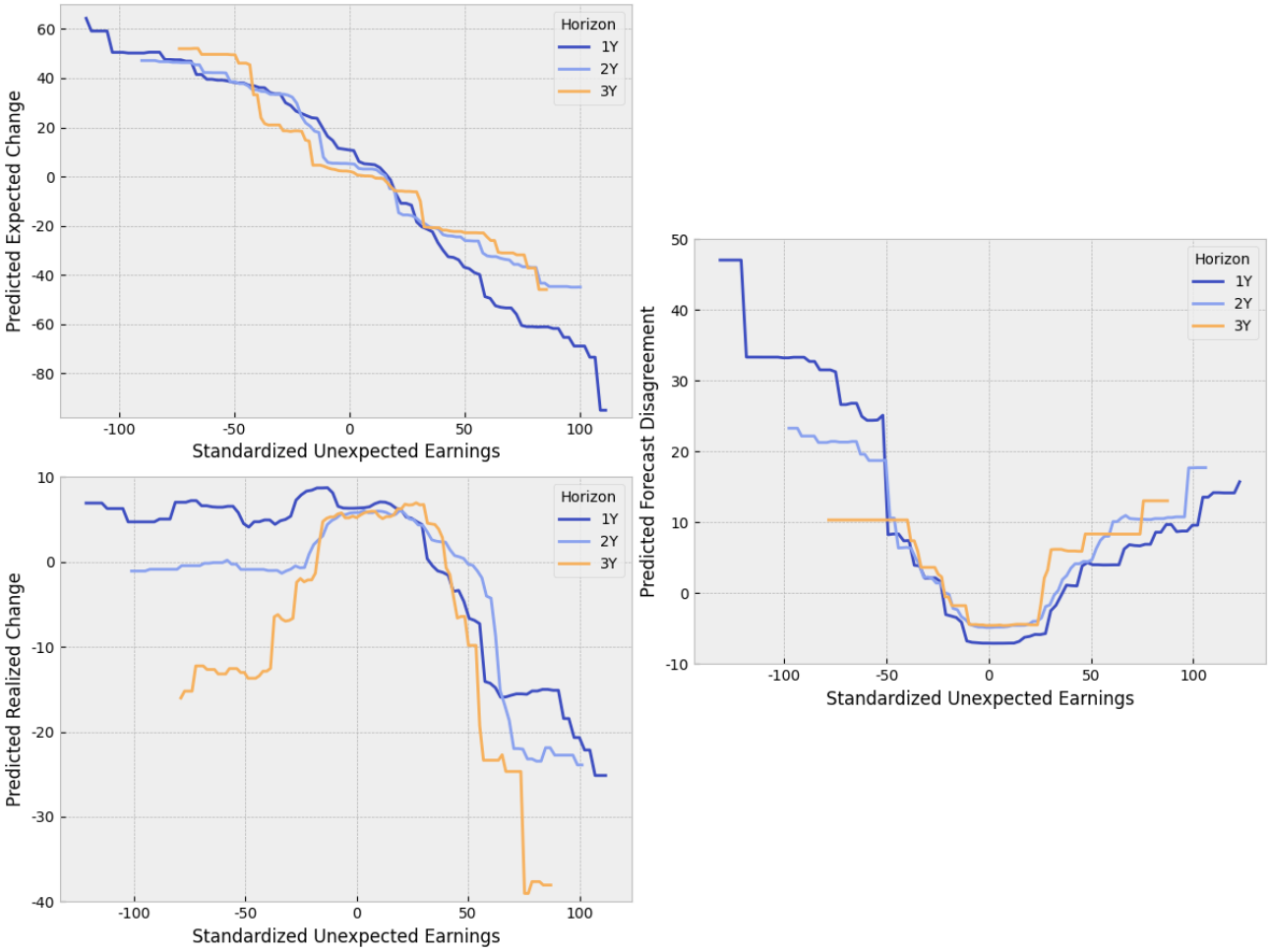


FIGURE 6: **Geometry of Original and Morphed Earnings Calls**

This figure shows UMAP-computed (McInnes et al., 2018) two-dimensional projections of text embeddings for original and morphed earnings call transcripts. Each panel is a 30×30 bins two-dimensional histogram; darker cells indicate higher local transcript density. The UMAP algorithm approximately preserves neighborhood structure, so transcripts with similar topics lie near one another. The last row shows the original (un-morphed) embeddings for comparison, and the 2D coordinates are to be interpreted in the comparison across morphing directions, so corresponding regions tend to represent the same topics across embeddings.



FIGURE 7: **Linguistic Dimensions in Analysts Beliefs and Realized Earnings**

This chart reports the average predicted treatment effects (PTEs) of the linguistic dimensions (described in Section 5) on *Analyst Expected Earnings Change* (analyst consensus after the earnings call minus the value just realized, scaled by price) and *Realized Change in Earnings* (the value to be realized in one, two, or three years minus the value just realized, scaled by price). PTEs are computed as the difference between the output of the trained model, $\widehat{g}_{o,h}^{(ST)}(x^S, e^T)$, when earnings call transcripts are morphed to emphasize dimension d and its output when evaluated on the original transcript embeddings. *Fundamental News* corresponds to the effect associated with an interquartile change in the empirical distribution of Standardized Unexpected Earnings (SUE), as predicted by the same model. The *Difference* bar reports the gap between the *Analyst Expected Earnings Change* and *Realized Change in Earnings*. All reported PTEs are statistically significant at the 1% level. Inference is based on transcript-level variation in predicted treatment effects: for each linguistic dimension, I test whether the mean transcript-level PTE differs from zero using a two-sided t-test. Observations are pooled across all forecast horizons.

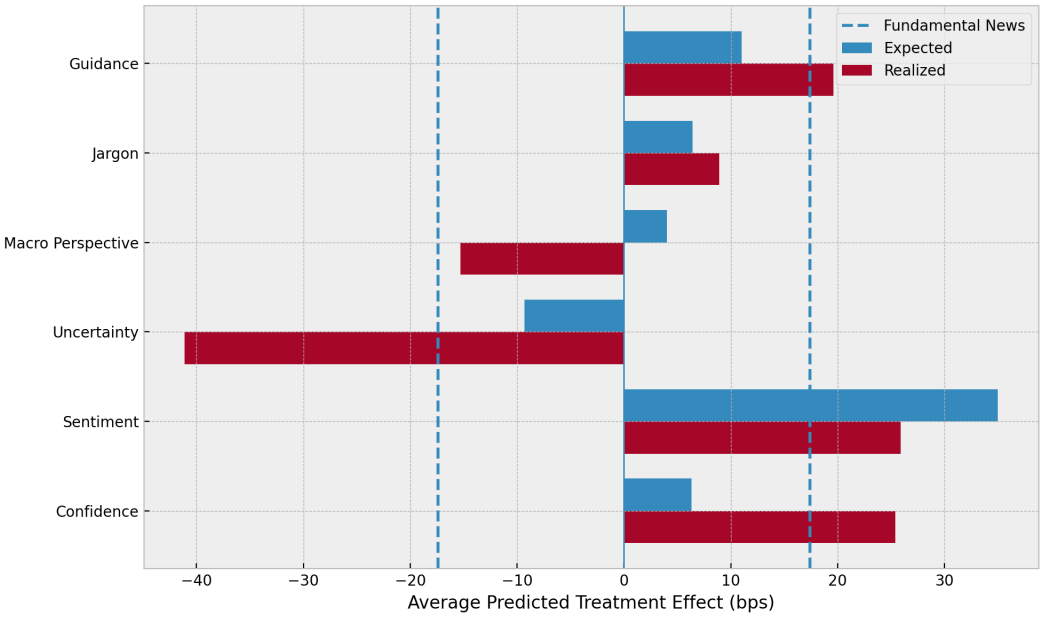
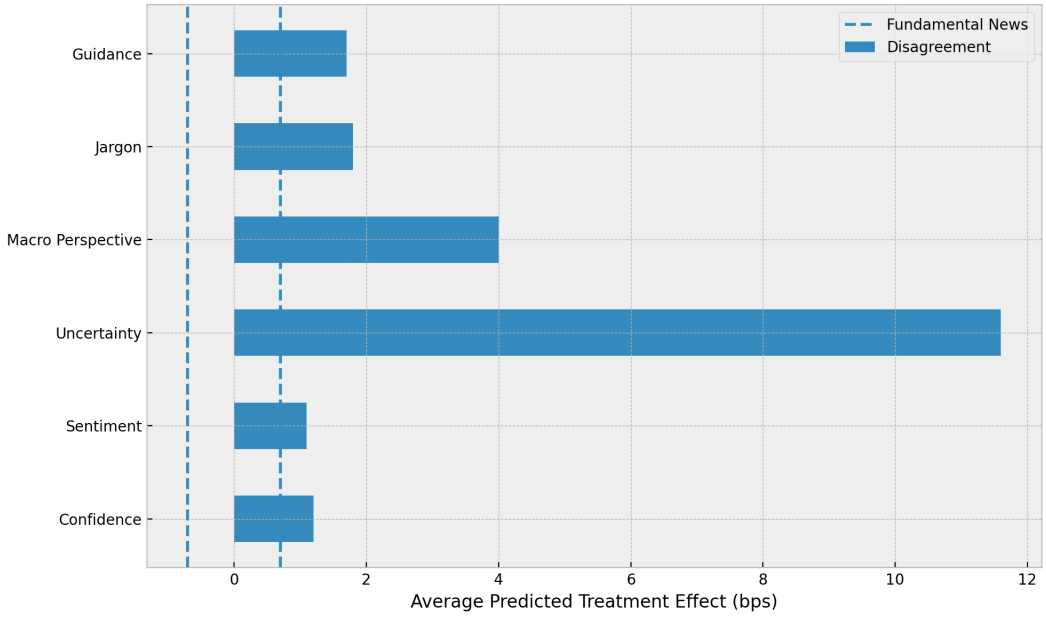


FIGURE 8: **Linguistic Dimensions and Forecast Disagreement**

This chart reports the average predicted treatment effects (PTEs) of the linguistic dimensions (described in Section 5) on *Forecast Disagreement* (the standard deviation of analyst-level forecast changes, scaled by price). PTEs are computed as the difference between the output of the trained model, $\widehat{g}^{(ST)}_{DISP,h}(x^S, e^T)$, when earnings call transcripts are morphed to emphasize dimension d and its output when evaluated on the original transcript embeddings. *Fundamental News* corresponds to the effect associated with an interquartile change in the empirical distribution of Standardized Unexpected Earnings (SUE), as predicted by the same model. All reported PTEs are statistically significant at the 1% level. Inference is based on transcript-level variation in predicted treatment effects: for each linguistic dimension, I test whether the mean transcript-level PTE differs from zero using a two-sided t-test. Observations are pooled across all forecast horizons.



8 Tables

TABLE 1: **Distribution of Sample Observations Across Forecast Horizons**

This table reports, for the one-, two-, and three-year ahead forecast horizons, the number of observations, their percentage share of the total sample, the number of companies, the number of brokers, and the number of analysts employed by those brokers. The sample covers the period 2008–2023.

Horizon	# Obs	% Obs	# Companies	# Brokers	# Analysts
1Y	34,798	45	5,338	611	8,242
2Y	27,808	36	4,654	563	7,388
3Y	15,234	19	3,240	367	3,662

TABLE 2: **Summary Statistics for Target Variables**

This table reports summary statistics for the three target variables used in the forecasting exercises. The variables are: (i) *Analyst Expected Earnings Change* (analyst consensus after the earnings call minus the value just realized, scaled by price), (ii) *Forecast Disagreement* (the standard deviation of analyst-level forecast changes, scaled by price), and (iii) *Realized Change in Earnings* (the value to be realized in one, two, or three years minus the value just realized, scaled by price). All variables are trimmed at the 5% and 95% levels. For clarity and comparability, all quantities are reported in basis points.

(A) Expected Change in Earnings

Horizon	Count	Mean	Std	Min	p25	p50	p75	Max
1Y	31318	344.95	376.83	-1227.68	221.45	410.90	568.75	1029.61
2Y	25028	517.20	303.90	-565.25	361.98	528.82	696.09	1270.89
3Y	13714	619.48	304.38	-227.74	433.20	607.21	798.32	1465.52

(B) Forecast Disagreement

Horizon	Count	Mean	Std	Min	p25	p50	p75	Max
1Y	29769	81.89	101.60	4.38	17.45	40.69	101.47	547.08
2Y	23491	103.83	114.33	9.86	29.56	58.81	130.34	615.78
3Y	9648	128.64	146.76	8.85	34.43	70.45	159.90	775.18

(C) Realized Change in Earnings

Horizon	Count	Mean	Std	Min	p25	p50	p75	Max
1Y	31316	264.91	500.37	-1716.17	124.27	376.07	556.79	1152.66
2Y	25017	363.07	476.75	-1228.29	148.23	422.05	646.46	1470.19
3Y	13707	439.19	479.46	-1005.35	177.69	469.70	724.51	1662.47

TABLE 3: **Improvement in Forecast Accuracy from Textual Features**

This table reports the results of the Clark and West (2007) nested forecast comparison test assessing whether textual measures provide statistically significant out-of-sample gains when predicting both fundamentals and analyst behavior at one-, two-, and three-year forecast horizons. The three target variables are: (i) *Analyst Expected Earnings Change* (analyst consensus after the earnings call minus the value just realized, scaled by price), (ii) *Forecast Disagreement* (the standard deviation of analyst-level forecast changes, scaled by price), and (iii) *Realized Change in Earnings* (the value to be realized in one, two, or three years minus the value just realized, scaled by price). The comparison involves two specifications: a benchmark model including stock characteristics and financial-statement variables, and an expanded model that additionally incorporates textual features. The one-sided alternative hypothesis tests whether the text-augmented model delivers superior forecasting performance relative to the benchmark. **Mean Squared Error (MSE) Reduction due to Text** reports the percentage decrease in test-set MSE attributable to textual information (adjusted for noise induced by model expansion), and **C&W t-stat** provides the corresponding test statistic.

(i) Expected Change in Earnings

Horizon	MSE Reduction due to Text (%)	C&W t-stat
1Y	9.70	11.67
2Y	12.13	11.06
3Y	8.10	7.32

(ii) Forecast Disagreement

Horizon	MSE Reduction due to Text (%)	C&W t-stat
1Y	9.11	10.50
2Y	10.82	9.89
3Y	12.55	7.64

(iii) Realized Change in Earnings

Horizon	MSE Reduction due to Text (%)	C&W t-stat
1Y	6.97	10.30
2Y	9.21	11.52
3Y	8.93	8.51

TABLE 4: **Original and Morphed Earnings Call Excerpts by Linguistic Dimension**

This table provides illustrative pairs of excerpts corresponding to the morphing exercise detailed in Section 5. The first column contains the passages from the original earnings call, while the second column reports the morphed versions produced by the LLM when prompted to intensify a specific linguistic dimension.

Dimension	Original Excerpt	Morphed Excerpt
Forward Guidance	Looking ahead, we expect steady demand in our core markets and continued margin discipline. While some uncertainty remains, we're confident in our long-term strategy.	Looking ahead, we anticipate revenue growth next quarter, driven by accelerating demand in our core markets. Operating margins should expand as we execute our cost initiatives.
Jargon	We have streamlined operations to better align production with customer demand.	We've optimized our end-to-end operational footprint through agile capacity realignment to ensure cross-functional synergies across key demand verticals.
Macro Focus	Our Q2 results were primarily shaped by the success of our new retail partnerships.	Our Q2 results reflected broader macro dynamics–particularly the uptick in consumer confidence and stabilizing broader supply chains–which supported retail expansion across regions.
Uncertainty	We are encouraged by our progress and expect continued momentum next quarter.	While we are encouraged by our progress, we remain cautious about potential headwinds–including raw material inflation, geopolitical tensions, and slower–than-expected demand recovery in Europe.
Sentiment	This quarter's performance was in line with our expectations.	This quarter's performance was excellent, surpassing expectations and reaffirming the strength of our business model.
Confidence	We think our new product line will perform well, though it is still early in the rollout.	Our new product line is already exceeding expectations, and we are absolutely confident it will continue to drive strong performance as adoption scales.

TABLE 5: **Text vs. Fundamentals Importance Across Linguistic Morphs**

This table reports how different linguistic morphs affect the relative contribution of textual features versus firm fundamentals in the machine learning model predicting *Analyst Expected Earnings Change* (analyst consensus after the earnings call minus the value just realized, scaled by price). Feature importance is measured using SHAP values: for each observation, absolute SHAP values are summed within the text and fundamentals feature groups and normalized so that their shares sum to 100%. By construction, the Text Share and Fundamentals Share add to 100% in each column. The table compares the original (unmorphed) text with its morphed counterpart for each linguistic dimension. The final column reports the change in the Text Share (in percentage points) induced by the morph. *t*-statistics testing whether this change differs from zero are reported in parentheses, with standard errors clustered at the firm level.

	Text Share (%)		**Fundamentals Share (%)**		
Linguistic Morph	**Original**	**Morphed**	**Original**	**Morphed**	**Δ Text Share (pp)**
Forward Guidance	40.7	49.2	59.3	50.8	8.5***
					(13.2)
Jargon	40.7	43.3	59.3	56.7	2.6***
					(5.8)
Macro Focus	40.7	46.5	59.3	53.5	5.8***
					(11.2)
Uncertainty	40.7	59.6	59.3	40.4	18.9***
					(22.4)
Sentiment	40.7	46.1	59.3	53.9	5.4***
					(10.6)
Confidence	40.7	51.8	59.3	48.2	11.1***
					(16.8)

TABLE 6: **Changes in Attention to Earnings Surprises by Linguistic Dimension**

This table examines how linguistic morphs alter the model's attention to earnings surprises, measured by Standardized Unexpected Earnings (SUE), that captures the surprise component of the most recently announced earnings relative to analyst consensus. The Original SUE Share reports SUE's baseline relative importance, computed as SUE's share of the total absolute SHAP value among all fundamental features under unmorphed transcripts. The Morphed SUE Share reports the corresponding share after applying each linguistic morph. Δ Relative SHAP (pp) measures the change in SUE's relative importance, expressed in percentage points. Δ Absolute SHAP reports the change in the raw (unnormalized) absolute SHAP value associated with SUE, capturing shifts in the overall weight placed on earnings surprises. *t*-statistics test whether these changes differ from zero and are reported in parentheses, with standard errors clustered at the firm level. The baseline row reports values for the original (unmorphed) transcripts. Statistical significance at the 1%, 5%, and 10% levels is denoted by ***, **, and *, respectively.

Dimension	**Original SUE Share (%)**	**Morphed SUE Share (%)**	**Δ Relative SHAP (pp)**	**Δ Absolute SHAP**	**t-stat**
Baseline					
Original	4.8	–	–	0.029	–
Linguistic Morphs					
Forward Guidance	4.8	3.1	-1.7***	-0.013	(-9.1)
Jargon	4.8	4.2	-0.6**	-0.004	(-2.9)
Macro Focus	4.8	6.4	1.6***	0.005	(7.8)
Uncertainty	4.8	3.2	-1.6***	-0.012	(-8.4)
Sentiment	4.8	7.2	2.4***	0.008	(11.3)
Confidence	4.8	5.9	1.1***	0.006	(5.2)

TABLE 7: **Determinants of Linguistic-Dimension Treatment Effects on Analyst Expectations**

This table reports OLS regression results where the dependent variable is the Predicted Treatment Effect (PTE) on Analyst Expected Earnings Change for each linguistic dimension, measured in basis points. PTEs are computed as $\widehat{g}^{(ST)}_{EC,h}(x^S_{i,t}, \widetilde{e}^{T,d}_{i,t}) - \widehat{g}^{(ST)}_{EC,h}(x^S_{i,t}, e^T_{i,t})$, where $\widehat{g}^{(ST)}_{EC,h}(\cdot)$ is the gradient boosting model trained on hard-information controls and earnings call embeddings for Analyst Expected Earnings Change at horizon h. VIX_t is the level of the CBOE Volatility Index on the earnings call date. Book-to-Market is computed from the most recent financial statements. Analyst Coverage is the number of analysts issuing forecasts within 15 days of the earnings call. Earnings Volatility is the standard deviation of quarterly earnings over the prior 12 quarters. Firm Age is years since IPO. Standard errors clustered by firm are in parentheses. ***, **, * denote significance at the 1%, 5%, and 10% levels, respectively. Observations are pooled across all forecast horizons.

Dependent Variable: PTE on Analyst Expected Earnings Change (bps)						
	(1) Forward Guidance	(2) Jargon	(3) Macro Focus	(4) Uncertainty	(5) Sentiment	(6) Confidence
Panel A: Market-Level Variables						
$\mathbb{1}(\text{VIX}_t > 20)$	0.892^{*}	0.187	0.876^{***}	-4.218^{***}	1.847^{**}	0.623
	(0.478)	(0.298)	(0.254)	(0.987)	(0.743)	(0.512)
Panel B: Firm-Level Variables						
$\text{Book-to-Market}_{i,t}$	0.534^{**}	-0.423^{**}	0.087	-1.456^{***}	2.134^{***}	1.287^{***}
	(0.213)	(0.178)	(0.098)	(0.312)	(0.387)	(0.298)
$\text{Log(Market Cap)}_{i,t}$	-0.156	0.312^{***}	-0.148^{**}	0.623^{***}	-0.412^{**}	-0.287^{*}
	(0.112)	(0.089)	(0.053)	(0.145)	(0.167)	(0.156)
$\text{Analyst Coverage}_{i,t}$	-0.023	0.045^{***}	-0.019^{**}	0.087^{***}	-0.056^{**}	-0.034
	(0.018)	(0.014)	(0.008)	(0.023)	(0.027)	(0.024)
$\text{Earnings Volatility}_{i,t}$	0.456	-0.234	0.592^{***}	-3.245^{***}	1.567^{**}	0.823^{*}
	(0.398)	(0.287)	(0.182)	(0.687)	(0.612)	(0.487)
$\text{Firm Age}_{i,t}$	-0.008	0.018^{**}	-0.008^{*}	0.034^{**}	-0.023^{*}	-0.012
	(0.009)	(0.007)	(0.004)	(0.014)	(0.012)	(0.011)
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	28,456	28,456	28,456	28,456	28,456	28,456
R^2	0.068	0.104	0.143	0.172	0.131	0.089

References

Anilowski, C., Feng, M., and Skinner, D. J. (2007). Does earnings guidance affect market returns? the nature and information content of aggregate earnings guidance. *Journal of Accounting and Economics*, 44:36–63.

Araci, D. T. (2019). Finbert: Financial sentiment analysis with pre-trained language models. *Working Paper*.

Arif, S., Marshall, N. T., Schroeder, J. H., and Yohn, T. L. (2019). A growing disparity in earnings disclosure mechanisms: The rise of concurrently released earnings announcements and 10-ks. *Journal of Accounting and Economics*, 68:101221.

Bastianello, F. (2022). Time-series and cross-section of risk premia expectations: Evidence from financial analysts. *SSRN Electronic Journal*.

Bastianello, F., Décaire, P. H., and Guenzel, M. (2024). Mental models and financial forecasts. *SSRN Electronic Journal*.

Ben-David, I. and Chinco, A. (2024). Expected eps × trailing p/e. *NBER Working Paper*.

Bhagwat, V., Cookson, J. A., Dim, C., and Niessner, M. (2025). The market's mirror: Revealing investor disagreement with llms. *SSRN*.

Bianchi, F., Ludvigson, S. C., and Ma, S. (2024). What hundreds of economic news events say about belief overreaction in the stock market. *NBER Working Paper*.

Binsbergen, J. H. V., Han, X., and Lopez-Lira, A. (2023). Man versus machine learning: The term structure of earnings expectations and conditional biases. *The Review of Financial Studies*, 36:2361–2396.

Bordalo, P., Gennaioli, N., Porta, R. L., and Shleifer, A. (2019). Diagnostic expectations and stock returns. *The Journal of Finance*, 74:2839–2874.

Bordalo, P., Gennaioli, N., Porta, R. L., and Shleifer, A. (2024). Belief overreaction and stock market puzzles. *Journal of Political Economy*, 132:1450–1484.

Bordalo, P., Gennaioli, N., Porta, R. L., and Shleifer, A. (2025). Finance without exotic risk. *Journal of Financial Economics*, 173:104145.

Bouchaud, J. P., Krüger, P., Landier, A., and Thesmar, D. (2019). Sticky expectations and the profitability anomaly. *The Journal of Finance*, 74:639–674.

Bozanic, Z., Roulstone, D. T., and Buskirk, A. V. (2018). Management earnings forecasts and other forward-looking statements. *Journal of Accounting and Economics*, 65:1–20.

Bybee, J. L. (2025). The ghost in the machine: Generating beliefs with large language models. *Working Paper*.

Call, A. C., Hribar, P., Skinner, D. J., and Volant, D. (2024). Corporate managers' perspectives on forward-looking guidance: Survey evidence. *Journal of Accounting and Economics*, 78:101731.

Charles, C. and Kendall, C. (2025). Causal narratives. *SSRN Electronic Journal*.

Charles, C. and Sui, P. (2025). Marketwide memory. *SSRN Electronic Journal*.

Chen, Z. and Huang, J. J. (2024). Memory and beliefs in financial markets: A machine learning approach. *Working Paper*.

Clark, T. E. and West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138:291–311.

Cohen, L., Malloy, C., and Nguyen, Q. (2020). Lazy prices. *The Journal of Finance*, 75:1371–1415.

Cohen, L. and Nguyen, Q. (2024). Moving targets. *SSRN Electronic Journal*.

Dahlquist, M. and Ibert, M. (2024). Equity return expectations and portfolios: Evidence from large asset managers. *The Review of Financial Studies*, 37:1887–1928.

De La O, R., Han, X., and Myers, S. (2025). The return of return dominance: Decomposing the cross-section of prices. *Journal of Financial Economics*, 169:104059.

De La O, R. and Myers, S. (2021). Subjective cash flow and discount rate expectations. *The Journal of Finance*, 76:1339–1387.

De La O, R. and Myers, S. (2024). Which subjective expectations explain asset prices? *The Review of Financial Studies*, 37:1929–1978.

De la Parra, D. and Gallemore, J. (2024). The sound of uncertainty: Examining managerial acoustic uncertainty in conference calls. *SSRN Electronic Journal*.

De Rosa, L. (2024). (Deep) Learning Analyst Memory. *SSRN Electronic Journal*.

Dessaint, O., Foucault, T., and Fresard, L. (2024). Does alternative data improve financial forecasting? the horizon effect. *The Journal of Finance*, 79:2237–2287.

Diether, K. B., Malloy, C. J., and Scherbina, A. (2002). Differences of opinion and the cross section of stock returns. *The Journal of Finance*, 57:2113–2141.

Dzieliński, M., Wagner, A., and Zeckhauser, R. J. (2017). Straight talkers and vague talkers: The effects of managerial style in earnings conference calls. *NBER Working Paper*.

Décaire, P. H. and Guenzel, M. (2023). What drives very long-run cash flow growth expectations? *SSRN Electronic Journal*.

Décaire, P. H., Sosyura, D., and Wittry, M. D. (2024). Resolving estimation ambiguity. *SSRN Electronic Journal*.

Gabaix, X., Koijen, R. S. J., Richmond, R., and Yogo, M. (2025). Asset embeddings. *SSRN Electronic Journal*.

Giglio, S., Kelly, B., and Xiu, D. (2022). Factor models, machine learning, and asset pricing. *Annual Review of Financial Economics*, 14:337–368.

Giglio, S., Maggiori, M., Rillo, J., Stroebel, J., Utkus, S., and Xu, X. (2025). Investor beliefs and expectation formation. *SSRN Electronic Journal*.

Giglio, S., Maggiori, M., Stroebel, J., and Utkus, S. (2021). Five facts about beliefs and portfolios. *American Economic Review*, 111:1481–1522.

Glasserman, P. and Lin, C. (2023). Assessing look-ahead bias in stock return predictions generated by gpt sentiment analysis. *Journal of Financial Data Science*, 6:25–42.

Gormsen, N. J. and Huber, K. (2025). Corporate discount rates. *American Economic Review*, 115:2001–49.

Greenwood, R. and Shleifer, A. (2014). Expectations of returns and expected returns. *The Review of Financial Studies*, 27:714–746.

Hailiang Chen, B.-H. H. and Peng, Z. (2025). Why do investors like short-leg securities? evidence from a textual analysis of buy recommendations. *The Review of Financial Studies*.

Hansen, A. L., Horton, J. J., Kazinnik, S., Puzzello, D., and Zarifhonarvar, A. (2024). Simulating the survey of professional forecasters. *SSRN*.

Hassan, T. A., Hollander, S., Lent, L. V., and Tahoun, A. (2019). Firm-level political risk: Measurement and effects. *The Quarterly Journal of Economics*, 134:2135–2202.

Hirshleifer, D. and Teoh, S. H. (2003). Limited attention, information disclosure, and financial reporting. *Journal of Accounting and Economics*, 36:337–386.

Hoberg, G. and Maksimovic, V. (2015). Redefining financial constraints: A text-based analysis. *The Review of Financial Studies*, 28:1312–1352.

Hoberg, G. and Manela, A. (2025). The natural language of finance. *Foundations and Trends in Finance*, 14:244–365.

Hoberg, G. and Phillips, G. (2010). Product market synergies and competition in mergers and acquisitions: A text-based analysis. *The Review of Financial Studies*, 23:3773–3811.

Hoberg, G. and Phillips, G. (2016). Text-based network industries and endogenous product differentiation. *Journal of Political Economy*, 124:1423–1465.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? *WP*.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. (2024). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43.

Huang, X., Teoh, S. H., and Zhang, Y. (2014). Tone management. *The Accounting Review*, 89:1083–1113.

Jensen, T. I., Kelly, B., and Pedersen, L. H. (2023). Is there a replication crisis in finance? *Journal of Finance*, 78:2465–2518.

Kahneman, D. and Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, 39:341–350.

Kazinnik, S. and Sinclair, T. M. (2025). Fomc in silico: A multi-agent system for monetary policy decision modeling. *SSRN Electronic Journal*.

Ke, S. (2024). Analysts' belief formation in their own words. *SSRN Electronic Journal*.

Kothari, S. P., So, E., and Verdi, R. (2016). Analysts' forecasts and asset pricing: A survey. *Annual Review of Financial Economics*, 8:197–219.

Kravet, T. and Muslu, V. (2013). Textual risk disclosures and investors' risk perceptions. *Review of Accounting Studies*, 18:1088–1122.

Law, K. K. F. (2023). Good-bye `i/b/e/s` (or not?). *Journal of Financial Reporting*, 8:41–61.

Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., Bhattacharjee, A., Jiang, Y., Chen, C., Wu, T., Shu, K., Cheng, L., and Liu, H. (2024). From generation to judgment: Opportunities and challenges of llm-as-a-judge. *Arxiv*.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, 45:221–247.

Li, K., Mai, F., Shen, R., and Yan, X. (2021). Measuring corporate culture using machine learning. *The Review of Financial Studies*, 34:3265–3315.

Livnat, J. and Mendenhall, R. R. (2006). Comparing the post–earnings announcement drift for surprises calculated from analyst and time series forecasts. *Journal of Accounting Research*, 44:177–205.

Lopez-Lira, A. and Tang, Y. (2023). Can chatgpt forecast stock price movements? return predictability and large language models. *SSRN Electronic Journal*.

Loughran, T. and Mcdonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66:35–65.

Ludwig, J., Mullainathan, S., and Rambachan, A. (2025). Large language models: An applied econometric framework. *NBER Working Paper*.

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S. I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence 2020 2:1*, 2:56–67.

Lundberg, S. M. and Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 2017-December:4766–4775.

Lv, L. (2024). The value of information from sell-side analysts. *Arxiv*.

Lyle, M. R., Riedl, E. J., and Siano, F. (2023). Changes in risk factor disclosures and the variance risk premium. *The Accounting Review*, 98:327–352.

Ma, Y., Ropele, T., Sraer, D. A., and Thesmar, D. (2024). A quantitative analysis of distortions in managerial forecasts. *SSRN Electronic Journal*.

Mamaysky, H., Shen, Y., and Wu, H. (2023). Credit information in earnings calls. *Arxiv*.

Manning, B. S., Zhu, K., and Horton, J. J. (2024). Automated social science: Language models as scientist and subjects. *SSRN Electronic Journal*.

Mayew, W. J. and Venkatachalam, M. (2012). The power of voice: Managerial affective states and future firm performance. *Journal of Finance*, 67:1–43.

McInnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *Working Paper*.

Meursault, V., Liang, P. J., Routledge, B. R., and Scanlon, M. M. (2023). Pead.txt: Post-earnings-announcement drift using text. *Journal of Financial and Quantitative Analysis*, 58:2299–2326.

Miller, B. P. (2010). The effects of reporting complexity on small and large investor trading. *The Accounting Review*, 85:2107–2143.

Nagel, S. and Xu, Z. (2022). Asset pricing with fading memory. *The Review of Financial Studies*, 35:2190–2245.

Peng, L. and Xiong, W. (2006). Investor attention, overconfidence and category learning. *Journal of Financial Economics*, 80:563–602.

Price, S. M. K., Doran, J. S., Peterson, D. R., and Bliss, B. A. (2012). Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance*, 36:992–1011.

Sarkar, S. K. (2025). Economic representations. *Working Paper*.

Sarkar, S. K. and Vafa, K. (2024). Lookahead bias in pretrained language models. *SSRN Electronic Journal*.

Sautner, Z., Lent, L. V., Vilkov, G., and Zhang, R. (2023). Firm-level climate change exposure. *The Journal of Finance*, 78:1449–1498.

Siano, F. (2025). The news in earnings announcement disclosures: Capturing word context using llm methods. *Management Science*.

Song, W. and Stern, S. (2025). Firm inattention and the efficacy of monetary policy: A text-based approach. *The Review of Economic Studies*, 92:3438–3469.

Stolborg, C. and Greenwood, R. M. (2025). Bubble beliefs. *SSRN Electronic Journal*.

Appendix

A Identification of the latest version of earnings call transcripts

I obtain the full text and associated metadata for earnings call events from the Capital IQ Transcripts database. Capital IQ Transcripts contains the full text of approximately 8,000 public-company events, including earnings calls, M&A calls, and shareholder meetings. From this database, I extract all available earnings-call transcripts.

Each transcript undergoes four stages of editorial review: *Preliminary* (real-time cleanup of voice-to-text with basic spellcheck), *Edited* (accuracy pass with added metadata), *Proofing* (external review for errors/classifications), and *Audited* (final quality assurance).

Different versions are indicated by `transcriptCollectionTypeId`. Each version has its own `transcriptId`, and metadata fields—such as the event date or title—may exhibit minor revisions across versions even when they refer to the same underlying event. When multiple versions are available, I consider the most recent timestamp as typically the most accurate.

Because each event is indexed by `keydevid` and may pertain to multiple companies, I construct the dataset at the (`keydevid`, company) level. For each event–company pair, I retain the most up-to-date transcript—operationalized as the highest-quality version in the editorial sequence and, within versions, the most recent timestamp.

As an additional safeguard, for any firm–event-date pair with multiple transcripts, I retain only the most up-to-date version (i.e., the latest, highest-quality transcript).

Finally, I link firms across data sources by using the Capital IQ mapping table to connect the Capital IQ transcript identifier `companyId` to Compustat's `gvkey`.

B Methodology for Language Generation

Language generation is a high-dimensional statistical process that depends on the careful tuning of hyper-parameters governing the model's stochastic output. Unlike conventional econometric models, there are no universally accepted guidelines for tuning these parameters, as their effectiveness depends on complex interactions and the specific textual context. I adopt standard settings widely used in practice. The temperature parameter controls the entropy of the sampling distribution; lower values yield more deterministic output. I set `temperature` to 0.7, a commonly used value in production settings such as ChatGPT. The `top p` parameter, set to 0.95, constrains sampling to the smallest set of candidate tokens whose cumulative probability exceeds 95%, thereby excluding implausible continuations. These parameters jointly influence the diversity, coherence, and rhetorical variation of the generated text, and thus have meaningful implications for the empirical content and realism of simulated earnings calls.

C Prompt Descriptions for Language Generation

This appendix documents the set of textual prompts used to systematically modify the tone and content of earnings call transcripts in the experimental framework. Each prompt targets a distinct rhetorical or informational dimension commonly observed in corporate disclosures. The prompts were operationalized using a large language model to simulate variation in managerial communication styles.

Forward Guidance: Strengthens forward-looking guidance by prompting the use of clear, strategic, and directional language. The generated text highlights managerial expectations, priorities, and actions while avoiding ambiguity or generic phrasing.

Prompt:
`Clarify and strengthen the company's forward-looking guidance. Use assertive, strategic language to highlight expectations, priorities, and upcoming actions. Where appropriate, specify directionality, expected outcomes, and the rationale behind future decisions. Avoid vague or generic phrasing.`

Jargon: Increases the use of technical, domain-specific terminology common in institutional financial communication. Examples include references to "margin expansion," "revenue mix," and "TAM." The prompt maintains semantic fidelity while enriching the financial register.

Prompt:
`Use more technical, domain-specific language typical of institutional earnings calls and financial reporting. Incorporate sector-relevant terminology (e.g., margin expansion, revenue mix, sequential growth, TAM, end-market demand) without changing the underlying meaning or facts.`

Macro Focus: Modifies the text to emphasize macroeconomic conditions, market-wide dynamics, and high-level industry developments. The prompt links firm-specific performance to broader economic indicators such as inflation, GDP, and regulatory trends.

Prompt:
`Revise the discussion to emphasize broader macroeconomic trends, market-wide`

dynamics, and high-level industry shifts. Draw links between the company's results and wider economic forces such as inflation, GDP, regulatory environments, or sector-wide developments.

Uncertainty: Prompts the model to introduce or amplify references to business risks and external uncertainties. These may include regulatory pressures, market volatility, or operational dependencies. The tone remains professional but more cautious.

Prompt:
Highlight and emphasize relevant risks, uncertainties, and operational or market challenges. Introduce cautious language where appropriate, noting external dependencies, regulatory pressures, or volatility in end markets. Maintain a professional tone while acknowledging areas of concern.

Sentiment: Enhances the optimistic tone of the text. The prompt increases the emphasis on business momentum, growth opportunities, and positive outlook, often by highlighting recent wins and expressing confidence in future performance.

Prompt:
Make the emotional tone more positive and optimistic. Emphasize momentum, exciting opportunities, and enthusiasm about the company's future. Highlight wins and express confidence in continued success.

Confidence: Instructs the model to adopt the tone of confident and assertive executives. The prompt removes hedging language and emphasizes decisiveness, authority, and control over business operations and strategy.

Prompt:
Rewrite the text in the tone of confident and assertive executives during an earnings call. Use decisive language, remove hedging expressions, and clearly convey control over business strategy and performance.

These prompts enable structured variation in linguistic tone and content, allowing for controlled analysis of linguistic effects on perception and interpretation of earnings disclosures.

D LLM as a Judge

While large language models enable scalable counterfactual text generation, they remain susceptible to hallucinations that can undermine empirical validity. For example, a synthetically altered earnings call transcript might inadvertently report incorrect earnings figures, thereby biasing the predictive accuracy metrics reported in this paper. Even with carefully engineered prompts, instruction-tuned chat variants can introduce spurious references to unrelated firms, stray numerical data, or extraneous statements drawn from their training corpus. Without an ex post validation mechanism, these unintended deviations threaten to produce misleading inferences and spurious results.

This unintended behavior, commonly referred to as "hallucination" in the ML literature, has been documented extensively by computer scientists and practitioners (e.g., Huang et al., 2024). Relying on manual inspection to verify whether prompt-driven modifications have been executed correctly is both laborious and inherently subjective, exposing the analysis to evaluator bias. To overcome these limitations and ensure the internal validity of the counterfactual text experiments, I implement an automated "judge" using another instance of the model described above, Meta's LLaMA 3 70B-Instruct. LLM-as-a-judge functions as an automated validation layer: a pretrained LLM evaluates each morphed transcript against the original, assessing whether quantitative content is preserved and whether the intended linguistic shift is clear, and flagging any spurious insertions (e.g., hallucinated figures or false statements) (Li et al., 2024).

By comparing these scores across versions, I ask the LLM to classify each morphed transcript into one of three categories:

1. **Yes** — the morphing was executed correctly and the language modification is clear and evident;
2. **I am unable to determine** — the morphing quality is uncertain;
3. **No** — the morphing is inadequate.

I discard every variant that fails to meet the highest quality standard, thereby ensuring that the predictive analyses are based exclusively on precisely controlled linguistic change. The details on the prompt I use are in Section E.

This methodology helps create and validate credible textual counterfactuals by establishing an objective and reproducible mechanism for assessing the quality of generated texts. The LLM-as-a-judge framework acts as a consistency filter that ensures counterfactuals preserve the informational core of the original transcript while exhibiting a controlled and interpretable linguistic modification. By systematically scoring and classifying each counterfactual according to predefined semantic and stylistic dimensions, the judge model operationalizes the notion of "credible" counterfactuals—texts that differ meaningfully in the intended attribute (e.g., tone or sentiment) yet remain faithful to the underlying economic message.

In doing so, the approach mitigates two critical threats to internal validity. First, it minimizes the risk of *semantic drift*, that is, the introduction of unintended changes to factual or contextual content that could distort the interpretation of downstream results. Second, it limits *evaluation bias* by replacing subjective human judgment with an automated and transparent scoring procedure. Because the same pretrained model family is used for both generation and evaluation, the validation process benefits from consistent representational semantics across tasks, enhancing the reliability of the comparisons.

E Prompt Descriptions for LLM-as-a-Judge

This appendix outlines the prompt template used to evaluate whether a morphed version of an earnings call transcript preserves factual content while exhibiting a significant change in tone.

Each prompt consists of the original (Text A) and the morphed (Text B) version. The large language model is instructed to assess two criteria: (i) that all numerical information and structural elements (e.g., order, scope) are retained, and (ii) that Text B exhibits a clear tonal shift (e.g., more confident or more cautious).

The standardized prompt format is as follows:

```
You are an expert in financial communication.
Below are two excerpts from earnings calls.

Text A: [original excerpt]
Text B: [morphed excerpt]
```

```
Does Text B preserve the same numbers and structure as Text A,
while clearly changing the tone?

Respond with one of the following:
1. Yes: the morphing was executed correctly
and the language modification is clear and evident,
2. Not sure
3. No: the morphing is inadequate
```

All evaluations are performed in a zero-shot setting. All responses labeled "No: the morphing is inadequate" are discarded from the analysis and reproduced in another instance.