

# Analysts' Belief Formation in Their Own Words

Shikun (Barry) Ke\*

*Job Market Paper*

[Link to the latest version](#)

First draft: November 16, 2024. This draft: November 23, 2025

## Abstract

I study how equity analysts form subjective beliefs about firms' earnings using their own written text from over 1.1 million equity research reports. Using large language models, I identify the topics discussed by analysts and represent topic-level information using textual embeddings. I introduce a novel text-instrumented Coibion-Gorodnichenko regression to uncover analysts' over- and underreaction to specific information. Using this new procedure, I find pervasive underreaction in short-term earnings forecasts across topics, whereas overreaction in long-term forecasts is concentrated in qualitative, intangible topics rather than quantitative, statistical ones. Revisions driven by qualitative information in long-term earnings forecasts strongly predict future stock returns. Finally, I use textual data to investigate the behavioral mechanisms underlying the documented misreactions. The empirical results suggest that overconfidence is an important driver of the overreaction to qualitative information, while herding appears to be important in explaining the overall underreaction observed in short-term forecasts.

*Keywords:* Analyst reports, earnings forecast, belief formation, overreaction, underreaction, large language model

---

\*Yale School of Management (barry.ke@yale.edu). I am indebted to my dissertation committee, Nick Barberis, Bryan Kelly, Alp Simsek, and Tobias Moskowitz, for their guidance and continued support. I also thank Stefano Giglio, Kelly Shue, Paul Goldsmith-Pinkham, Kaushik Vasudevan, Jeff Pontiff, Theis Jensen, Leland Bybee, Ben Matthies, Manish Jha, Seth Pruitt, Dan Bradley, Lawrence Jin, Tim Christensen, Ahmed Guecioueur, Lin Peng, Zhengyang Xu, Snehal Banerjee, Xindi He, Felix Wilke, Ernst Maug, as well as participants at the Spring 2025 NBER Behavioral Finance Working Group Meeting, the Oxford SMLFin Seminar, the 2025 Midwest Finance Association Conference, the 2025 Eastern Finance Association Conference, the MPWZ text-as-data workshop, the 2025 FIRS Conference, the Hong Kong Conference for Fintech, AI, and Big Data in Business, the 2025 Helsinki Finance Summit on Investor Behavior, the 2025 European Finance Association Conference, the 2025 CQA Annual Meeting, Texas A&M Mays Business School, and Yale for their comments. I thank Jared Flake and Kelvin Law for sharing the IBES brokerage matching data.

# 1 Introduction

There is a large body of research on subjective beliefs captured in survey data. Researchers examining these survey-based measures of beliefs have documented significant deviations from rationality, including overreaction and underreaction to information. These systematic biases in survey-based belief measures are also linked to important economic phenomena in macroeconomics and finance, such as business cycles and variation in asset prices.<sup>1</sup>

However, several open questions remain regarding the process of *belief formation*: What information do people pay attention to? What information do people over- or underreact to? And what are the behavioral mechanisms that explain over- or underreactions in beliefs? Answering these questions is essential for understanding people’s beliefs, yet traditional numerical or categorical survey responses offer limited guidance. Addressing these questions requires opening the black box of people’s own thinking.

In this paper, I study belief formation by examining how people explain their own beliefs. Specifically, I study the subjective beliefs of sell-side equity research analysts regarding stock earnings, using the written text from over 1.1 million equity research reports. These reports provide detailed analyses, forecasts, and recommendations on stocks to help investors make informed decisions. The quantitative measures they contain, such as earnings forecasts and price targets, are collected by aggregators like the Institutional Brokers’ Estimate System (IBES) and serve as standard proxies for subjective expectations in empirical research.<sup>2</sup> The written text in these reports, which provides the analysts’ rationale for their quantitative forecasts, thus offers valuable context for studying the formation of subjective beliefs.

I use large language models (LLMs) to extract and condense information from analyst reports, and I make three contributions to the understanding of subjective belief formation. First, my descriptive results show how the information on which analysts focus varies across time, stocks, and forecast horizons. Second, as my main contribution, I introduce a novel text-instrumented Coibion-Gorodnichenko regression to identify over- or underreaction to specific information. Leveraging this approach, I find that in their long-term earnings forecasts, analysts overreact to qualitative information but not to quantitative information.<sup>3</sup> In contrast, underreaction in short-term fore-

---

<sup>1</sup>Papers that study survey-based measures of subjective expectations include [Malmendier and Nagel \(2011\)](#), [Greenwood and Shleifer \(2014\)](#), [Coibion and Gorodnichenko \(2015\)](#), [Malmendier and Nagel \(2016\)](#), [Bordalo et al. \(2020\)](#), [Giglio et al. \(2021\)](#), [Lochstoer and Muir \(2022\)](#), [Bianchi et al. \(2022\)](#), [Nagel and Xu \(2022\)](#), [Nagel and Xu \(2023\)](#), [Bordalo et al. \(2024b\)](#), [Bordalo et al. \(2024c\)](#), [Bianchi et al. \(2024a\)](#), [Cui et al. \(2024\)](#), [Bhandari et al. \(2025\)](#), [Li et al. \(2025\)](#), and [Thesmar and Verner \(2025\)](#). See [Adam and Nagel \(2023\)](#) and [Giglio et al. \(2025\)](#) for recent reviews of the use of survey-based subjective expectation data in asset pricing.

<sup>2</sup>See [Chen et al. \(2013\)](#), [Bouchaud et al. \(2019\)](#), [Delao and Myers \(2021\)](#), [Cassella et al. \(2023\)](#), [Delao and Myers \(2024\)](#), [Bordalo et al. \(2024b\)](#), and [Bordalo et al. \(2024a,c\)](#) for use of IBES EPS forecasts as subjective cash-flow expectations, and [Bastianello \(2022\)](#), [Jensen \(2024\)](#), [Andrei et al. \(2023\)](#), and [Cao et al. \(2024\)](#) for use of IBES price targets as subjective return expectations.

<sup>3</sup>Intangible qualitative information pertains, for example, to a firm’s product development, corporate leadership,

casts is pervasive across all topics. Third, I explore the behavioral mechanisms for over- and underreaction manifested in analysts’ language. I find evidence consistent with overconfidence being an important driver of overreaction to qualitative information and herding behavior playing an important role in explaining short-term underreaction.

**Text-instrumented Coibion-Gorodnichenko regression.** Empirical research on over- and underreaction in survey data usually relies on the classic Coibion-Gorodnichenko regression (Coibion and Gorodnichenko (2015), henceforth CG regression) of forecast errors on forecast revisions. Using this approach, it has been documented that analysts’ short-term earnings forecasts exhibit underreaction to information (Bouchaud et al. (2019)), while their long-term forecasts exhibit overreaction (Bordalo et al. (2019, 2024c)). However, the classic CG regression does not attribute the source of over- and underreaction, because the forecaster’s information set is unobserved. The main objective of this paper is to use the text in analyst reports to study this key open question: to what kinds of information do analysts over- or underreact?

To address this question, I begin with a stylized model that relates belief formation and text, which provides a conceptual framework for identifying the sources of over- and underreaction. In the model, the analyst forecasts a firm’s future earnings based on a set of disaggregated noisy signals. I show that the original CG regression only identifies a weighted average of misreactions to individual signals and suffers from negative bias when measured beliefs contain non-informational driven variation.

I thus introduce a novel two-stage “text-instrumented” CG (TICG) regression, which identifies over- and underreaction to specific textual topics in the underlying self-disclosed text. This approach provides a much more granular view of where the over- and underreaction arises. In the first stage of the TICG regression, I regress forecast revisions on numerical representations of information in various topics. In the second stage, for each individual topic, I regress forecast errors on the instrumented forecast revisions while controlling for textual information in other topics. Intuitively, it captures excessive or insufficient forecast revisions *driven by information related to a specific topic*. I show that under certain assumptions, this procedure can identify over- or underreaction to specific topics, uncovering patterns that the original CG regressions based solely on numerical forecasts cannot capture.<sup>4</sup>

**Information extraction with LLMs.** To conduct the study, I construct a new dataset that links analysts’ earnings forecasts to their original research reports from January 1998 to September

---

or research and development (R&D); quantitative tangible information relates to profitability or financial conditions. See Section 5 below for details.

<sup>4</sup>The main identification assumption is that the text provides “sufficient statistics” regarding analysts’ belief revision (see Assumption 2), which assumes that the text related to the topic is sufficiently rich to accurately and completely describe the information that drives analyst’s truly held beliefs. I discuss potential violations of this condition in Section 2.5.

ber 2023. I then employ LLMs to extract and summarize the detailed content of the reports. LLMs better capture semantics and interpret natural language more human-like than traditional methods of textual analysis, such as dictionary-based searches or topic modeling. I categorize the information in analyst reports into 28 topics, and I capture topic-specific information using high-dimensional embedding vectors of the text associated with each topic. These embeddings represent textual information in numerical form and encode far richer semantic meaning than word counts or univariate sentiment measures. This enables researchers to account for sufficient variation in forecast revisions, which is crucial for valid inference in TICG regressions.<sup>5</sup>

Using the extracted topics, I begin my empirical analysis by analyzing variation in analysts’ focus on different types of information, measured by topic weights in analyst reports. I document substantial time-series variation: analysts’ focus on firm-specific operational topics—such as mergers and acquisitions (M&A), profitability, corporate leadership, and business strategy—is procyclical, whereas their focus on the macroeconomy, risk, and debt is strongly countercyclical. I also identify additional patterns of analysts’ information focus. For example, analysts discuss topics related to firms’ financial condition and risk more for value firms, while they tend to emphasize business strategy and M&A for growth firms. Analysts also focus on different topics when making short-term and long-term forecasts. Finally, the topic-weight distributions are more concentrated for analysts with less experience and broader coverage, suggesting a processing-capacity constraint when analysts face multiple topics simultaneously.

Furthermore, I provide evidence that these patterns are consistent with a top-down attention allocation model, in which agents focus more on topics that are most relevant to future outcomes. I find supporting evidence that analysts increase their discussion on the macroeconomy topic for a firm when that firm’s future earnings growth becomes more sensitive to the current macroeconomic conditions. I find similar patterns based on firm-level characteristics: those that predict future earnings growth more strongly receive higher weights in analyst reports. These results validate that the topics capture information closely related to analysts’ beliefs, and that analysts are actively choosing which topics are important to focus on.

**Dissecting over- and underreaction.** I then move on to my main analysis, where I use TICG regressions to study what kinds of information analysts over- or underreact to.

I begin by estimating the first-stage TICG regression, which decomposes forecast revisions into different topics using textual embeddings. This analysis addresses the question, “What information do analysts react to?” I find that textual embeddings have strong explanatory power for variation in forecast revisions, with an  $R^2$  of 0.136 for short-term (1–2 year) forecasts and 0.178 for long-term (3–5 year) forecasts, after controlling for analyst-, firm-, and time-fixed effects. Across topics,

---

<sup>5</sup>In contrast, univariate sentiment measures do not capture sufficient variation in forecast revisions, which leads to a weak first stage in TICG regression.



analysts react more strongly to profitability and other accounting-based information in their short-term forecasts, whereas they respond more to information about the macroeconomy, international markets, and R&D in their long-term forecasts. The strong predictability from textual embedding vectors also indicates that these embeddings can serve as strong instruments for forecast revisions.

I then proceed to the second stage of the TIGC regression to address the question, “What information do analysts over- or underreact to?” Figure 1 summarizes the key empirical findings from the TIGC regressions. Each dot in the plot represents a topic, and the y-axis corresponds to the topic-level TIGC regression estimate. Intuitively, it captures the predictability of forecast error from forecast revisions *driven by a specific topic*. For example, a TIGC coefficient of  $-0.3$  for the R&D topic means that when an analyst revises up their EPS forecast by \$1.0 per share *due to R&D information*, they over-adjust the EPS forecasts by \$0.3.<sup>6</sup> Therefore, the TIGC coefficient provides a direct measure of topic-level over- or underreaction.

In Figure 1, I document two new facts about analysts’ misreactions to information. First, for short-term earnings forecasts, analysts underreact to almost all topics, as reflected in positive TIGC coefficients. In contrast, long-term earnings forecasts exhibit much greater heterogeneity: analysts only overreact to a subset of topics (negative coefficients).

Second, for both short- and long-term forecasts, analysts tend to overreact more (or underreact less) to qualitative topics. I establish this by using an LLM to determine if a piece of information is qualitative based on whether it contains numerical content or statistics. I find that topics related to intangible aspects of a firm, such as business strategies, marketing, product development, and R&D, are highly qualitative, while topics tied to accounting-based information, such as profitability, cost, debt, and cash, are more quantitative. Figure 1 depicts a negative relationship between the topic-level TIGC regression coefficient (y-axis) and the share of qualitative information in each topic (x-axis). This pattern indicates that overreaction in long-term earnings forecasts is primarily driven by an overreaction to qualitative information, which is both statistically and economically significant.

Motivated by this observation, I construct two measures of long-term earnings forecast revisions: one in response to qualitative topics and the other in response to quantitative topics. I find that qualitative revisions strongly and negatively predict forecast errors, whereas quantitative revisions do not. My estimates indicate that approximately 25.7% of the variation in long-term earnings forecast revisions *due to qualitative information* reflects excessive adjustment (overshooting the targets), whereas only 2.0% of the long-term forecast revision variation *due to quantitative information* represents misreaction.

---

<sup>6</sup>More formally, since the EPS forecasts are scaled by price to ensure stationarity (Bouchaud et al., 2019; Van Binsbergen et al., 2023; De Silva and Thesmar, 2024), a  $-0.3$  TIGC coefficient means that when an analyst increases their EPS forecast by 1% relative to the current price due to a topic, this revision overshoots the realized EPS number by 0.3% relative to the current price on average.

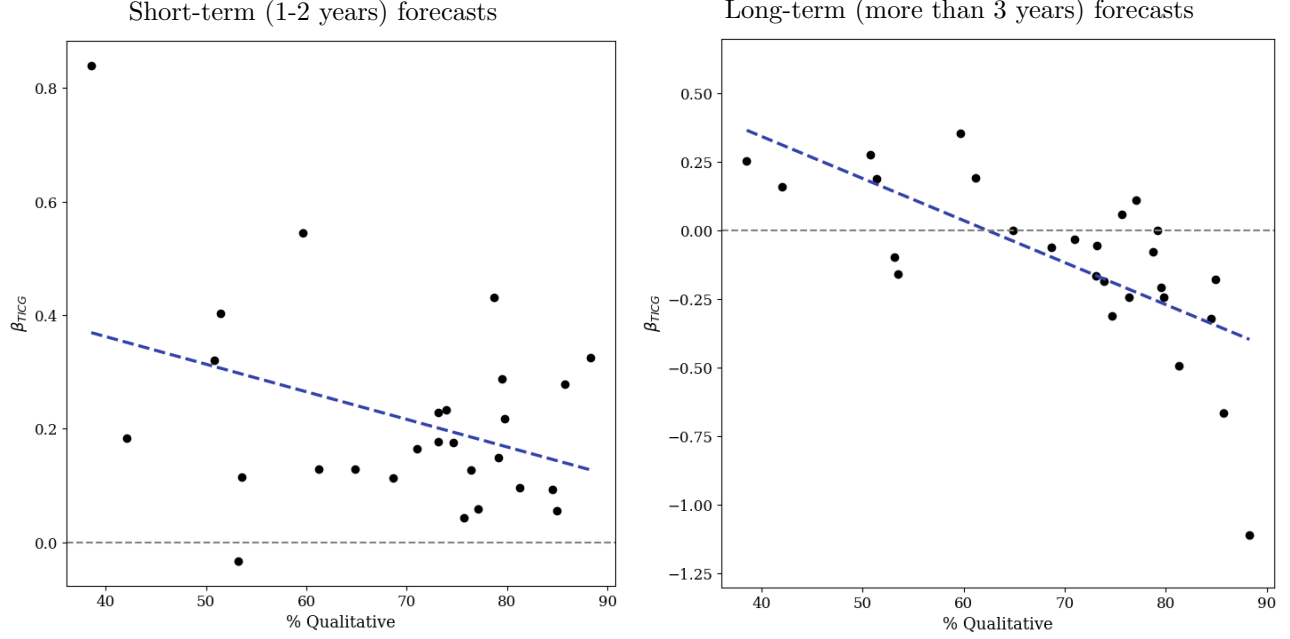


Figure 1: Text-instrumented CG coefficients versus topic qualitiveness

*Notes:* The scatter plots show the relationship between text-instrumented CG-regression coefficients and topic qualitiveness. Each dot represents one topic, and a positive (negative)  $\beta_{TICG}$  indicates underreaction (overreaction) to that topic. The x-axis shows the fraction of qualitative information tuples in a topic. For further detail see Section 5.

I conduct several robustness tests to validate the above result. First, I show that the TICG regression results are robust to the dimensionality of the embedding vectors used. Second, I find similar results using the original CG regression but interacting forecast revisions with topic weights: when analysts focus more on qualitative topics like product development, marketing, or R&D, the original CG coefficients for long-term forecasts become more negative, indicating larger overreaction.<sup>7</sup> Third, I estimate TICG regressions in subsamples that are constructed based on analyst and broker features related to institutional distortions. While the degree of overreaction to qualitative topics varies across subsamples, the overall negative relationship remains consistent with the pattern depicted in Figure 1. This suggests that institutional incentives are unlikely to explain this pattern. In particular, overreaction is stronger among less experienced analysts and those working at smaller brokerage firms, suggesting that this tendency reflects a behavioral bias that diminishes with experience.

In addition, the overreaction to qualitative information has significant implications for asset pricing. I find that revisions of long-term earnings forecast due to qualitative information negatively predict stock returns in the cross-section, whereas revisions due to quantitative information

<sup>7</sup>In general, the topic weight-interacted CG-regression coefficients do not have the structural interpretation of a misreaction to specific signals. However, it is a convenient diagnosis of the original CG coefficient's dependence on topic weights.

lack this predictive power. A long-short portfolio that bets against qualitative earnings revisions achieves an annualized Sharpe ratio of 0.85 and has significant alphas to prominent factor models. In addition, the returns of such a portfolio also explain the long-term behavioral mispricing factor in [Daniel et al. \(2020\)](#).

**Uncovering the behavioral mechanisms in text.** After using the TICG regressions to identify the information to which analysts over- and underreact, I turn to the question: What behavioral mechanisms underlie these documented misreactions?

The self-disclosed description of the belief formation process, such as the text in analysts’ reports, offers an opportunity to study the mechanisms of belief formation. Specifically, for different behavioral mechanisms or psychological motivations, researchers can look for direct evidence in the language and assess how well each explains the observed behavioral biases. Building on this premise, I examine which behavioral mechanisms can empirically account for the two misreaction patterns I document: (1) an overreaction to qualitative information in long-term forecasts, and (2) a distinct overall tendency to underreaction that mainly affects short-term forecasts.

I explore three overreaction mechanisms: overconfidence, memory effects, and overinference from weak signals. Each mechanism can potentially explain why analysts overreact more to qualitative than to quantitative information. Analysts may overreact because they are overconfident in their ability to interpret qualitative information, which allows for more room for interpretation and thus greater potential for overestimation of the precision of a signal ([Odean, 1998](#)). As regards memory effects, qualitative information may trigger stronger associative recall of similar past events, which can distort probability assessments and lead to overreaction ([Graeber et al., 2024](#); [Enke et al., 2024](#)). Analysts may also overreact to qualitative topics if these provide weaker signals than quantitative ones, and analysts overinfer from weak signals when uncertain about their strength ([Augenblick et al., 2025](#)).

I conduct the analysis by prompting the LLM to look for verbal evidence of overconfidence and memory mechanisms in analysts’ written language. Specifically, I look for assertive language (such as “we are very confident” or “without doubt”) as indications of confidence and references to past events or episodes as indications of memory. I also directly predict future earnings growth using embedding vectors of qualitative topics to assess whether they are indeed weaker than quantitative ones. I find that when analysts use more assertive language in their subjective statements, their revisions due to qualitative information predict forecast errors more negatively, indicating a stronger overreaction to qualitative information. This suggests that analysts’ overconfidence in their ability to interpret qualitative information is an important mechanism explaining overreaction to qualitative information in long-term forecasts.

I also examine whether two underreaction mechanisms—sticky beliefs and herding—can explain

the overall underreaction in short-term forecasts. In the sticky belief model, analysts adhere to their previous forecasts, leading to underreaction to new information (Bouchaud et al., 2019). I capture this behavior by prompting the LLM to detect references to an analyst’s prior forecast in the report. In the herding model, analysts follow the consensus forecast and ignore their own signals, which also results in underreaction in beliefs (Valchev and Gemmi, 2023; Banerjee et al., 2025). I examine this tendency by prompting the LLM to identify references to consensus forecasts (such as “street estimates” or “consensus opinions”) in the reports. I find that the more frequently analysts mention consensus forecasts in their reports, the more positive the CG coefficients become, suggesting herding is an important driver of analyst underreaction.<sup>8</sup>

Motivated by these findings, I develop a stylized belief-formation model within the broader conceptual framework to capture the empirical patterns. In the model, analysts are overconfident when interpreting qualitative topics, which I model as their believing that qualitative signals have higher precision than they actually do. Analysts also exhibit herding behavior as they face an additional incentive to stay close to consensus forecasts. This model explains the new empirical findings documented in the paper. It also predicts that the herding mechanism affects short-term forecasts more strongly than long-term forecasts because more analysts issue short-term forecasts than long-term forecasts, for which I find supportive evidence.

## 1.1 Related Literature

This paper connects to the extensive literature on subjective beliefs about earnings. Prominent models of belief formation concerning subjective cash-flow expectations include models that rely on conservatism and extrapolation (Barberis et al., 1998), sticky beliefs (Bouchaud et al., 2019), belief in the reversal of earnings growth (Delao and Myers, 2021), diagnostic expectations (Bordalo et al., 2024b), learning with fading memory (Nagel and Xu, 2022), and noisy cognitive defaults (De Silva and Thesmar, 2024). These various studies build belief-formation models based solely on the empirical patterns in analysts’ numerical earnings forecasts.

This paper, by contrast, directly examines how analysts explain the formation of their beliefs in their written text. By studying the text in analyst reports, I provide direct evidence of how analysts actually establish their beliefs. I do so I shed light on important aspects of belief formation that researchers cannot study using only numerical measures of beliefs, such as analysts’ information focus, their heterogeneous misreactions to information, or the underlying behavioral mechanisms. This approach is similar to that seen in an emerging literature in behavioral economics where open-ended surveys or textual data are used to understand people’s thinking (Choi, 2022; Andre

---

<sup>8</sup>Additionally, I classify forecast revisions into “herding” and “non-herding” types following Clement and Tse (2005), and show that herding forecasts are associated with more positive CG coefficients, supporting herding as the main underreaction mechanism.

et al., 2023, 2024; Haaland et al., 2024; Binetti et al., 2024; Laarits et al., 2025).<sup>9</sup> In particular, this paper is similar to those using surveys to distinguish between different observationally equivalent behavioral models or hypotheses (Liu et al., 2022; Chincio et al., 2022; Chen et al., 2023a).

The findings here contribute to the literature on the assessment of misreaction in analysts' forecasts. The TICG offers a new avenue for identifying the information to which analysts misreact, an aspect that traditional tests like CG regression cannot address. My findings suggest that analysts' overreactions depend on the type of information encountered, complementing existing research on misreaction in earnings forecasts that uses only numerical forecast data (Bouchaud et al., 2019; Delao and Myers, 2021; Chaudhry, 2023; Bordalo et al., 2024c,a; De Silva et al., 2025). I also provide evidence that overconfidence and herding are important drivers of the observed misreaction (Odean, 1998; Gervais and Odean, 2001; Clement and Tse, 2005; Ben-David et al., 2013; Valchev and Gemmi, 2023; Banerjee et al., 2025).

This paper also relates to the study of textual information in analysts' written reports.<sup>10</sup> Despite their apparent relevance, analyst reports have not been systematically used to study the drivers of analysts' subjective expectations. Several recent studies have used analyst's reports to study how they establish price targets: Decaire and Guenzel (2023) and Decaire and Graham (2024) explore subjective terminal growth and discount rates in DCF models, and Ben-David and Chincio (2024) focus on price multiples. Bastianello et al. (2024) study how analysts' attention to different topics relates to their valuation model and how they explain differences in price targets. Chen et al. (2023a) use analyst reports to determine why investors invest in stocks with low expected returns. In contrast, this paper focuses on earnings expectations, which are a convenient setting for computing forecast revisions and extending the classic CG regression framework to study over- or underreaction to information. In this sense, this paper also contributes to the extensive literature using CG regressions as the main tool to detect belief misreaction.<sup>11</sup>

Finally, this paper contributes to the emerging agenda on the use of generative artificial intelligence (AI) and LLMs in behavioral finance research. I show that using LLMs in a RAG framework allows researchers to flexibly extract information from unstructured documents, which can be important for studying subjective beliefs. More broadly, this paper connects to the extensive literature in which machine learning is employed in financial research (see Kelly et al., 2023)).

---

<sup>9</sup>More broadly, this paper connects to the emerging literature employing new data about beliefs in financial markets, for example, (Cookson and Niessner, 2020; Cookson et al., 2024; Coutts et al., 2023; Dahlquist and Ibert, 2024; Coutts et al., 2024; Gormsen and Huber, 2024, 2025).

<sup>10</sup>Asquith et al. (2005), Huang et al. (2014), and Huang et al. (2018) study investor reactions to analyst reports by looking at short-term movements in stock prices. Recent work by Bellstam et al. (2021) employs topic modeling to measure corporate innovation in analyst reports. Li et al. (2023) use a RAG framework similar to mine to assess corporate culture. Chi et al. (2024) identify the use of alternative datasets in analyst reports and study how this affects forecasts.

<sup>11</sup>Examples include Coibion and Gorodnichenko (2015), Bordalo et al. (2020), Wang (2021), Angeletos et al. (2021), Kohlhas and Walther (2021), d'Arienza (2021), Afrouzi et al. (2023), De Silva and Thesmar (2024), Kelly et al. (2024), De Silva et al. (2025), and Halperin and Mazlish (2025).

In particular, I contribute to the literature on the use of machine learning to study analysts’ earnings forecasts (So, 2013; Van Binsbergen et al., 2023; De Silva and Thesmar, 2024) by showing that their written texts are an important and informative source for the study of systematic forecasting mistakes.<sup>12</sup>

**Outline.** The remainder of this paper is structured as follows. Section 2 presents a conceptual framework for studying belief formation through texts. Section 3 describes the analyst-report data and the LLM-based process for information extraction. Section 4 documents the results concerning variation in information focus. Section 5 examines over- and underreaction to specific information, with a focus on how misreaction differs based on the qualitateness of the topic. Section 6 employs textual evidence to investigate the behavioral mechanisms behind over- and underreaction. Section 7 concludes.

## 2 A framework for belief formation in texts

I first present a theoretical belief-formation framework in which the analyst incorporates disaggregated information sets. I highlight that the original CG regression is ill-suited to studying the drivers of over- and underreaction, and that this challenge can be mitigated by incorporating self-disclosed text describing the process of belief formation. I then introduce the TICG regression, a new procedure that leverages textual embeddings to identify the misreaction of analysts to different types of information. This section provides the conceptual foundation for the empirical analysis that follows.

### 2.1 Model Setup

The framework is motivated by Kacperczyk et al. (2016) and Kohlhas and Walther (2021). For exposition, I consider a model in which one analyst makes a forecast for one firm in a static setting. Let  $y_{t+1}$  denote the firm’s future earnings that the analyst is trying to forecast. The analyst observes a set of  $K$  signals  $s_{k,t}$  of the future earnings, which takes the following form:

$$s_{k,t} = y_{t+1} + v_{k,t}. \quad (1)$$

I assume both  $y_{t+1}$  and  $v_{k,t}$  are Gaussian and  $v_{k,t}$  are independent noise. I use  $F_t[\cdot]$  to represent the analyst’s subjective belief at time  $t$ , and I use  $\mu_{t-1} = F_{t-1}[y_{t+1}]$  to denote the analyst’s prior belief at time  $t - 1$  when entering period  $t$ . For simplicity, I assume  $\mu_{t-1}$  is correct (rational), which means the forecast error is exclusively driven by misreaction to information. This allows me to

---

<sup>12</sup>Other recent works using machine learning to study beliefs include Bianchi et al. (2022), Goetzmann et al. (2022), Bali et al. (2023), Bybee (2023), Charles and Sui (2024), Jensen (2024), Bianchi et al. (2024b), van Binsbergen et al. (2024), Fedyk et al. (2024), Chen (2025), Jha et al. (2025), Sarkar (2025), and Bini et al. (2025).

interpret the subsequent CG-type regression analysis as a study of misreaction to new information in  $s_{k,t}$ , which aligns with the common interpretation in the literature. It also rules out predictable forecast errors due to parameter learning (Farmer et al., 2024; Li et al., 2025).<sup>13</sup>

Given the structure of  $y_{t+1}$  and  $\{s_{k,t}\}_{k=1}^K$ , I model the revision in the analyst’s subjective forecast as the weighted combination of the new information from all  $K$  signals, plus a residual term  $\epsilon_t$  that captures the unmodeled variation in belief revisions:

$$Rev_t[y_{t+1}] := F_t[y_{t+1}] - F_{t-1}[y_{t+1}] = \sum_{k=1}^K \hat{w}_k(s_{k,t} - \mu_{t-1}) + \epsilon_t \quad (2)$$

where  $\hat{w}_k$  captures the weight that they place on signal  $k$ .  $\epsilon_t$  captures beliefs based on reasons other than information about these  $K$  components, such as unconditional bias, reactions to information not included in the  $K$  signals, distortion of institutional incentives, or simply noise in expectations.

Similarly, I use  $E_t[y_{t+1}]$  to denote the rational expectation at time  $t$  for the firm’s future earnings, the revision of which is given by the following:

$$E_t[y_{t+1}] - E_{t-1}[y_{t+1}] = \sum_{k=1}^K w_k(s_{k,t} - \mu_{t-1}), \quad (3)$$

where  $w_k$  is the rational (Kalman gain) on signal  $s_k$ , and note that  $\mu_{k,t-1}$  is assumed to be rational. Given this formulation, I interpret the differences  $w_k - \hat{w}_k$  to be the misreaction to information in signal  $k$ . If an analyst overreacts to information in signal  $k$ , they put excessive weight on  $s_k$  and  $w_k - \hat{w}_k < 0$ . Similarly, if they underreact to information in signal  $k$ , we have  $w_k - \hat{w}_k > 0$ . Without loss of generality, I assume that both  $w_k$  and  $\hat{w}_k$  are positive.<sup>14</sup>

## 2.2 Challenges in identifying misreaction to information

A widely used approach to the study of belief misreaction is to run the classic CG regression of forecast errors on forecast revisions, in the form of

$$y_{t+1} - F_t[y_{t+1}] = \beta_{CG} \cdot (F_t[y_{t+1}] - F_{t-1}[y_{t+1}]) + e_t. \quad (4)$$

The CG regression provides a general framework for empirically documenting over- or underreaction to information in forecasts.<sup>15</sup> A positive  $\beta_{CG}$  coefficient estimate is typically interpreted as

<sup>13</sup>In Table OA.10, I find that lagged forecasts do not have strong predictability of forecast errors in IBES, and that CG coefficient remains significant after controlling for lagged forecasts.

<sup>14</sup>This is innocuous because we can always “negate” the signal so that it makes a positive contribution to  $y$  when  $s_k$  increases. The assumption that  $w_k$  and  $\hat{w}_k$  have the same sign ensures a clear interpretation of over- and underreaction.

<sup>15</sup>I subsume the constant term for the purpose of exposition. In the empirical implementation, I include time-, stock- and analyst-fixed effects.



an underreaction: when the forecaster revises a forecast, a positive  $\beta_{CG}$  implies the revision is not sufficiently large, and the forecast undershoots the target. Conversely, a negative  $\beta$  is interpreted as an overreaction to information.

However, in a setting where the forecaster observes disaggregated signals, the baseline CG coefficient cannot be readily interpreted as a misreaction to any specific piece of information; it only identifies a weighted average of the misreactions across all signals. Moreover, the presence of non-information-driven beliefs (like the noise in beliefs) introduces a negative bias in the CG coefficient.<sup>16</sup> Formally, the CG coefficient in the model is characterized in the following proposition:

**Proposition 1** (Baseline CG coefficient with disaggregated signals)

*Suppose that the analyst forms their belief according to Eq. (2). Then, the CG coefficient in the regression  $y_{t+1} - F_t[y_{t+1}] = \beta_{CG} \cdot (F_t[y_{t+1}] - F_t[y_{t+1}]) + e_t$  is given by*

$$\beta_{CG} \propto \sum_{k=1}^K \gamma_k (w_k - \hat{w}_k) - \sigma_\epsilon^2, \quad (5)$$

where  $\gamma_k$  is a weight coefficient that depends on all  $\hat{w}_k$  and the covariance matrix of signals  $s_k$ , and  $\sigma_\epsilon^2$  is the variance of the non-informational component  $\epsilon_t$  in beliefs in (2).

Proposition 1 shows that  $\beta_{CG}$  only uncovers a weighted average of misreaction parameter  $w_k - \hat{w}_k$  and suffers from a negative bias when analysts' forecast revisions are not information-driven. The core issue is that the CG regression does not include actual information; the forecast revision is merely a proxy for information, and this proxy becomes difficult to interpret when analysts are reacting to multiple signals.

Researchers may also be tempted to directly regress forecast errors on signals they are able to observe, with the hope that the predictability of these will reveal misreaction to specific types of information.<sup>17</sup> However, this approach typically suffers from omitted variable bias.<sup>18</sup> For example, if a researcher regresses earnings-forecast errors on a stock's past returns and finds a positive coefficient, this does not necessarily imply underreaction to price changes; the analyst might instead be underreacting to past earnings growth, and the positive coefficient arises because of the correlation between earnings growth and past returns.<sup>19</sup>

<sup>16</sup>See Juodis and Kučinskas (2023) and De Silva and Thesmar (2024) for further discussion on the noise in expectations.

<sup>17</sup>For example, Hribar and McNnis (2012) regress analyst forecast errors on measures of investor sentiment, Walther and Willis (2013) use measures of investors' macroeconomic expectations, Faralli (2024) uses extreme weather, Ma (2025) uses an index of firm technological obsolescence, and Xu et al. (2025) uses a firm's fiscal risk. See also So (2013) for a study that predicts forecast errors using observable stock characteristics.

<sup>18</sup>See Proposition 3 in the Appendix for a formal argument.

<sup>19</sup>One approach to mitigating this concern is to use "exogenous" shocks to isolate variation in a single signal that is uncorrelated with others. For example, Angeletos et al. (2021) and Kučinskas and Peters (2024) use macroeconomic or inflation shocks derived from VARs, Chaudhry (2023) instruments earnings forecasts using flow-induced shocks to stock prices, and Derrien et al. (2025) considers ESG-related news. However, for many types of

In summary, studying belief misreaction to information is challenging when the forecaster’s information set is typically unobserved. The main contribution of this paper is to show that the textual descriptions underlying numerical forecasts are a significant step toward uncovering the belief-formation process, as I show in the approach developed in the following section.

## 2.3 Uncovering belief formation from text

**Operational formulation of belief formation.** While the classic formulation of belief formation in Eq. (2) allows researchers to easily define and interpret over- and underreaction, it is highly stylized and not easily mapped to real-world circumstances. Therefore, I define an equivalent, operational formulation that allows for an easier mapping to the text in analysts’ reports.

First, in reality, rather than a *single* signal  $s_k$ , analysts observe some general and broadly defined *informational content* related to different aspects of the firm. For example, in practice, there does not exist a univariate signal for “firm profitability”; analysts instead observe accounting numbers in corporate filings, listen to the CEO’s remarks on sales in earnings calls, talk to industry experts, and so on. I assume that the new information about aspect  $k$  observed by the analyst can be represented as a numerical vector  $\tilde{X}_t^k$ , which can be high dimensional to capture all relevant aspects of the new information related to aspect  $k$ .

Next, given the numerical representation  $\tilde{X}_t^k$  for  $k = 1$  to  $K$ , I model the analyst’s revision of  $y_{t+1}$  as the mapping from  $\{\tilde{X}_t^k\}_{k=1}^K$  to a scalar,

$$Rev_t[y_{t+1}] = \sum_{k=1}^K \lambda'_k \tilde{X}_t^k + \epsilon_t, \quad (6)$$

where  $\lambda_k$  captures how the analyst revises their belief in response to the new information about firm’s aspect  $k$ . I define this formulation to be equivalent to the classic formulation in Eq. (2) as follows.

**Assumption 1** (Equivalence of classic and operational formulation of belief formation)

*The classic formulation in Eq. (2) and the operational formulation in Eq. (6) are assumed to be equivalent, which means*

$$Rev_t[y_{t+1}] = \sum_{k=1}^K \hat{w}_k(s_{k,t} - \mu_{t-1}) + \epsilon_t = \sum_k \lambda'_k \tilde{X}_t^k + \epsilon_t \quad (7)$$

$$\hat{w}_k(s_{k,t} - \mu_{t-1}) = \lambda'_k \tilde{X}_t^k.$$

---

information, identifying valid exogenous shocks is a challenge.

### 2.3.1 Text as a sufficient statistic of belief revisions

Having stated the classic and operational formulations of belief formation, I briefly discuss how the theoretical framework motivates the extraction of information and the representation of text. More details on the text representation can be found in Section 3.

First, the  $K$  different aspects of a firm’s fundamentals can be captured by the textual topics that the analyst describes in their report. These topics can then be viewed as a collection of nouns. For example, when the analyst writes about “revenue,” “profit margins,” or “sales,” they are describing information about the firm’s profitability, whereas when they write about “inventory,” “products,” or “customers,” they are describing information about the firm’s business operations.

Next, for each topic  $k$ , I construct an embedding vector of the text related to this topic in the report, which I denote using  $X_t^k$ . An embedding vector is a high-dimensional vector that captures the semantic meanings of the text, therefore encodes the semantic information in the text.<sup>20</sup> I present the details of the embedding vector constructions in Section 3.2.

For the text to be informative about the belief formation process, I introduce the following assumption that relates the empirically measured embedding vector  $X_t^k$  to the informational content  $X_t^k$  that the analyst observes:

**Assumption 2** (Text as sufficient statistics of belief revision)

*The embedding vector of topic  $k$  is a sufficient statistic for the information that drives the analyst’s beliefs about component  $k$ ,*

$$X_t^k = \tilde{X}_t^k; \quad X_t^k \perp \epsilon_t \quad (8)$$

Assumption 2 essentially states that the text related to topic  $k$  is sufficiently rich, in the sense that the analyst has described all the relevant aspects of  $s_k$  that affect their belief distribution. The assumption implies that once we know the textual content of analyst’s reports about topic  $k$ , we can (perfectly) determine how they respond to information related to component  $k$ . While admittedly a strong assumption, it provides a useful benchmark for considering how a self-disclosed text relates to a person’s beliefs.

First, this assumption follows the tradition of interpreting survey data as evidence of genuine beliefs. Second, it reflect the fact that analyst reports must adhere to regulatory requirements, such as FINRA Rule 2241, which mandates that “purported facts in research reports [be] based on reliable information” and requires that conflicts of interest be properly managed to “promote objective and reliable research that reflects the truly held opinions of research analysts and to prevent the use of research reports or research analysts to manipulate or condition the market or

---

<sup>20</sup>See Section 4.8 of [Hoberg and Manela \(2025\)](#) for an introduction to embedding methods in finance and [Hanley and Hoberg \(2019\)](#), [Chen et al. \(2022\)](#), [Mazumder et al. \(2023\)](#), [Kim and Nikolaev \(2024\)](#), and [Sarkar \(2025\)](#) for recent applications to finance of embeddings to capture information in text.

favor the interests of the member or a current or prospective customer or class of customers.”<sup>21</sup> Also, buy-side investors often view the written content of these reports as more important than the quantitative forecasts, creating an incentive for analysts to provide truthful research-based justifications for their forecasts.<sup>22</sup>

In practice, however, there are many reasons why the text embeddings,  $X^k$ ’s, in analyst reports might not provide a sufficient statistic for analysts’ belief revisions, topic misclassification, analysts’ preferences for withholding relevant information, or ex-post justifying their beliefs in response to certain institutional incentives. In Section 2.5, I discuss these potential violations of Assumption 2. In particular, when such violations are limited, the TICG still identifies the *sign* of misreaction, that is, whether the analyst over- or underreacts, although the estimated magnitude of the misreaction will be biased.

## 2.4 Text-instrumented Coibion-Gorodnichenko Regression

I now introduce the two-step TICG regression as a tool for studying misreaction to specific information. In the first step, forecast revisions are projected onto the textual information associated with various topics, identifying how much of any revision is driven by each topic. In the second step, for each topic, forecast errors are regressed on the predicted revisions while controlling for the embeddings of other topics.

1. Given observed forecast revisions  $Rev_t[y_{t+1}]$  and the text embedding vectors  $X_t^k$  for topic  $k = 1, 2, \dots, K$ , regress

$$Rev_t[y_{t+1}] = \sum_{k=1}^K \lambda_k' X_t^k + u_t \quad (9)$$

Let  $\hat{\lambda}_k$  be a consistent estimator of  $\lambda_k$  and  $\widehat{Rev}_t[y_{t+1}] := \sum_{k=1}^K \hat{\lambda}_k' X_t^k$  the predicted revisions.

2. For each topic  $k$ , regress forecast errors on the predicted revisions while using the embeddings

---

<sup>21</sup>See <https://www.finra.org/rules-guidance/rulebooks/finra-rules/2241>. Similar requirements can be found in the US Securities and Exchange Commission (SEC) Regulation Analyst Certification: <https://www.sec.gov/rules-regulations/2003/02/regulation-analyst-certification>. Regulators enforce these rules and can impose fines or suspend and analyst’s certification, including in cases involving major investment banks like [Deutsche Bank](#) and [Barclays](#). Andrew J. Ceresney, a former director of enforcement for the SEC remarked to the [New York Times](#) that: “When research analysts tell clients to buy or sell a particular security, the rules require them to actually mean what they say. Analysts simply cannot express one view publicly and the opposite view privately.”

<sup>22</sup>For example, a survey conducted by the *Institutional Investor* magazines shows that the written report is the fifth most important consideration when investors vote for “All-Star” analysts, while earnings forecasts rank only twelfth. “In the end, stock ratings and target prices are just the skin and bones of analysts’ research. The meat of such reports is in the analysis, details, and tone.” ([Bloomberg: When a Stock’s Rating and Target Collide. 2002-04-24.](#)) See Table 1 in [Bradshaw \(2011\)](#) for further details on the importance of the *Institutional Investor* survey. Anecdotal evidence from analysts at long-short equity hedge funds confirm that reading sell-side analyst reports is an important part of their research to generate trade ideas.

of other topics  $k' \neq k$  as exogenous controls,

$$y_{t+1} - F_t[y_{t+1}] = \beta_{TICG}^k \widehat{Rev}_t[y_{t+1}] + \sum_{k' \neq k} \xi_{k'}' X_t^{k'} + e_t. \quad (10)$$

The following proposition shows that the text-instrumented CG-regression coefficient  $\beta_{TICG}^k$  can be interpreted as the degree of over- or underreaction to information about signal  $s_{k,t}$  in the belief formation framework introduced in Section 2.1.

**Proposition 2** (Identifying misreaction to granular information using the TICG regression)

*Given the two equivalent formulations of belief updating in Assumption 1 and Assumption 2, the TICG regression has the coefficient*

$$\beta_{TICG}^k = (w_k - \widehat{w}_k) / \widehat{w}_k.$$

*Therefore, with the assumption that  $\widehat{w}_k > 0$ , it identifies underreaction to information about  $k$  if  $\beta_{TICG}^k > 0$  and overreaction to information about  $k$  if  $\beta_{TICG}^k < 0$ .*

Proposition 2 demonstrates that the two-stage procedure in the TICG regression reveals misreaction to information at the topic level, offering a powerful means to diagnose the source of systematic forecast errors in analyst’s forecasts. Intuitively,  $\beta_{TICG}^k$  measures, on average, how much the analyst’s forecast revision overshoots or undershoots the realized outcome when they revise *only due to information about topic  $k$* . The assumption that  $\widetilde{X}_t^k = X_t^k$  for all topics enables researchers to directly control for the information related to other topics  $k' \neq k$  and uncover the forecast error predictability only from topic  $k$ . As an example, a  $\beta_{TICG} = -0.3$  for the R&D topic means that when the analyst revises his EPS forecast upward by \$1 only due to the firm’s R&D information, on average \$0.30 of that upward revision will be excessive. Therefore, unlike the original CG regression, the text-instrumented version explicitly incorporates the “information” observed by the analyst into the empirical procedure, enabling a more granular test of deviations from rationality.

**Estimation, pre-test, and inference.** Note that the first-stage (9) is essentially the first-stage regression of  $Rev_t[y_{t+1}]$  on  $X_t^k$  while controlling for  $X_t^{k'}$  for  $k' \neq k$ , we can estimate TICG regressions using the standard two-stage least square (TSLS) estimator. A distinct issue in this setting is that, when implementing the TICG regressions, capturing the textual information related to topic  $k$  typically requires that the embedding vector  $X^k$  be high dimensional. This introduces econometric challenges similar to those in a “many instrumental variables” setting. I follow the standard approach in the literature (Angrist and Krueger, 1995; Angrist et al., 1999) and apply jackknife estimation in the first stage. Another concern is that textual embedding vectors might contain substantial noise, rendering them “weak instruments.” I address this following Mikusheva

and Sun (2022, 2024) and compute the F-statistic for many weak instruments. The results indicate that weak-instrument bias is not a significant concern in the empirical implementation. Further detail on the estimation, inference, and pre-tests is provided in Appendix B.7.

## 2.5 Potential violations of the sufficient statistics Assumption

In this paper, I assume the text of analysts’ reports provides a sufficient description of the true drivers of their beliefs, as stated in Assumption 2. If there are distortions, such as measurement error in the text representation, topic misclassification, or incentives for analysts to engage in strategic communication, this assumption will not hold. I model these violations as a general form of measurement error in the text-embedding vectors that are potentially correlated with the “truly observed” informational and non-informational drivers of analysts’ beliefs. Specifically, I model the text embedding  $X_t^k$  as

$$X_t^k = \tilde{X}_t^k + E_t^k, \quad (11)$$

where  $E_t^k$  captures the distortions in text embeddings, which can be correlated with the true informational content  $X_t^k$ , the non-informational drivers  $\epsilon_t$ , and across topics. In Appendix B.1, I discuss how different violation types can be represented in this way.

When there exists measurement errors  $E_t^k$ ,  $\beta_{TICG}^k$  will provide a biased inference concerning the true misreaction parameter  $b_k$ .<sup>23</sup> I provide an explicit formula for  $\beta_{TICG}^k$  under the general measurement-error structure in Appendix B.1. In particular, if we are only interested in the *sign* of the misreaction and determining whether the analyst is over or underreacting, then the TICG regression still identifies the correct signs when the violation of Assumption 2 is not too large relative to the magnitude of the misreaction parameter  $b_k$ .

**Simulation exercise.** I validate the efficacy of TICG regressions and study the impact of potential violations of Assumption 2 by conducting simulations that mimic the empirical settings; these are set out in Appendix B.2. The simulation results show first, that it is necessary to use a jackknife estimator in the TICG to deal with the high-dimensional embedding vectors. Second, they show that the procedure can allow for a certain degree of violation of Assumption 2 and still identify the sign of misreaction. Finally, they provide a sense of which empirical patterns are diagnostic of potential violations. For example, when  $X_t^k \perp \epsilon_t$  is violated, we would expect to see topics that explain more variation in forecast revisions having a more negative  $\beta_{TICG}$ . This is at odds with the empirical result in Section 5 that topics with a lower level of explained variance have a more negative  $\beta_{TICG}$ . This indicates that the violation of  $X_t^k \perp \epsilon_t$  is unlikely to be a key driver of the empirical results in Section 5.

---

<sup>23</sup>Quantifying and correcting biased inference due to measurement error in variables constructed by generative AI is an active area of econometric research, for example, Battaglia et al. (2025) and Ludwig et al. (2025). These methods usually require a human-labeled validation dataset.

**Empirical robustness.** In addition to showing a certain degree of robustness through theory and simulation, I also attempt to address these issues empirically. First, to mitigate topic misclassifications and missing information, I use a multi-step information-extraction process using LLMs. I use explicit prompts to determine if the information extraction is completed, and employ multiple LLMs for topic classification to minimize idiosyncratic noise.

Second, I address institutional distortions empirically by controlling for observable institutional features. In the baseline implementation of the TICG regressions, I include time-, stock-, and analyst-fixed effects, which account for various incentive distortions at these levels. In Section 5.4, I conduct robustness checks that incorporate additional brokerage- and analyst-level characteristics, such as underwriting affiliation, reputational capital, investment-banking pressure, broker loyalty, and institutional ownership. If distortions due to institutional incentives are tied to these features (for example, affiliated analysts may be more inclined to overreact), one would expect heterogeneous misreaction. I find that while there is a certain degree of heterogeneity across broker and analyst characteristics, the main misreaction patterns remain.

## 3 Data and methodology

### 3.1 Data

#### 3.1.1 Analyst reports

I obtain sell-side equity research analyst reports from Investext, an aggregator provided by Mergent Online. Investext stores these reports in PDF format and each publication is accompanied by metadata that include the report title, publisher (investment bank or brokerage firm), authors (sell-side analysts), company names and tickers mentioned, as well as Investext’s custom tags such as report style, category, and subject. The sample starts on 01/01/1998 and ends on 09/30/2023. See Appendix A.1 for details on the downloaded reports.

For each downloaded analyst report, I match the stocks mentioned with their CUSIP codes based on the ticker and company name provided by Investext (see Appendix A.3). I then match the report authors with IBES analyst codes (`amaskcd`) using authors’ first and last names and their stock coverage. Additionally, I match contributors of the reports with IBES estimator IDs (`estimator`) based on name resemblance and analyst affiliations (see Appendix A.4 for details). I focus on 115 brokerage firms for which the mapping between Investext contributor and IBES estimator IDs is validated by the matching tables provided by Flake (2023) and Law (2023). These brokerage firms include major investment banks JP Morgan, Credit Suisse, Deutsche Bank, Wells Fargo, and RBC, as well as independent research firms such as Argus Research and Wolfe Research.



Finally, I match analyst reports to the announcements in the IBES unadjusted detail file. Specifically, I attempt to match each announcement to IBES (as identified by the `cusip-estimator-anndats`) with research reports written by the same institution about the same stock within  $\pm 1$  business day of the announcement date (`anndats`).<sup>24</sup> Appendix A.6 details the matching process. As a result, I am able to match 818,780 EPS announcements made by the 115 institutions with 1,144,952 unique analyst reports, covering about 50% of all announcements made by the 115 institutions in IBES and 23% of the total IBES announcements in the sample period.<sup>25</sup> In A.6.1 I report the summary statistics of the matched sample; unconditional moments for key forecast variables suggest the matched sample is representative of the full IBES sample.

The analyst reports are stored in PDF format, which is unstructured and requires parsing. I process these reports using the deep-document-understanding pipeline developed by InfiniFlow. This pipeline uses various deep learning models in computer vision and textual analysis, including optical character recognition, layout recognition, text recognition, and table recognition. Following these steps ensures that textual information is preserved during parsing, thereby maintaining data quality. See Appendix A.2 for details on the parsing procedure.

### 3.1.2 Subjective expectation data

I obtain analyst forecast data from IBES, a widely used source for studying equity analysts' subjective expectations. My primary focus is analysts' beliefs regarding firm earnings, measured through their EPS forecasts for future fiscal years (FY1 to FY5). I obtain the individual forecasts from the IBES unadjusted detail file. The process of cleaning and processing of IBES data largely follows the literature and is detailed in Appendix A.5.<sup>26</sup>

## 3.2 Information extraction from analyst reports

Implementing the TICG regressions requires that the information related to different topics be extracted and represented as embedding vectors. Traditional methods of textual analysis in eco-

---

<sup>24</sup>Previous studies, for example, Huang et al. (2014) and Huang et al. (2018) match forecasts to reports published between IBES `anndats` and `revdats`, which is the most recent date on which the analyst reviewed the forecast and deemed it still effective. Matching only based on `anndats` allows me to focus on the reports that contain analysts' forecast revisions, rather than those that reiterate their previous forecasts.

<sup>25</sup>The number of reports exceeds the number of announcements because multiple reports can be associated with a single announcement. For example, a broker might issue a report on an EPS announcement and on the following day publish a research digest summarizing the contents of the original report. Since there is no way to enhance the match without reading the reports, I include all matched observations in my sample, which admittedly increases the noise in the data.

<sup>26</sup>A number of papers (La Porta, 1996; Bordalo et al., 2024a,c) focus on IBES estimates of long-term growth (LTG), which is the "expected annual increase in operating earnings over the company's next full business cycle" with a typical forecast horizon of 3-5 years. Given that IBES directly collects LTG measures from contributing analysts and that the actual LTG is not directly available, it is challenging to merge this with analyst reports and run CG-type regressions. I note, however, that the overreaction pattern in FY3 to FY5 numerical forecasts is consistent with the literature.

nomics and finance that rely on a bag-of-words approach are ill-suited for this type of information extraction.<sup>27</sup> In this paper, I extract information related to different topics using LLMs in a RAG framework. The framework has two components: a “retrieval” component in which the most relevant textual chunks are retrieved based on their contextual similarity to a user-defined query, and a “generation” component in which an LLM is used to answer the question based on these relevant chunks. Compared to bag-of-words approaches, RAG retrieves information based on contextual meaning in a process that mimics human comprehension and judgment. In the generation step an LLM is used to answer the user-specified questions based on the retrieved context.

I implement a multi-step procedure in the generation stage to ensure the validity and completeness of LLM-based information extraction. My primary tool is Llama-3.1-8b, a recent open-source model developed by Meta AI. This delivers state-of-the-art performance on a wide range of textual tasks and is comparable to closed-source alternatives. Its open-source nature allows flexible local deployment, unrestricted inference, and complete control over the generation process. For several key steps, I also use additional open-source models (e.g., DeepSeek and Phi) to validate results.

Figure 2 depicts the multi-step information extraction. I highlight four important steps in the procedure. Further details are provided in Appendix OA.1.

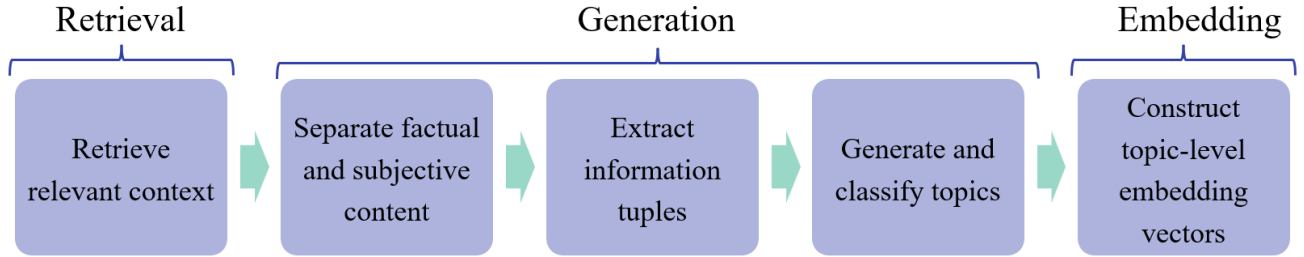


Figure 2: Illustration of information extraction from analyst reports

Notes: For a more detailed illustration see Figure OA.2 and detailed steps in Appendix OA.1.

First, I use the LLM to separate *factual* from *subjective* statements in analysts’ reports.<sup>28</sup> Typically, a report includes objective descriptions of recent firm developments and the analyst’s own outlook. As an example, consider two sentences:

- “TSLA reported lower EV unit delivery due to supply chain disruption in Q1.”
- We believe TSLA will trade lower as TSLA will face more challenges in EV delivery.”

<sup>27</sup>Examples of dictionary-based approaches include Tetlock (2007), Loughran and McDonald (2011), Hassan et al. (2019), and examples of topic modeling in economics and finance include Hansen et al. (2018), Bybee et al. (2024), Ke et al. (2024), and Freyaldenhoven et al. (2025). See Gentzkow et al. (2019) for a review of these traditional approaches to textual analysis.

<sup>28</sup>All prompts used are set out in Appendix OA.5.

While both sentences concern Tesla’s EV delivery, the first conveys information about an objectively verifiable event that has occurred and entered the analyst’s information set, whereas the second reflects a forward-looking belief. Separating factual from subjective content is crucial: the former captures what the analyst is reacting to, while the latter reveals how the analyst translates that information into beliefs. The framework in Section 2 formalizes the new signals of different components, which is connected to the *factual* information concerning different topics. I thus focus on the factual content in the following analyses.

Second, I use the LLM to extract information from the factual content in the form of a (*Noun*: *Description*) tuple, where *Noun* is a word or phrase of an object or concept, and *Description* is the phrase or sentence that describes and that captures the information related to it. This process produces 70,307,231 tuples with 9,528,530 unique nouns.<sup>29</sup>

Third, I use the LLM to generate topic labels from the extracted nouns by feeding *Nouns* to LLMs sequentially. I use open-sourced multiple LLMs to generate topic labels to guard against model instability. Different models generate 50 to 627 topic labels with certain overlaps and distinctions, but these topic lists are sufficiently small to allow human inspection. Based on manual review, I consolidate the labels into 153 topics, which I further aggregate into 28 meta-topics; the latter are referred to as “topics” in the subsequent sections. I then use the LLM to classify each noun as associated with one of these topics. Topic definitions are in Table OA.13, and in Appendix OA.9 I show the word cloud for each topic.

Finally, I generate text embeddings for the 28 topics by using the embedding vectors of the *Noun* and the *Description* in the extracted information tuples. For example, consider an information tuple  $s = (\text{“net margin”, “increased by 6\%”})$ . I represent the textual information associated with this tuple as

$$X_s = [\text{Embedding}(\text{“net margin”}), \text{Embedding}(\text{“increased by 6\%”})],$$

where  $\text{Embedding}(\cdot)$  denotes the embedding output. For embedding, I use ChronoBERT-v1999 developed by He et al. (2025), which produces a 768-dimensional vector for each input string. Due to the computational burden, I reduce the dimensionality by extracting the first 100 principal components of the embedding vectors for the *Nouns* and the *Descriptions*. The choice to use 100

---

<sup>29</sup>Table OA.12 in Appendix OA.6 sets out an example separating factual from subjective content and detailing the process of information extraction. In the original context, the analyst from Credit Suisse writes detailed factual information about BBWI’s loyalty program and its selling, general, and administrative (SG&A) expenses. The LLM can identify and extract the factual information accurately. Notice that, in the middle of the passage, the analyst writes “*we think it will be difficult to hold SG&A dollars flat*”, which is clearly their subjective view. The LLM can successfully identify and summarize this subjective content. In the last part of the example, the financial economic concepts extracted by the LLM are set out, along with their descriptive adjectives. Appendix OA.7 reports more examples of LLM responses, which further validate that the LLM provides satisfactory results that separate factual information from subjective statements and extracts meaningful information.

components is based on the fact that these capture more than 70% of the total variation in the original embeddings (see Figure OA.18), and I show the robustness of my main empirical results in embedding dimensions in Section 5.4. Each information tuple is thus represented by a 200-dimensional vector, comprising 100 dimensions each from the *Noun* and *Description* embeddings. The topic-level embedding vector is then computed as the average across information tuples in the topic.

### 3.3 Remarks on LLM implementation

While LLMs have demonstrated remarkable advantages over traditional natural language processing methods and are increasingly adopted in research, concerns remain regarding the validity of their outputs. In this section, I address several of these concerns.

**Traditional topic models.** First, one might argue that using LLMs for topic generation is excessive, given the availability of traditional methods of topic modeling. In Appendix OA.10, I show that alternative approaches, such as latent Dirichlet allocation and embedding-based clustering, perform poorly. Pre-tests using Hopkins statistics reveal little clustering structure in the term-document frequency and embedding matrices. In addition, the performance scores of these models are low, and they produce topic assignments that differ substantially from the LLM-based classifications.

**Look-ahead bias.** Since most LLMs are trained on large internet-based corpora, researchers have raised concerns about look-ahead bias (e.g. Glasserman and Lin, 2023; Sarkar and Vafa, 2024). I note that, in this paper, LLMs are used primarily for information extraction and summarization, where look-ahead bias should not be a major concern.

Look-ahead bias remains a concern if the textual embedding vector  $X$  contains forward-looking information. In such cases,  $X$  and the resulting predicted revisions may predict forecast errors simply because they encode information about the unforecastable component, rather than because they reflect behavioral mistakes. It will also be a concern for the results regarding return predictability in Section 5.3. I mitigate this by using the 1999-snapshot of ChronoBERT developed by He et al. (2025) as the embedding model. Trained solely on financial news up to 1999, this model ensures that the embedding vectors are free from look-ahead bias.

**Reproducibility.** A common concern with LLMs is the lack of reproducibility due to the stochastic nature of text generation, where outputs are sampled from a probability distribution over tokens. I ensure deterministic behavior by setting all randomness-controlling parameters (`temperature`, `top_k`, and `top_p`) to 0, so that the model produces identical outputs for a given input. The fact that many closed-source LLMs (e.g., ChatGPT) are continuously updated with

new training data and user interactions means that future users may not obtain the same outputs as the model weights evolve over time (e.g. [Chen et al., 2023b](#)). This concern does not, however, apply to open-source models like Llama 3.1 that are used in this paper, where the researcher retains complete control over the model weights.

**Agreement in topic assignment across LLMs.** LLMs vary in their architecture, training data, and parameterization, which can lead to differences in their outputs for downstream tasks. This raises potential concerns about robustness when relying on a single model for textual analysis. In addition to primarily using Llama-3.1-8b, I address this concern by using 11 other open-source LLMs for topic generation and assignment. These include widely used models from the Llama, DeepSeek, Mistral, and Phi families. Table [OA.1](#) shows that while different models may produce different total-topic numbers, the pairwise cosine similarities between the embedding vectors of topic labels remain consistently high (above 0.94), suggesting strong agreement across models on the topics used to classify the nouns.

I also demonstrate that the topic-classification results are robust to the inherent randomness of LLM output generation and variation across model instances. For the 10,000 most frequently occurring nouns, I use 11 LLMs to classify them into 153 sub-topic labels. Each model is run with `temperature = 1` to allow for some stochastic behavior, and the classification task is repeated 100 times. Figure [OA.12](#) shows a high degree of agreement across models, particularly among more recent, larger models. Final topic assignments are determined by majority vote across the 11 models, further reducing the impact of randomness.

**Completeness and uncertainty in output.** The completeness of the extracted information and topics is ensured by running sanity-check prompts at each key generation step, with the LLM explicitly prompted to verify whether any relevant content is missing. I proceed to the next step only if the LLM confirms that the output is complete. I assess the uncertainty in the LLM outputs and benchmark them against a closed-source model by employing ChatGPT-4o to evaluate the topic classification of the 1,000 most frequently occurring nouns with the “inner probability” method proposed in [Chen et al. \(2025\)](#) to quantify the evaluation uncertainty. Over 93.4% of the topic assignments are deemed correct, with an average inner probability of 98%, indicating a high level of confidence in the LLM-generated classifications.

## 4 Variation in information focus

Having extracted the topics, I proceed with my first set of main empirical results to demonstrate that there is substantial variation in analysts’ information focus. As introduced in Section [2.3](#), I measure analyst  $i$ ’s information focus on topic  $k$  when forecasting firm  $j$ ’s EPS at time  $t$  as the

fraction of topic  $k$ -related nouns within the factual content of the analyst’s report:

$$m_{i,j,t}^k = \frac{\sum_{noun \in A_{i,j,t}^{factual}} \mathbf{I}(noun \in C^k)}{\sum_{noun \in A_{i,j,t}^{factual}} \mathbf{I}}, \quad (12)$$

where  $noun \in A_{i,j,t}^{factual}$  is a noun in the factual content  $A_{i,j,t}^{factual}$  written by analyst  $i$  for firm  $j$  at time  $t$ ,  $\mathbf{I}$  is an indicator function and  $\mathbf{I}(noun \in C^k)$  is an indicator function that the  $noun$  belongs to topic  $k$ . The top panel of Figure A.2 shows the full-sample topic weights for the 28 topics identified by the LLMs. Understandably, analysts devote most of their discussion (24.5% of the total weight) to firm profitability, followed by valuation (6.1%), asset management (5.9%), and cost (5.9%). I then explore how their focus varies along various dimensions.

## 4.1 Time-series variation

Analysts’ information focus exhibits significant variation over time and across business cycles. I construct an aggregate measure of the focus on topic  $k$  as the average of  $m_{i,j,t}^k$  across all analysts and firms at time  $t$ . Figure 3 plots the aggregate focus on several selected topics over time, with additional topics shown in Figure A.3.

There are two significant patterns to note from these time-series plots. First, analysts’ focus on certain topics exhibits strong time trends. For example, analysts have decreased their emphasis on profitability and costs over the past 20 years. The discussions of R&D and business strategies were frequent during the tech bubble, dropping off before steadily increasing again in recent years. Second, analysts shift their information focus markedly during recessions. Panels (c) and (d) of Figure 3 show that during recessions analysts increase their discussions of firms’ financial conditions, the macroeconomy, and risks.

As a formal statistical test of how information focus responds to business cycles, I run a time-series regression of changes in aggregate focus on each topic against changes in log industrial production. The regression results are presented in Figure OA.4. I find strong heterogeneity in the cyclicity of analysts’ information focus across topics. For example, analysts focus on firm-specific intangible information—such as M&A, profitability, corporate leadership, and products—is strongly procyclical. In contrast, analysts tend to focus more on firm-level financial conditions (cash, debt, and funding), as well as on the macroeconomy and risk-related topics during recessions.<sup>30</sup> These patterns are consistent with the idea that, in economic downturns, firms’ future fundamentals become more sensitive to financial distress and aggregate risk than would otherwise be the case. Accordingly, analysts appear to pay more attention to these variables during recessions, in line with the predictions of rational inattention models (Kacperczyk et al., 2016).

---

<sup>30</sup>In Figure OA.5, I find similar results in a panel regression setting with firm- and analyst-fixed effects.



Notably, analysts' countercyclical focus on macroeconomic information aligns with recent findings that firm managers also exhibit countercyclical attention to macroeconomic conditions (Song and Stern, 2021; Flynn and Sastry, 2024).

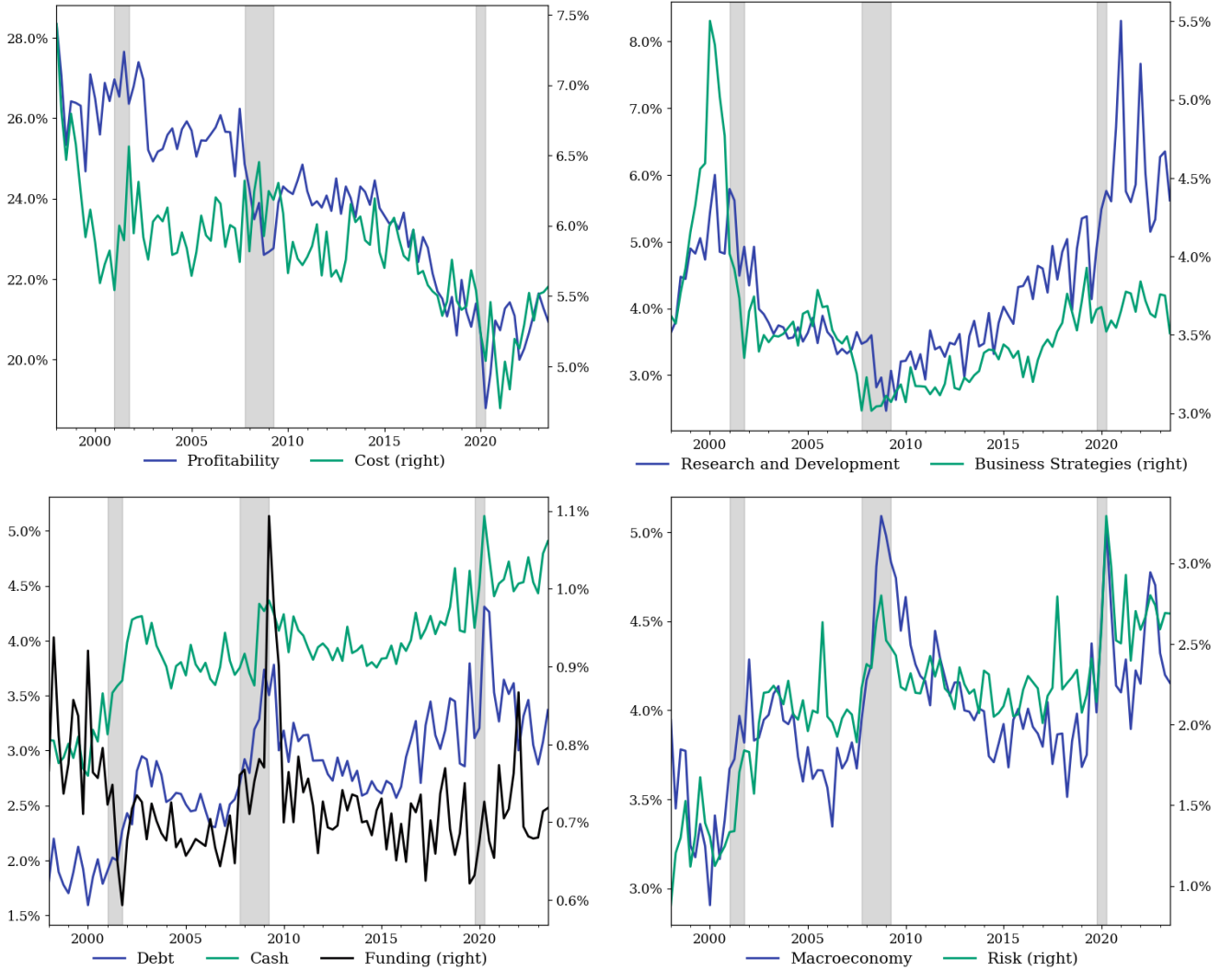


Figure 3: Time-series variation in topic weight

*Notes:* This figure shows the quarterly average weights of selected topics, computed as the average across all analyst reports published in a given quarter. Additional time-series plots are reported in Figure A.3. Shaded region denotes recession periods identified by NBER.

## 4.2 Additional results on variation in topic weights

In Appendix C, I conduct additional analyses on topic-weight variation. In the cross-section, I find that there are substantial variation across stocks even within industries. In particular, analysts tend to focus more on M&A, business strategies, R&D, and product-related topics for growth stocks, but there are no significant differences in the focus on profitability for growth versus value stocks. This suggests that high valuations are primarily supported by information related to



intangibles rather than earnings performance.

I also find that there is significant variation in topic weights depending on the time horizons of analysts' earnings forecasts. When forecasting long-term earnings, analysts focus on a firm's R&D and when they forecast short-term earnings, they tend to focus on a firm's profitability and financial conditions. These differences align with the differences in topic relevance at different horizons.

Finally, I explore differences in attention across analysts covering the same firm, and find that analyst-specific attention choice exists but is limited compared to firm-specific attention choice. I also find that analysts' topic concentration—measured by the Herfindahl-Hirschman Index (HHI) of the 28 topic weights—is lower for analysts with more extensive experience but higher for analysts covering more stocks. This suggests the presence of a processing-capacity constraint for analysts when processing information related to different topics.

### 4.3 Evidence of top-down attention allocation

One way to interpret topic weight  $m_{ijt}^k$  is that it captures the analyst's attention to topic  $k$ . Motivated by this interpretation, in Appendix B.6 I solve for a rational inattention model in terms of which the analyst chooses the focus of their attention to maximize their expected forecast accuracy, but incurs a cost for acquiring a more precise signal. The model predicts that the analyst should allocate more attention to a fundamental component with higher predictive power for future earnings.

Given this prediction, in Appendix D, I explore whether earnings predictability based on a topic increases when it receives greater attention from analysts—a central prediction of a top-down attention-allocation model. I indeed find evidence of such changes in predictability. For example, when analysts talk more about the macroeconomy in their reports, the firm's future earnings growth is indeed more predictable based on current macroeconomic condition. I also find similar patterns using stock characteristics. This result indicates that the topics in analyst reports are not random: analysts actively choose to focus on more important information in their reports.

## 5 Dissecting over- and underreaction to information

In this section, I present my main results regarding analysts' misreactions to different types of information, using the TICG regression introduced in Section 2.4.

As a starting point, I confirm that, consistent with the literature, analysts underreact in their short-term forecasts but overreact in their long-term forecasts by estimating the original CG

regression

$$Error_{i,t}[y_{j,t+h}] = \beta_{CG} \cdot Rev_{i,t}[y_{j,t+h}] + \eta_i + \eta_j + \eta_t + e_{i,j,t}, \quad (13)$$

where  $y_{j,t+h}$  is firm  $j$ 's  $h$ -year-ahead EPS at time  $t$ ,  $Error_{i,t}[y_{j,t+h}]$  is analyst  $i$ 's forecast error with respect to firm  $j$ 's  $h$ -year ahead earnings at time  $t$ ,  $Rev_{i,t}[y_{j,t+h}]$  is the forecast revision, and  $\eta_i, \eta_j, \eta_t$  denotes analyst-, firm-, and time (quarter)-fixed effects in the regression. Following [Bouchaud et al. \(2019\)](#) and [Van Binsbergen et al. \(2023\)](#), I scale both earnings and forecasts by the stock price in the month prior to the forecast revision.<sup>31</sup> As in the original CG-regression framework, a positive  $\beta_{CG}$  indicates that when analysts receive good news and revise their forecasts upward, their new forecasts remain low relative to the actual EPS, implying underreaction. Conversely, a negative  $\beta_{CG}$  indicates an overreaction to information.<sup>32</sup>

	Forecast horizon			
	Short-term (1-2 years)		Long-term (> 3 years)	
	(A)	(B)	(C)	(D)
	Original CG	Total TIGC	Original CG	Total TIGC
$\beta_{CG}$	0.176*	0.265*	-0.113*	-0.146*
	(22.55)	(39.41)	(-6.34)	(-8.14)
Fixed effects	✓	✓	✓	✓
First-stage F-stat		274.5		233.2
Adj. $R^2$ (first-stage)		0.136		0.178
Adj. $R^2$ (second-stage)	0.012	0.004	0.027	0.001
Num Obs.	793462	793462	176285	176285

Table 1: Coibion-Gorodnichenko regression of EPS forecasts

*Notes:* Columns A and C report the estimates from the original CG regression, Eq. (13). Columns B and D report the results of regressing forecast errors on predicted revisions as in Eq. (15). Analyst-, firm-, and quarter-fixed effects are included and  $t$ -statistics are reported in parentheses, with standard errors two-way clustered at the analyst-firm and firm-quarter levels. The first-stage F-statistics are computed following [Mikusheva and Sun \(2022\)](#); \* denotes statistical significance at the 1% level.

Columns (A) and (C) of Table 1 report the results with OLS estimation of the original CG regressions. I find a positive  $\beta_{CG}$  for short-term EPS forecasts, indicating underreaction ([Bouchaud](#)

<sup>31</sup>Outliers beyond the 1% and 99% percentiles are dropped, following [Kelly et al. \(2024\)](#).

<sup>32</sup>Another possible explanation for the negative CG coefficient is that the forecast revisions are in the *opposite* direction of the forecast error. Empirically, I find that about 35% of the long-term forecast revisions are in the wrong direction. Because the agent in many belief formation models is assumed to know the correct direction of revision (e.g. [Bordalo et al., 2020](#); [Kohlhas and Walther, 2021](#); [Augenblick et al., 2025](#)), I follow the tradition and interpret negative CG coefficients as overreactions. However, an alternative interpretation of TIGC coefficients is possible when considering reactions in the wrong direction.

et al., 2019), and a negative  $\beta_{CG}$  for long-term EPS forecasts, indicating overreaction (Bordalo et al., 2019, 2024c).<sup>33</sup>

As discussed in Section 2.2, the original CG regression cannot identify *which* types of information analysts over- or underreact to. In the remainder of this section, I use the new TICG regression to shed light on this issue.

## 5.1 Results from the text-instrumented CG regression

I then implement the TICG regression introduced in Section 2.4. Specifically, for analyst  $i$  issuing a forecast for firm  $j$  at time  $t$  for a  $h$  horizon, I estimate the two-stage regression for each topic  $k$  with an embedding vector  $X_{i,j,t}^k$  using the following specification

$$\begin{aligned} Rev_{i,t}[y_{j,t+h}] &= \sum_{k=1}^{27} \lambda'_k X_{i,j,t}^k + \eta_i + \eta_j + \eta_t + u_{i,j,t} \\ Error_{i,t}[y_{j,t+h}] &= \beta_{TICG}^k \cdot \widehat{Rev}_{i,t}[y_{j,t+h}] + \sum_{k' \neq k} \xi'_{k'} X_{i,j,t}^{k'} + \eta_i + \eta_j + \eta_t + e_{i,j,t} \end{aligned} \quad (14)$$

where  $\widehat{Rev}_{i,t}[y_{j,t+h}] = \sum_{k=1}^{27} \widehat{\lambda}'_k X_{i,j,t}^k$  and  $\eta_i, \eta_j, \eta_t$  denote analyst-, firm-, and time- (quarter-) fixed effects. I estimate the first-stage regression using the jackknife IV estimator studied in Mikusheva and Sun (2022) and Mikusheva and Sun (2024). I drop the “valuation” topic from analyses because its TICG coefficient likely captures the mechanical effect of analysts incorporating their biased forecasts into their subjective valuation, rather than analysts misreacting to information.<sup>34</sup>

### 5.1.1 First-stage results

I start by showing that the first-stage regression is strong in the data, that is, the textual embeddings capture significant variation in  $Rev_{i,t}[y_{j,t+h}]$ . I find the first-stage regression produces an  $R^2$  of 0.136 for short-term forecasts and 0.178 for long-term forecasts (after controlling for analyst-, firm-, and time-fixed effects).

Since the first-stage regression effectively decomposes the forecast revisions into different topics, in Figure 4, I present the variance decomposition of forecast revisions into various topics. I find that for short-term forecasts, information related to profitability has the greatest explanatory power for forecast-revision variation (accounting for 24% of the explained variation and 2.7% of the total variation), followed by information concerning a firm’s costs, inventory, taxes, and debt. In contrast, for long-term earnings forecasts, information about the macroeconomy, international

<sup>33</sup>I show in Tables OA.7 and OA.8 that the results are robust to alternative fixed-effect specifications and to more granular forecast horizons. Table OA.9 reports the results for the full IBES sample, where the pattern of underreaction in short-term forecasts and overreaction in long-term forecasts remains consistent.

<sup>34</sup>The main results are robust when the topic is included.

markets, and R&D plays the important role (each contributing 8.4% of the explained variation and 1.5% of the total variation).

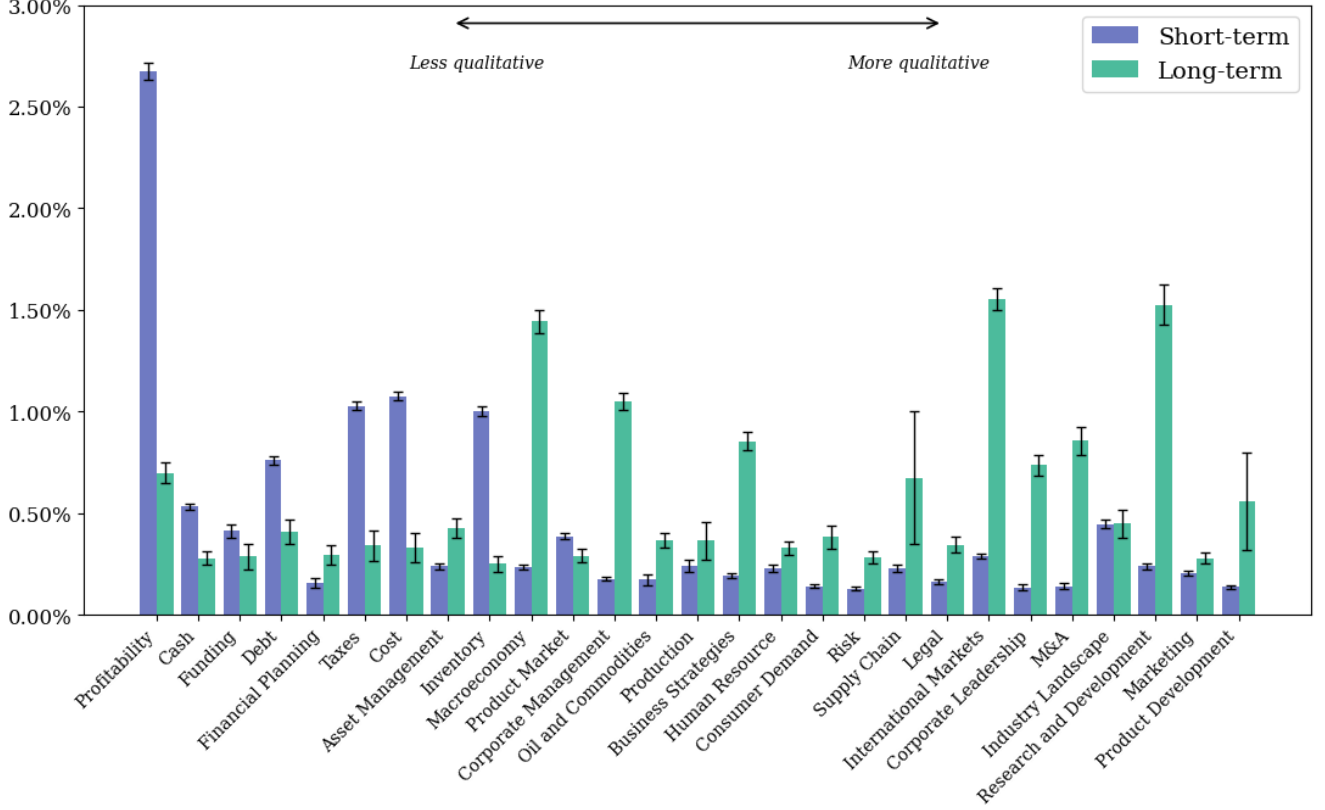


Figure 4: Variance decomposition of forecast revisions into different topics

*Notes:* This figure reports the variance decomposition of forecast errors using embedding vectors of different topics. Specifically, I estimate

$$Rev_{i,t}[y_{j,t+h}] = \sum_{k=1}^{27} \lambda'_k X_{i,j,t}^k + \eta_i + \eta_j + \eta_t + \epsilon_{i,j,t}$$

where  $X_{i,j,t}^k$  denotes the topic-level text embedding for topic  $k$  in report written by analyst  $i$  at time  $t$  covering firm  $j$ , and  $\eta_i, \eta_j, \eta_t$  denote analyst-, firm-, and time (quarter)-fixed effects. Predicted revision from each topic is computed using jackknife estimator. The variance contribution is then computed as  $\frac{Cov(\hat{\lambda}_k X_{i,j,t}^k, Rev_{i,t}[y_{j,t+h}])}{Var(Rev_{i,t}[y_{j,t+h}])}$  for topic  $k$  by regressing  $\hat{\lambda}_k X_{i,j,t}^k$  onto  $Rev_{i,t}[y_{j,t+h}]$ . Error bars represent 95% confidence intervals, with standard errors two-way clustered at the analyst-firm and firm-quarter levels.

These results suggest that tangible, accounting-based information plays a more important role in shaping analysts' short-term expectations, whereas intangible, qualitative topics are more influential in their long-term forecasts. This is also broadly consistent with the differences in information focus in short-term and long-term forecasts shown in Figure OA.8. For example, analysts devote more discussion to profitability when forming short-term forecasts and focus more on R&D when forming long-term forecasts.<sup>35</sup>

<sup>35</sup>In Figure OA.3, I show that there is a positive relationship between a topic's explanatory power for forecast

The strong explanatory power of forecast revisions from topic embeddings implies that these embeddings are strong instruments. Following the procedure in Mikusheva and Sun (2022), I find that when using all 27 topic embeddings, the first-stage F-statistics are 274.5 for short-term forecast revisions and 233.2 for long-term forecast revisions. Figure A.4 reports the topic-level F-statistics. I find that for all but one topic, the corresponding F-statistics comfortably exceed the weak many-instrument threshold of 4.14 proposed in Mikusheva and Sun (2022).

In Figure A.6, I report the correlation of topic-level revisions (correlations of  $\widehat{Rev}_{i,t}[\theta_{j,t+h}] = \widehat{\lambda}'_k X_{i,j,t}$ ). I find that the estimated topic-level revisions exhibit low correlations, with the average absolute pairwise correlation being 0.032 for short-term forecast revisions and 0.024 for long-term forecast revisions. This suggests that the topics capture distinct aspects of information relevant to analysts' forecast revisions.

### 5.1.2 Second-stage results

I now present the main results from the second stage of the TIGC regressions. I first run the “total” TIGC regression by regressing forecast errors on the aggregate predicted revision across all topics,

$$Error_{i,t}[y_{j,t+h}] = \beta_{TIGC} \cdot \widehat{Rev}_{i,t}[y_{j,t+h}] + \epsilon_{i,j,t}, \quad (15)$$

where  $\widehat{Rev}_{i,t}[y_{j,t+h}] = \sum_k \widehat{\lambda}_k X_{i,j,t}^k$  is the predicted forecast revision due all topics. These results are reported in Columns (B) and (D) in Table 1. I find the same statistically significant underreaction in short-term forecast revisions and overreaction in long-term revisions, suggesting that analysts are indeed misreacting to the information contained in the text of analyst reports.<sup>36</sup>

Next, moving on to the topic-level over- and underreaction, Figure 5 presents the TIGC estimates separately for short-and long-term EPS forecasts using (14). It reveals two novel findings. First, the top panel reports the results for short-term forecasts. I find that the underreaction is pervasive across topics: 26 out of the 27 topics have positive TIGC coefficients, and 15 are statistically significant at the 5% level.

---

revisions and the average topic weight in analyst reports. A rational inattention model predicts a one-to-one mapping between analysts' *attention* and *reaction* to different information. The positive relationship I document is directionally consistent with this prediction, but the fit is far from perfect, which indicates that there may be gaps between the extent to which an analyst chooses to pay attention and the extent to which they react to a piece of information.

<sup>36</sup>In Table A.1, I show that the residual in the first-stage regression also predicts forecast errors. This is likely reflecting additional belief misreaction to information not discussed the text in analyst reports.

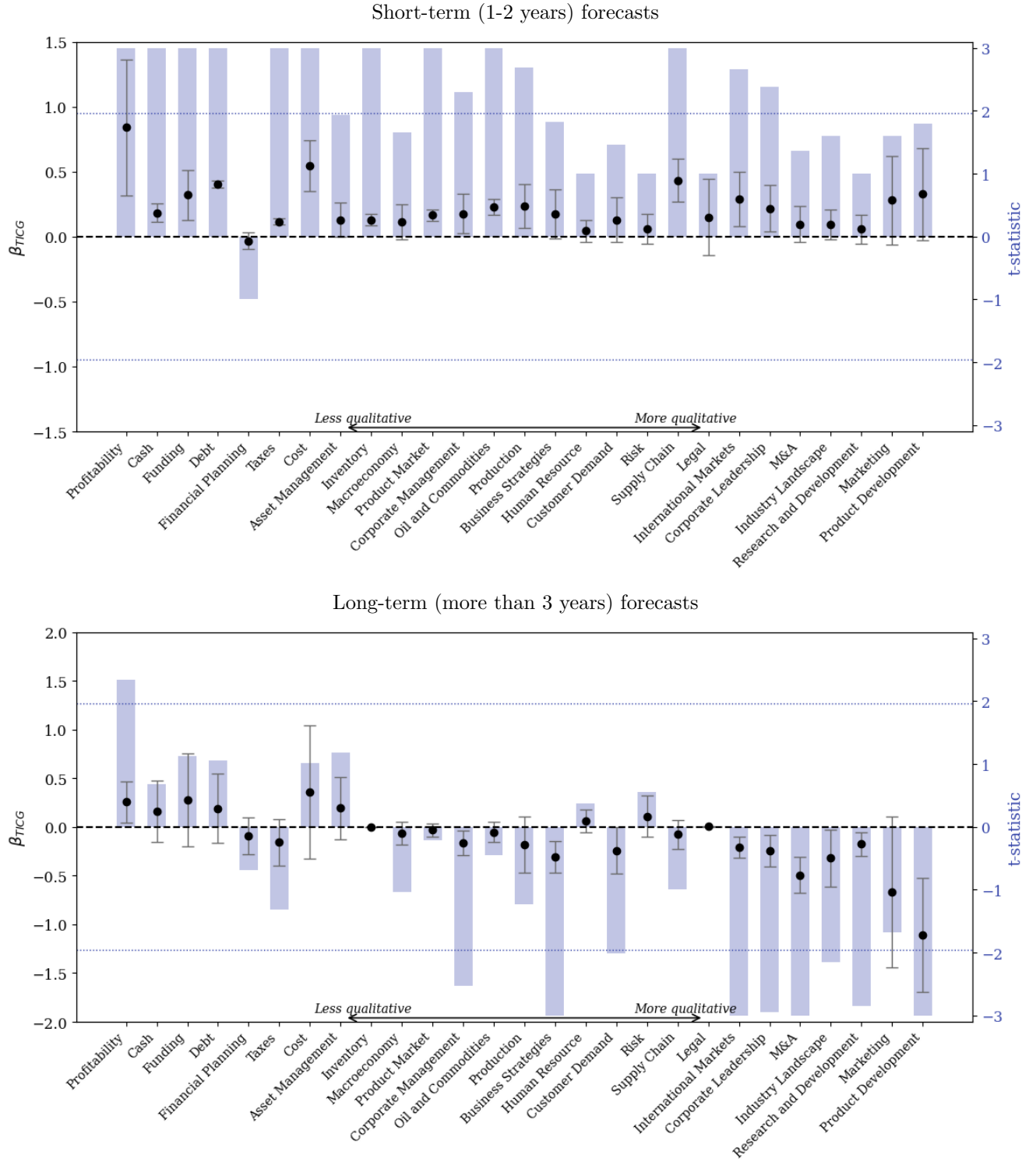


Figure 5: Text-instrumented CG-regression coefficients

*Notes:* This figure shows the TICG-regression coefficient estimates of each topic. The first-stage regression is estimated using a jackknife estimator, and the standard errors are computed using the variance formula in Appendix B.7.2. Analyst-, firm- and time-fixed effects are included. Gray error bars indicate 95% confidence intervals. The blue bars (right axis) show the  $t$ -statistics of each TICG coefficient, and the dashed lines indicate 5% significance levels;  $t$ -statistics with absolute values greater than 3.0 are truncated for visual purpose.

Second, and perhaps more interestingly, the TIGC coefficients for long-term forecasts exhibit much greater heterogeneity. I find that analysts still underreact to profitability information in their long-term forecasts, as evidenced by the positive and significant TIGC coefficient. In contrast, the TIGC coefficients for product development, marketing, M&A, business strategies, industry landscape, and corporate leadership topics are strongly negative. This indicates that the overreaction documented by the original CG regressions is primarily driven by an overreaction to information related to these topics.

In Figure A.5, I perform a “reduced form” forecast error variance decomposition using topic embeddings directly

$$Error_{i,t}[y_{j,t+h}] = \sum_{k=1}^K \xi'_k X_{i,j,t}^k + \eta_i + \eta_j + \eta_t + e_{i,j,t} \quad (16)$$

The results show that the R&D, product development, industry landscape, and M&A topics explain the largest share of variation in forecast errors for long-term earnings forecasts, consistent with the findings from the TIGC regressions.<sup>37</sup>

Taken together, these TIGC results detail how analysts misreact to different types of information. By highlighting which types of information are more susceptible to misreaction, these results offer guidance for theoretical models of earnings overreaction.<sup>38</sup>

## 5.2 Qualitative-quantitative gap in overreaction

Why do analysts overreact more to some topics than others? Although it is difficult to offer a definitive answer, in this section I provide evidence that this pattern may be driven by a differential response to qualitative and quantitative content. Specifically, I document a “qualitative-quantitative” gap in misreaction: analysts tend to overreact more (or underreact less) when processing qualitative, story-like information, and overreact less (or underreact more) when processing quantitative, statistical information.

First, I show that the heterogeneous misreaction patterns documented in the TIGC regressions align with the share of qualitative information tuples within each topic, as reported in the bottom panel of Figure A.2. Specifically, I use the LLM to classify whether a  $\{Noun: Description\}$  tuple

---

<sup>37</sup>This regression also yields higher forecast error predictability compared with the original CG or TIGC regressions, with an  $R^2$  of 0.035 for short-term forecast errors and 0.089 for long-term forecast errors. This finding suggests that, beyond the topic-level over- and underreaction captured by the TIGC regressions, the mappings  $\lambda_k$  and  $\xi_k$  likely point in different directions. Hence, analysts appear to exhibit misreaction even to more granular, within-topic information captured by different embedding dimensions.

<sup>38</sup>In Appendix OA.11, I provide anecdotal evidence of analysts overreacting to news related to the corporate leadership and business strategy of JCPenney, R&D news of General Motors, and to customer-demand news for Twitter.



contains statistical or numerical values.<sup>39</sup> As an example, the tuple (“net margin”, “increased by 6%”) will be classified as quantitative, while the tuple (“the CEO”, “announced product launch timeline”) will be classified as qualitative. I find that for topics such as product development, marketing, the industry landscape, and R&D, more than 80% of the information tuples are classified as qualitative by the LLM. In contrast, less than 40% of the profitability topic’s information tuples are classified as qualitative.

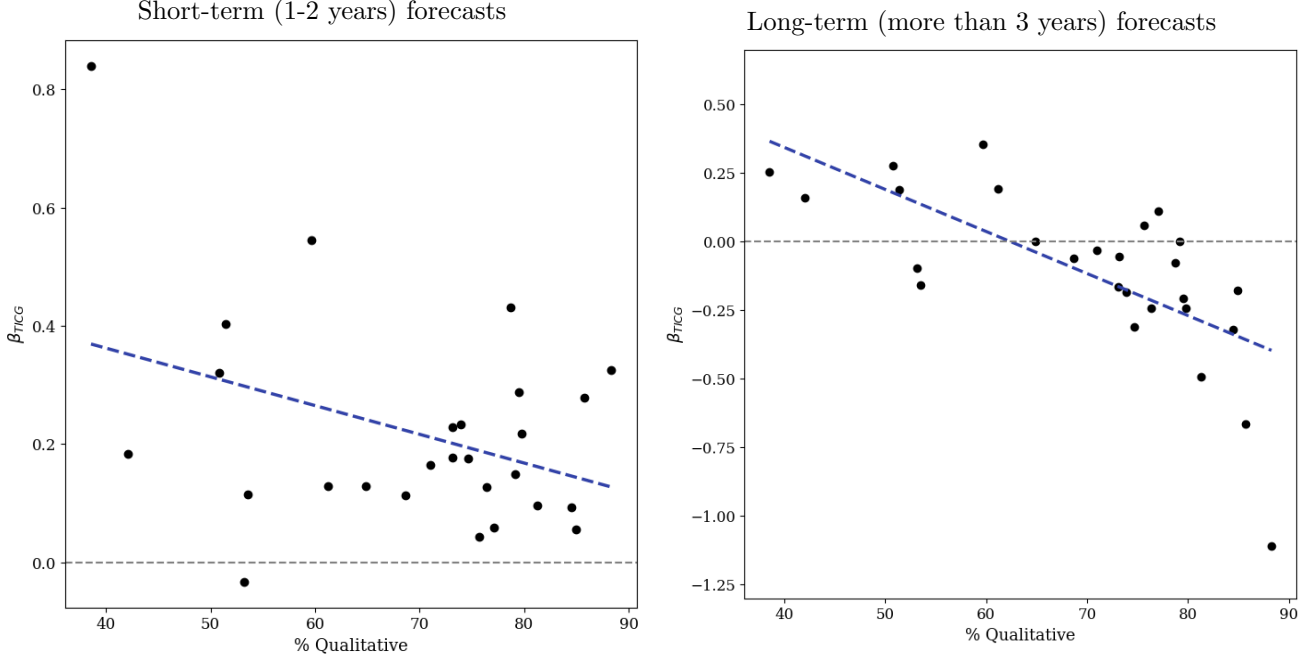


Figure 6: TICG coefficients versus fraction of qualitative information tuples

*Notes:* The scatter plots show the relationship between text-instrumented CG-regression coefficients and topic qualitiveness. Each dot represents one topic, and a positive (negative)  $\beta_{TICG}$  indicates underreaction (overreaction) to that topic. The x-axis shows the fraction of qualitative information tuples in a topic.

Importantly, the differences in qualitative information weights appear to explain the cross-topic variation in TICG coefficients. Figure 6 presents scatter plots of  $\beta_{TICG}$  for different topics against the share of qualitative information in each. I find a strong negative relationship between a topic’s qualitative content and its TICG coefficient.<sup>40</sup> This negative relationship is more pronounced in the long-term forecast setting, where the topics that are the most qualitative in nature are those primarily responsible for the negative coefficient in the original CG regression. A similar pattern is also observable in short-term forecasts, but there is another underreaction that causes TICG coefficients to become more positive across all topics.<sup>41</sup>

<sup>39</sup>The prompt for determining whether an information tuple is qualitative is Prompt #10 in Appendix OA.5.

<sup>40</sup>Figure A.9 depicts analogous scatter plots using the coefficients from the topic-weight-interacted CG regressions and a similarly strong negative relationship between topic qualitiveness and CG coefficients, conditional on the topic weights.

<sup>41</sup>The difference between short- and long-term forecasts is not the result of analysts paying more or less attention

### 5.2.1 Overreaction is primarily driven by qualitative information

As a more direct test of the impact of the qualitativenss of information on overreaction in long-term forecasts, I construct a measure of long-term forecast revision in response to qualitative (quantitative) information  $Rev_{i,t}^{qual}[y_{j,LT}]$  ( $Rev_{i,t}^{quant}[y_{j,LT}]$ ) as the sum of predicted revision associated with topics with high (low) shares of qualitative information tuples,

$$\widehat{Rev}_{i,t}^{qual}[y_{j,LT}] := \sum_{k \in \mathcal{K}_{qual}} \widehat{Rev}_{i,t}[y_{j,LT}]^k; \quad \widehat{Rev}_{i,t}^{quant}[y_{j,LT}] := \sum_{k \in \mathcal{K}_{quant}} \widehat{Rev}_{i,t}[y_{j,LT}]^k, \quad (17)$$

where  $\mathcal{K}_{qual}$  ( $\mathcal{K}_{quant}$ ) denotes the top (bottom) half of the list of topics ranked by qualitative tuple share. Given these two explicit measures, it is possible to directly examine how overreaction in long-term forecasts is driven by revisions with respect to qualitative and quantitative information. I do so by running the regression

$$y_{j,t+h} - F_{i,t}[y_{j,t+h}] = \beta_{TICG}^{qual} \cdot \widehat{Rev}_{i,t}^{qual}[y_{j,t+h}] + \beta_{TICG}^{quant} \cdot \widehat{Rev}_{i,t}^{quant}[y_{j,t+h}] + \epsilon_{i,j,t}. \quad (18)$$

Intuitively, this regression examines the total TICG regression, Eq. (15), and whether the overreaction in long-term forecasts is driven by a revision in response to qualitative or quantitative information. Table 2 shows the results. Column 1 reproduces the total TICG regression result in Table 1, which indicates an overreaction in analysts' long-term earnings forecasts. This overreaction is assessed in Columns 2 to 4, which show that it is mainly driven by analysts' revisions with respect to qualitative information  $\widehat{Rev}_{i,t}^{qual}[y_{j,t+h}]$ , rather than their revision in response to quantitative information. Such overreaction to qualitative information is both statistically and economically significant. My estimates indicate that 25.7% of the variation in long-term earnings forecast revisions *due to qualitative information* reflects excessive adjustment (overshooting the targets), whereas only 2.0% of the variation *due to quantitative information* represents misreaction.

Finally, I directly examine the original CG-regression coefficient conditional on the fraction of qualitative information in the analyst's report. Specifically, I estimate the following regression:

$$y_{j,t+h} - F_{i,t}[y_{j,t+h}] = \beta_{CG} \cdot Rev_{i,t}[y_{j,t+h}] + \beta_{CG}^{qual} \cdot Rev_{i,t}[y_{j,t+h}] \times Qual\_Frac_{i,j,t} + \epsilon_{i,j,t}. \quad (19)$$

The coefficient  $\beta_{CG}^{qual}$  thus captures how misreaction varies with the degree to which analysts are exposed to information presented in qualitative form. The negative  $\beta_{CG}^{qual}$  in the last column in Table 2 implies that the more analysts are exposed to information presented in qualitative form, the more negative their CG coefficient, indicating greater overreaction.

---

to qualitative than quantitative information. The average share of qualitative tuples is 0.63 for short-term forecasts and 0.65 for long-term forecasts. This suggests that the results are indeed capturing differences in *misreaction*, not differences in *attention*.

Forecast horizon: Long-term ( $> 3$ Years)					
Dependent variable: $y_{j,t+h} - F_{i,t}[y_{j,t+h}]$					
$\widehat{Rev}_{i,t}[y_{j,t+h}]$	-0.070*				
	(-8.14)				
$\widehat{Rev}_{i,t}^{qual}[y_{j,t+h}]$		-0.257*		-0.257*	
		(-2.70)		(-2.69)	
$\widehat{Rev}_{i,t}^{quant}[y_{j,t+h}]$			-0.022	-0.020	
			(-0.36)	(-0.33)	
Revision					0.219*
					(2.50)
Revision $\times$ Qual_Frac					-0.496*
					(-3.87)
Fixed effects	✓	✓	✓	✓	✓
Adj. $R^2$	0.000	0.005	0.000	0.005	0.032
Num obs.	176285	176285	176285	176285	176285

Table 2: Impact of qualitative information on overreaction

*Notes:* This table reports the results of forecast-error prediction regression, Eqs. (18) and (19) for long-term forecasts. The instrumented forecast revision  $\widehat{Rev}_{i,t}[y_{j,t+h}]$  is defined in Eq. (15), and  $\widehat{Rev}_{i,t}^{qual}[y_{j,t+h}]$  ( $\widehat{Rev}_{i,t}^{quant}[y_{j,t+h}]$ ) are instrumented forecast revisions using qualitative (quantitative) topics defined in Eq. (17). Analyst-, firm-, and quarter-fixed effects are included. Standard errors are two-way clustered at the analyst-firm and firm-quarter levels,  $t$ -statistics are reported in parentheses, and \* denote statistical significance at the 1% levels.

### 5.3 Asset pricing implication of overreaction to qualitative information

Thus far I have demonstrated that the overreaction in analysts' long-term earnings forecasts is mainly driven by their overreaction to qualitative information. In this section, I show that this overreaction also strongly affects stock returns.

**Cross-sectional return predictability.** Recent work by [Bordalo et al. \(2024a,c\)](#) has established that analysts' revisions to LTG forecasts predict long-term returns with a negative sign. In a similar spirit, I run a Fama-MacBeth regression of future returns on analysts' forecast revisions

(20)

$r_{j,t \rightarrow t+h} = \beta_1 Rev_t[y_{j,LT}] + \beta_2 \widehat{Rev}_t^{qual}[y_{j,LT}] + \beta_3 \widehat{Rev}_t^{quant}[y_{j,LT}] + \epsilon_{j,t}$ , where  $r_{j,t \rightarrow t+h}$  is firm  $j$ 's (annualized) cumulative return from year  $t$  to year  $t+h$ ,  $Rev_t[y_{j,LT}]$  is the stock-level numerical long-term EPS revision made by analysts in year  $t$ , and  $\widehat{Rev}_t^{qual}[y_{j,LT}]$  and  $\widehat{Rev}_t^{quant}[y_{j,LT}]$  are stock-level

long-term forecast revisions driven by qualitative and quantitative information following the construction in Eq. (17).

	Dependent variable: Future stock returns								
	$r_{j,t+1}$			$r_{j,t+2}$			$r_{j,t+3}$		
$Rev_t[y_{j,LT}]$	-0.007 (-0.33)	0.001 (0.06)	-0.003 (-0.14)	-0.016 (-0.74)	-0.013 (-0.54)	-0.014 (-0.56)	-0.009 (-0.49)	-0.005 (-0.26)	-0.009 (-0.43)
$\widehat{Rev}_t[y_{j,LT}]$	-0.061*** (-2.53)			-0.034* (-1.86)			-0.030** (-2.25)		
$\widehat{Rev}_t^{qual}[y_{j,LT}]$	-0.081*** (-4.62)			-0.077*** (-6.66)			-0.062*** (-4.09)		
$\widehat{Rev}_t^{quant}[y_{j,LT}]$	-0.028 (-1.28)			-0.006 (-0.37)			-0.006 (-0.48)		
Avg. $R^2$	0.008	0.014	0.018	0.012	0.018	0.023	0.007	0.012	0.018
Nobs	16831	16831	16831	16828	16828	16828	16723	16723	16723

Table 3: Fama-MacBeth regression of future stock returns on forecast revisions

*Notes:* This table reports results of the Fama-MacBeth regression of future 1- to 3-year stock return on long-term forecast revisions. Stock-level long-term forecast revisions are the 3-5 years ahead EPS forecast revisions averaged across analysts in a given year.  $Rev_t[y_{j,LT}]$  denotes numerical forecast revisions,  $\widehat{Rev}_t[y_{j,LT}]$  denotes text-instrumented forecast revisions, and  $\widehat{Rev}_t^{qual}[y_{j,LT}]$  ( $\widehat{Rev}_t^{quant}[y_{j,LT}]$ ) denotes text-instrumented forecast revisions in response to qualitative (quantitative) topics.  $t$ -statistics are reported in parentheses with Driscoll–Kraay standard errors using optimal bandwidth and \*, \*\*, and \*\*\* denote statistical significance at 10%, 5% and 1% level.

Table 3 shows the result for predictions of 1- to 3-year future returns. I find, first, that analysts’ numerical forecast revisions do not have strong return predictability, but the instrumented forecast revisions (as defined in Eq. (15)) predict future returns strongly with a negative sign. This indicates that  $\widehat{Rev}_t[y_{j,LT}]$  better captures genuine belief revision and has a more profound impact on stock prices. Second, I find that this strong predictability is entirely driven by the revisions in response to qualitative information, whereas revisions driven by quantitative topics do not predict future returns. This result confirms that the overreaction to qualitative information has a significant impact on stock prices.

In Table A.3, I show that the return predictability is robust to controlling for a set of well-known return predictors, including market capitalization, book-to-market ratio, past 12-month return, asset growth, and profitability growth.<sup>42</sup> In addition, Table A.4 shows that forecast revisions driven by qualitative information have an effect on contemporaneous stock returns that is twice as large as the effect of forecast revisions of the same magnitude, which suggests that the predictability reflects excess price movements and is consistent with belief overreaction.

<sup>42</sup>These stock characteristics are obtained from Jensen et al. (2023).

**Betting against qualitative revisions.** Given the strong negative return predictability of qualitative revisions in long-term forecasts, a trading strategy can be constructed to exploit the associated mispricing. I do so by estimating a TICG regression at the end of each year  $t$  starting from 2004 using previous long-term forecasts and textual embeddings, and construct a long-short portfolio by going long on the stocks with the lowest 20% of  $\widehat{Rev}_t^{qual}[y_{j,LT}]$  and shorting those with the highest 20% of  $\widehat{Rev}_t^{qual}[y_{j,LT}]$  in year  $t$ .<sup>43</sup> The black line in Figure 7 shows that betting against analysts’ qualitative revision delivers significant investment value. In comparison, similar long-short portfolios based on analysts’ numerical forecast revisions or revisions due to quantitative topics do not perform as well. I report summary statistics of these long-short portfolios in Table A.5, which shows that the strategy based on qualitative revisions achieves an annualized Sharpe ratio of 0.85 with significant alphas relative to the other two revision strategies and prominent factor models.

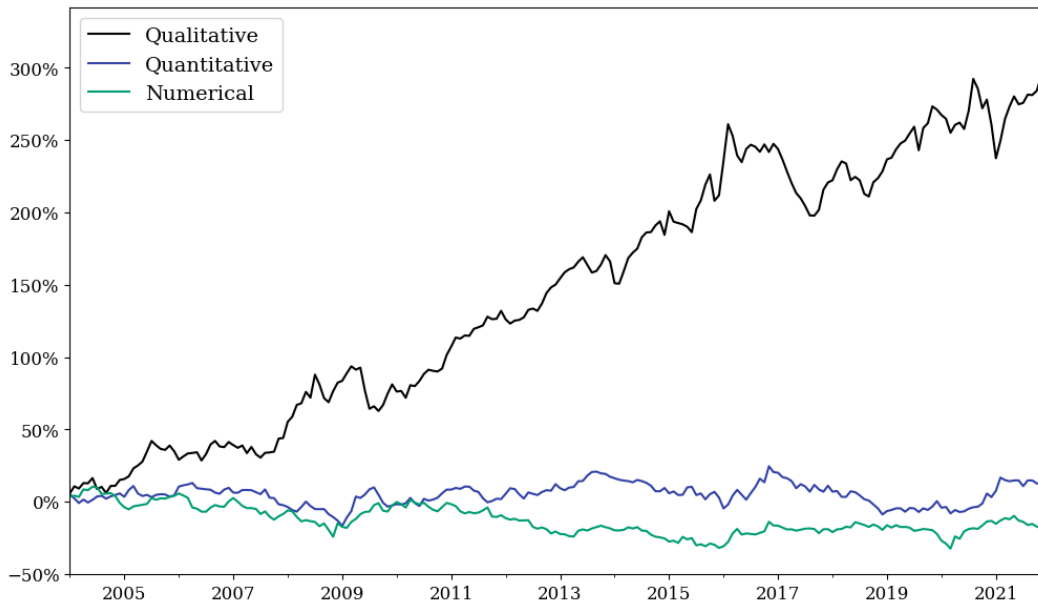


Figure 7: Long-short portfolio cumulative returns based on forecast revisions

*Notes:* This plot shows the long-short quintile-portfolio cumulative returns by sorting stocks based on predicted long-term revisions in response to qualitative and quantitative information, and numerical long-term revisions from IBES in a given year. Each portfolio is scaled to have an ex-post annual volatility of 10%.

**Explaining behavioral factors.** Finally, the long-short portfolio based on qualitative forecast revisions also helps explain the mispricings previously identified in the asset pricing literature. For example, Daniel et al. (2020) construct two behavioral factors that capture mispricing at different horizons and use these to explain the cross-section of expected returns. In Table A.6, I find that the long-term behavioral factor—which exploits persistent mispricing related to share

<sup>43</sup>The expanding window estimation of TICG regressions and the use of chronologically consistent embedding model (ChronoBERT-v1999) ensures that the result is truly out of sample.

issuance—can be spanned by the long-short portfolio based on qualitative revisions. This suggests that such persistent mispricing can be explained by overreactions to qualitative information.

## 5.4 Robustness tests and additional results

**Robustness to embedding dimensions.** In my baseline implementation, I use the first 100 principal components of the embedding vectors for each *Noun* and *Description* in information tuples to construct a 200-dimension embedding representation of the information related to topic  $k$ . Figure A.7 shows that the TICG-regression results are robust when using the first 50, 80, and 120 principal components in the embedding representations.

**Forecast persistence and learning from price.** In some cases, researchers might include the lagged forecast at  $t - 1$  in the CG regression to control for the persistence in belief distortions.<sup>44</sup> Table A.2 finds that controlling for the lagged forecast does not change the negative coefficient on  $\widehat{Rev}_{i,t}^{qual}[y_{j,t+h}]$ , which indicates that such overreaction to qualitative information is not just capturing persistence in belief distortions.

Chaudhry (2023) suggests that the overreaction in earnings forecasts might be related to analysts learning from stock price movements. The observed overreaction to qualitative information might reflect analysts using qualitative information to justify their forecast revisions learned from stock price movements. To investigate this, in Table A.2 I control for past 1-year, 3-year, and 5-year stock returns, and I find that the overreaction to qualitative information is not subsumed by these controls of price movements. This suggests that the overreaction is not due to analysts using qualitative information to justify their learning from price.

**Qualitative-quantitative gap within topic.** In Figure 6, I show that analysts overreact differently depending on the qualitateness of the topic. As a robustness check, I also investigate whether analysts overreact more to qualitative information *within* a topic than to quantitative information by running the following regression

$$y_{j,t+h} - F_{i,t}[y_{j,t+h}] = \sum_k \beta_k \widehat{Rev}_{i,t}^k[y_{j,t+h}] + \sum_k \beta_k^{qual} \widehat{Rev}_{i,t}^k[y_{j,t+h}] \times Qual\_Frac_{i,j,t}^k + \epsilon_{i,j,t}.$$

Effectively, this regression studies how the topic-level TICG coefficient changes when the topic contains more qualitative information tuples in the report. Figure A.10 shows that  $\beta_k^{qual}$  is mostly negative, indicating that analysts also overreact more to qualitative information within a topic.

---

<sup>44</sup>Such forecast error predictability naturally arises in models such as the diagnostic expectation of persistent variables (Bordalo et al. (2024c)), which Halperin and Mazlish (2025) describes as “over-extremity.”

**Topic-weight-interacted CG regression.** As an alternative gauge of how analysts misreact to different topics, I estimate a “topic-weight-interacted” CG regression. Specifically, I interact  $Rev_{i,t}[y_{j,t+h}]$  with topic weights in the underlying analyst reports. While this regression does not allow the structural interpretation afforded by the TICG regression, it provides an alternative description of how the original CG-regression coefficient changes as analysts’ topic focus changes. Figure A.8 reports the results. I find that the overreaction in analysts’ long-term forecasts is more pronounced when they focus on topics like product development, marketing, and R&D. Figure A.9 shows that there is also a clear negative relationship between the qualitative nature of a topic and its impact on the original CG-regression coefficient. This again aligns with the TICG-regression results.

**Alternative construction of revision to qualitative information.** In my baseline construction of  $\widehat{Rev}_{i,t}^{qual}[y_{j,t+h}]$  and  $\widehat{Rev}_{i,t}^{quant}[y_{j,t+h}]$ , I exploit the full sample’s cross-topic qualitative heterogeneity, ignoring variation in information quality at the report level. I construct alternative measures of  $\widehat{Rev}_{i,t}^{qual}[y_{j,t+h}]$  and  $\widehat{Rev}_{i,t}^{quant}[y_{j,t+h}]$  by aggregating topic-level revisions with the topic-level quality at the report level. Table A.7 shows that the results in Table 2 are robust to the definition of revisions with respect to qualitative and quantitative information.

**Connection with the overreaction in intangible returns.** The overreaction to qualitative information is also connected to the fact that a stock’s return is strongly and negatively related to its “intangible” return, the component of the past return that is orthogonal to the firm’s past performance (Daniel and Titman, 2006; Jiang, 2010). I conduct two exercises to show that qualitative and intangible information are closely connected. First, I compute the tangible and intangible returns, and calculate their variances in bins based on weight of qualitative information tuples at the stock-year level.<sup>45</sup>

Figure A.11 shows that intangible returns are more volatile when analysts discuss qualitative information in their reports, while there is little change in the volatility of tangible returns. Second, I regress tangible and intangible returns on long-term  $\widehat{Rev}_{j,t}^{qual}$  and  $\widehat{Rev}_{j,t}^{quant}$  (computed as the average across analysts for each stock-year). Table A.8 shows that the revisions due to qualitative information are more strongly related to intangible returns than revisions due to quantitative information, and vice versa for tangible returns. These results suggest that the documented overreaction to intangible returns is likely driven by the overreaction to qualitative information.

<sup>45</sup>Following Daniel and Titman (2006) and Jiang (2010), I compute tangible returns as the book returns (changes in log book equity) from year  $t - 1$  to year  $t$ ,  $r_{j,t}^{tangible} = \log(BE_{j,t}) - \log(BE_{j,t-1})$ . The intangible return is computed as the residual in the cross-sectional regression  $r_{j,t} = \gamma_0 + \gamma_1 \log(BE/ME_{j,t-1}) + \gamma_2 r_{i,t}^{tangible} + v_{j,t}$  and  $r_{j,t}^{intangible} = \widehat{v}_{j,t}$ . I aggregate  $Qual\_Frac$ ,  $\widehat{Rev}_{i,t}^{qual}[y_{j,t+h}]$  and  $\widehat{Rev}_{i,t}^{quant}[y_{j,t+h}]$  to the stock-year level. Variances are computed after controlling for stock- and year-fixed effects.



**Sample splitting based on regulation change.** The regulation landscape with respect to sell-side research has changed since the 2000s, mainly to address conflicts of interest and institutional incentives distorting the content of analyst reports.<sup>46</sup> To study whether the overreaction to qualitative information may be driven by institutional incentives, I redo the plots in Figure 6 for the samples before and after 2016, which corresponds to the start of FINRA 2241, a major regulation covering sell-side analysts. Figure A.12 shows that the qualitative nature of analyst reports is quite stable over time, and that the finding of a greater overreaction to qualitative topics in long-term forecasts is robust before and after 2016. The result is thus less likely to be driven by conflicts of interest or strategic communications.

**Institutional incentives.** A common concern in the literature regarding sell-side analysts' forecasts is that rather than reflecting the analysts' true beliefs, these may be distorted by institutional incentives. As explained in Section 2.5, such institutional incentives will bias the estimates of the true-belief effects.

Quantifying how institutional incentives distort analysts' forecasts is difficult and beyond the scope of this paper. I nevertheless show that these incentives are unlikely to be the reason analysts overreact more to qualitative than other information. Following the literature on sell-side brokerage incentives (Lin and McNichols, 1998; Michaely and Womack, 1999; Ljungqvist et al., 2006, 2007; Malmendier and Shanthikumar, 2014), I compute several measures of institutional features, including issuance and M&A-deal affiliations, pressure from investment banks, brokers' reputational capital, broker royalty index, brokerage size, and institutional ownership.<sup>47</sup> Figure A.14 shows that the negative relationship documented in Figure 6 remains robust across these subsamples, indicating that institutional distortions play a minimal role in the results.

**Heterogeneity across analysts.** I also explore differences across analysts in their degree of overreaction to qualitative information. Figure A.15 presents scatter plots depicting the relationship between TICG coefficients and topic qualitateness across subsamples, defined by analyst characteristics. The negative relationship documented in Figure 6 remains robust across these subsamples. However, the relationship weakens for analysts with more experience and those at larger brokerage firms. This suggests that analysts learn to correct belief distortions over time as they become more skilled in the profession.

**Focus on qualitative information.** Given that the overreaction to qualitative information is stronger for long-term forecasts, one might wonder whether this is because analysts focus more on qualitative information when making these. While Figure OA.8 shows that there are indeed

---

<sup>46</sup>Noticeable changes in US regulation include Reg FD in 2000, NASD 2711 and NYSE 472 in 2002, Regulation AC in 2003, and FINRA 2241 in 2015.

<sup>47</sup>The computation of these variables is detailed in Appendix E.

differences in topic weights when analysts make short-term versus long-term forecasts, I find no significant differences in the qualitative versus quantitative information focus of these. This suggests that the result is driven by misreaction to information, rather than information selection.

**Taking stock.** In this section, I have studied what information analysts misreact to in their earnings forecasts and found that there is substantial heterogeneity in the degree of misreaction across topics. I document two novel patterns: (1) analysts overreact more to qualitative, story-like information than to quantitative information, and (2) there is another underreactive force in the short-term earnings forecasts that operates across topics.

## 6 Uncovering behavioral mechanisms using textual data

The textual data capturing analysts’ beliefs facilitates the documentation of novel empirical patterns—such as heterogeneous overreactions to qualitative versus quantitative information—and provides a new empirical environment for testing established theories of belief formation. Researchers can search in natural language for textual evidence of any proposed mechanism. Assuming that the text genuinely reflects analysts’ process of belief formation, how analysts articulate and justify their beliefs provides a way to directly identify the behavioral forces underlying anomalies in observed beliefs.

In this section, I employ LLMs to find textual evidence of mechanisms that can help explain two facts documented in Section 5: (1) an overreaction concentrated around qualitative information in long-term forecasts, and (2) a separate underreaction that affects all topics in short-term forecasts. I focus on three candidate mechanisms for overreaction (overconfidence, memory effects, and overinference from weak signals) and two candidate mechanisms for underreaction (sticky beliefs and confirmation bias, and herding). These mechanisms are chosen because they are particularly amenable to study using textual data; what follows is by no means a complete evaluation of all possible behavioral mechanisms.<sup>48</sup> Nevertheless, this section offers a first step in this direction, and demonstrates that the self-disclosed text has the potential to advance our understanding of the behavioral mechanisms underlying misreaction.

Based on textual evidence, I find that the observed overreaction to qualitative information is consistent with the operation of overconfidence as a mechanism: analysts’ subjective statements tend to take a stronger tone and use more definitive language—for example “we are very confident,” “undoubtedly,” or “definitely”—when they encounter qualitative information. This pattern

---

<sup>48</sup>Other behavioral mechanisms that can generate overreaction include representative heuristics (Kahneman and Tversky, 1972; Bordalo et al., 2019), base-rate neglect (Kahneman and Tversky, 1973), and incorrect belief in the law of small numbers (Rabin, 2002; Jin and Peng, 2023). Other mechanisms that can generate underreaction include inattention (Sims, 2003) and gradual information diffusion (Hong and Stein, 1999).

suggests that an analyst’s overreaction to intangible signals stems primarily from their overconfidence when interpreting qualitative information. Regarding underreaction in short-term forecasts, I present suggestive textual evidence that this is driven by herding behavior among analysts.

## 6.1 What explains the qualitative-quantitative gap in overreactions in long-term forecasts?

**Overconfidence.** The overconfidence hypothesis posits that individuals systematically overestimate the precision of their own judgments (Barberis, 2018).<sup>49</sup> Overconfidence has been widely invoked to explain excess trading volume and persistent disagreement in financial markets (e.g. Barber and Odean, 2000; Grinblatt and Keloharju, 2001). Importantly, it also predicts that agents will overreact to information because their subjective precision is higher than the true precision of the signals (e.g. Daniel et al., 1998; Broer and Kohlhas, 2024; Adam et al., 2025; Li et al., 2025). In particular, analysts may exhibit greater overconfidence when processing qualitative than quantitative information, as the interpretive nature of the former leaves more room for subjective judgment. This flexibility in interpretation can lead to more strongly held personal convictions and, consequently, more overconfident behavior (Odean, 1998). Recent work by Filipovic and Wagner (2024) shows that managers tend to be overconfident in the value of M&A deals when they are discussing information about intangibles, which leads to lower abnormal returns after deal announcements.

I seek evidence of potential overconfidence by hypothesizing that overconfident analysts are more likely to use a stronger tone or more assertive language in their reports, particularly in their *subjective* statements. Based on this idea, I use an LLM to identify whether assertive language is present in the subjective sentences in each report. Specifically, I classify each chunk of subjective content as either containing assertive language (coded as 1) or not (coded as 0), using Prompt #12 in Appendix OA.5. I then construct a measure  $h^{confidence}$  for each report, defined as the fraction of subjective chunks that include assertive language. I present examples of language identified as confident in Appendix OA.12.1.

**Memory.** Recent studies highlight the important role of memory in belief formation (Bordalo et al., 2021, 2023; Wachter and Kahana, 2024; Jiang et al., 2024; Enke et al., 2024; Kwon and Tang, 2025). Specifically, new information may trigger analysts to recall similar past events or episodes, leading them to overweight the probability of outcomes implied by the new information. This memory channel may also help explain analysts’ tendency to overreact to qualitative infor-

---

<sup>49</sup>This form of overconfidence is typically referred to as “overprecision.” Another variant, “dismissiveness”, describes the tendency of individuals to discount the judgments or forecasts of others, including consensus views. However, this latter form of overconfidence seems inconsistent with the well-documented herding behavior among analysts.

mation. Recent work by Graeber et al. (2024) shows that qualitative content is more likely to elicit associative recall than statistical content.

In light of the above, I seek evidence of memory or associative recall by prompting the LLM to identify references to past events or episodes in analysts’ reports (using Prompt #13 in Appendix OA.5). To distinguish genuine memory recall from descriptions of recent developments, I instruct the LLM to exclude current events and those in the recent past. Specifically, each chunk in an analyst report takes the value of 1 if the LLM detects a reference to a prior event or episode, and 0 otherwise. I then construct a report-level measure,  $h^{memory}$ , defined as the fraction of chunks in which the LLM identifies such references. I present examples of identified memory language in Appendix OA.12.2, which shows that the LLM correctly identifies when analysts are actively recalling information from the past.

**Overinference of weak signals.** Another potential mechanism for overreaction is seen when people are unsure of the true strength of the signal and update their beliefs as if the signal had “intermediate strength.” This implies that they will overinfer from (and subsequently overreact to) weak signals, while at the same time underreacting to strong signals (Khaw et al., 2021; Ba et al., 2024; Augenblick et al., 2025). Along similar lines, people may overreact to transitory signals and underreact to persistent signals if they perceive those signals to have “intermediate persistence” (Wang, 2021; Afrouzi et al., 2023).

While this mechanism for overreaction is somewhat similar to the overconfidence mechanism, a key distinction is that the former predicts that the degree of overreaction will depend on the strength of the signals. Under this hypothesis, analysts overreact more to qualitative than quantitative information because the former type is a weaker predictor of long-term earnings. Evaluating the earnings predictability across topics thus provides a suggestive test of this mechanism. I explore this using textual embeddings of each topic,  $X^k$ , to predict realized growth in future earnings and examine whether the variation in predictive power aligns with the pattern of misreaction.

### 6.1.1 Empirical results

I begin by examining whether the presence of more qualitative information is associated with greater overconfidence or memory recall by running the following regression:

$$h_{i,j,t}^o = \beta \cdot Qual\_Frac_{i,j,t} + \epsilon_{i,j,t}, \quad (21)$$

where  $h^o \in \{h^{overconfidence}, h^{memory}\}$  and  $Qual\_Frac$  represents the share of information tuples in the analyst’s report classified as qualitative.

	$h^{overconfidence}$	$h^{memory}$
Qual_Frac	0.089* (25.32)	0.042* (17.02)
Fixed effects	✓	✓
Adj. $R^2$	0.534	0.553
Num obs.	391641	391641

Table 4: Impact of qualitative information on  $h^{overconfidence}$  and  $h^{memory}$

*Notes:* This table reports the impact of the qualitiveness of information on measures of overconfidence and memory effects. Qual\_Frac is the fraction of information tuples classified as qualitative in the analyst’s report,  $h^{overconfidence}$  is the fraction of chunks of subjective statements containing assertive language identified by the LLM, and  $h^{memory}$  is the fraction of factual chunks that describes historical episodes as identified by the LLM. Analyst-, stock-, and time (quarter)-fixed effects are included.  $t$ -statistics are reported in parentheses, with standard errors being two-way clustered at the analyst-firm and firm-quarter level; \* denotes statistical significance at the 1% level.

In Table 4, I find that analysts who include more qualitative content in their writing also use more assertive language (higher  $h^{overconfidence}$ ) and are more likely to reference past events (higher  $h^{memory}$ ) than those who do not. These findings suggest that both mechanisms could explain the observed overreaction to qualitative information.

Next, I examine how overconfidence and memory influence the overreaction to qualitative information by interacting terms between forecast revisions due to qualitative information  $\widehat{Rev}_{i,t}^{qual}[y_{j,t+h}]$  (as defined in Eq. (17)) with  $h^{overconfidence}$  and  $h^{memory}$ :

$$\begin{aligned}
y_{j,t+h} - F_{i,t}[y_{j,t+h}] = & \beta_1 \cdot \widehat{Rev}_{i,t}^{qual}[y_{j,t+h}] + \beta_2 \cdot \widehat{Rev}_{i,t}^{qual}[y_{j,t+h}] \times h_{i,j,t}^{overconfidence} \\
& + \beta_3 \cdot \widehat{Rev}_{i,t}^{qual}[y_{j,t+h}] \times h_{i,j,t}^{memory} + \epsilon_{i,j,t}.
\end{aligned} \tag{22}$$

In this regression, the coefficients  $\beta_2$  and  $\beta_3$  capture how the presence of overconfident or memory-related language in analysts’ reports influences how they overreact to qualitative information.

Forecast horizon: Long-term (>3 years)						
Dependent variable: $y_{j,t+h} - F_{i,t}[y_{j,t+h}]$						
$\widehat{Rev}_{i,t}^{qual}[y_{j,t+h}]$	-0.034*	-0.050*	0.175*			
	(-6.39)	(-8.63)	(-5.91)			
$\widehat{Rev}_{i,t}^{qual}[y_{j,t+h}] \times h^{overconfidence}$	-0.072*		-0.075*			
	(-3.08)		(-3.13)			
$\widehat{Rev}_{i,t}^{qual}[y_{j,t+h}] \times h^{memory}$		0.006	0.022			
		(0.18)	(0.63)			
<i>Revision</i>				0.037	-0.053*	0.007
				(0.37)	(-2.45)	(0.30)
<i>Revision</i> $\times$ $h^{overconfidence}$				-0.534*		-0.537*
				(-5.71)		(-5.63)
<i>Revision</i> $\times$ $h^{memory}$					-0.158	0.015
					(-1.41)	(0.13)
Fixed effects	✓	✓	✓	✓	✓	✓
Adj. $R^2$	0.009	0.008	0.009	0.025	0.025	0.025

Table 5: Impact of overconfidence and memory on overreaction

*Notes:* This table reports how  $h^{overconfidence}$  and  $h^{memory}$  affect overreaction. The first three columns show the results for revisions related to qualitative information, Eq. (22), and the last three columns show the results for total variation, Eq. (23).  $h^{overconfidence}$  is the share of chunks of subjective statements that the LLM identified as containing assertive language and  $h^{memory}$  is the share of factual chunks identified as describing historical episodes. Analyst-, firm- and quarter-fixed effects are included.  $t$ -statistics are reported in parentheses, with standard errors being two-way clustered at the analyst-firm and firm-quarter level; \* denotes statistical significance at 1% level.

The results, presented in Columns 1 to 3 of Table 5, show that the interaction term  $\widehat{Rev}_{i,t}^{qual}[y_{j,t+h}] \times h^{overconfidence}$  has a negative predictive coefficient for forecast errors. This suggests that analysts' overreaction to qualitative information is greater when they use more assertive language in their reports. In contrast, there is no significant effect from the interaction with  $h^{memory}$ .

In a similar fashion, I investigate how  $h^{overconfidence}$  and  $h^{memory}$  affect the overall negativity in the original CG regressions by interacting the original forecast revisions with  $h^{overconfidence}$  and  $h^{memory}$ ,

$$\begin{aligned}
y_{j,t+h} - F_{i,t}[y_{j,t+h}] = & \beta_1 \cdot Rev_{i,t}[y_{j,t+h}] + \beta_2 \cdot Rev_{i,t}[y_{j,t+h}] \times h_{i,j,t}^{overconfidence} \\
& + \beta_3 \cdot Rev_{i,t}[y_{j,t+h}] \times h_{i,j,t}^{memory} + \epsilon_{i,j,t}.
\end{aligned} \tag{23}$$

In this regression, a negative  $\beta_2$  indicates that analysts overreact more (the original CG regression is more negative) when analysts exhibit greater overconfidence, and a negative  $\beta_3$  indicates that analysts overreact more when they engage more in recalling past episodes. The results in the last

three columns of Table 5 show that the more confident the language used by analysts, the more they overreact, but this is not the case when they reference past events more. These results again suggest that overconfidence is an important driver of the observed overreaction.

### 6.1.2 Additional results on overconfidence

In addition to the direct evidence related to the predictability of forecast errors, I examine how overconfidence and memory effects relate to analysts' experience. The first two columns of Table A.13 report the coefficients for the regression of  $h^{overconfidence}$  and  $h^{memory}$  on analyst experience, defined as the number of quarters since the analyst first issued a forecast in the IBES dataset. I find that the overconfidence measure is negatively associated with experience, indicating that less experienced analysts are more likely to use assertive language and be identified as overconfident. In contrast, the memory measure is positively associated with experience, suggesting that the more experienced an analyst the more likely they are to reference historical events in their reports. These findings support the interpretation that the overconfidence measure captures belief distortions, while memory-related language may reflect accumulated knowledge that contributes to more precise, rather than more biased, forecasts.

Second, in Table A.9, I show that when interacting  $h^{overconfidence}$  with revisions related to quantitative information  $\widehat{Rev}_{i,t}^{quant}[y_{j,t+h}]$ , there are no significant effects. This suggests that it is analysts' overconfidence when interpreting qualitative information that drives the results.

Given the finding that analysts' overreaction is greater when  $h^{overconfidence}$  is high, that is, when analysts use more assertive and definitive language, one might be concerned that this does not constitute genuine overconfidence, but rather describes a larger movement in their beliefs. This concern is mitigated by the finding in Table A.10 that the coefficient of the regression of  $h^{overconfidence}$  on absolute forecast revision is not statistically significant. This suggests that more extensive revisions do not mechanically cause analysts to use more definitive language.<sup>50</sup> I also repeat regression, Eq. (23) but split  $h_{i,j,t}^{overconfidence}$  into its lagged value  $h_{i,j,t-1}^{overconfidence}$  and its change  $\Delta h_{i,j,t}^{overconfidence} = h_{i,j,t}^{overconfidence} - h_{i,j,t-1}^{overconfidence}$ . If the CG coefficient is negative because, mechanically,  $h^{overconfidence}$  is high when the current revision is extensive, then we would expect no effect from the lagged overconfidence measure. Table A.11 shows, however, that both changes and the lagged overconfidence measure predict more negative CG coefficients.

**Overinference from weak signals.** Finally, I examine whether overinference from weak signals could be the underlying behavioral mechanism driving overreaction. The mechanism's central prediction is that qualitative information tends to have weaker predictive value than quantitative information for future fundamentals but analysts overestimate its precision. I empirically assess

---

<sup>50</sup>Intuitively, analysts can say things like they are "very confident" that the future EPS "will not be impacted by current events" and not change their forecasts drastically.



this prediction by estimating the following predictive regression:

$$\Delta y_{j,t+h} = \beta'_k X_{j,t}^k + \epsilon_{j,t+h}, \quad (24)$$

where  $\Delta y_{j,t+h}$  denotes the  $h$ -year EPS growth, and  $X_{j,t}^k$  represents the average embedding vector of information tuples related to topic  $k$  for firm  $j$  at time  $t$ , constructed as the average across all analysts' reports published in that quarter. Figure A.16 displays the scatter plots of the relationship between topic-level predictive  $R^2$  and the fraction of qualitative tuples within each topic. I find a positive relationship: topics with a higher share of qualitative content tend to have greater predictive power for future earnings growth. This pattern contradicts the overinference hypothesis, which would predict lower forecast accuracy for qualitative (i.e., weaker) signals.

**Summary of overreaction mechanisms.** Taken together, a greater share of qualitative information is associated with both higher levels of overconfidence and memory recall. However, the overconfidence measure seems to be associated with a more negative CG coefficient in long-term forecasts. This suggests that overconfidence is an important driver of the overreaction to qualitative information in long-term forecasts. I find limited support for memory effects and overinference as the underlying mechanisms for overreaction.

## 6.2 What explains the overall underreaction in short-term forecasts?

**Sticky beliefs.** Bouchaud et al. (2019) attribute the positive CG coefficient to belief stickiness. In their model, subjective beliefs are formed as a convex combination of the rational belief and the agent's prior belief from the preceding period. This stickiness component, which anchors current to past beliefs, generates forecast underreaction. In the literature, sticky beliefs have been attributed to psychological features like confirmation bias (Rabin and Schrag, 1999; Pouget et al., 2017) and conservatism (Barberis et al., 1998).

Since belief stickiness can be understood as an overweighting of past forecasts, I hypothesize that the reports of analysts with stickier beliefs are more likely to reference their own prior earnings forecasts. On this basis, I prompt the LLM to identify whether an analyst refers to their previous EPS forecast in a report (using Prompt #14 in Appendix OA.5). Based on this classification, I then construct a binary measure,  $h^{sticky}$ , which is equal to 1 if the LLM detects a reference to the analyst's prior forecast, and 0 otherwise. In Appendix OA.12.3, I provide snippets of analysts' mentions of their past beliefs, and in many cases, the analysts state that the current news confirms their prior beliefs.

**Herdling.** There is substantial empirical evidence that sell-side analysts tend to herd their forecasts and recommendations around the consensus, often disregarding their privately held infor-

mation (Trueman, 1994; Welch, 2000; Clement and Tse, 2005; Jegadeesh and Kim, 2010). Such herding behavior may arise from career or reputational concerns, as modeled by Scharfstein and Stein (1990), or be related to information cascades (Bikhchandani et al., 1992). More recent studies, including Valchev and Gemmi (2023) and Banerjee et al. (2025), reexamine herding within the framework of overreaction and underreaction, particularly in the context of CG regressions, showing theoretically that herding can lead to a positive CG coefficient, indicating underreaction to information.

In herding models, agents place excessive emphasis on the consensus forecast. I capture this behavior empirically by prompting the LLM to identify whether an analyst refers to the consensus forecast within their report, as indicated by the use of such terms as “street estimates”, “consensus forecasts”, or “other analysts” (using Prompt #15 in Appendix OA.5). Based on this classification, I construct a measure,  $h^{herd}$ , for each report that takes the value of 1 if the LLM detects a reference to the consensus forecast, and 0 otherwise.

### 6.2.1 Empirical results

I examine whether the LLM-based measures of herding and sticky beliefs are related to the overall underreaction in analysts’ short-term forecasts by interacting  $h^{herding}$  and  $h^{sticky}$  with the predicted revision  $\widehat{Rev}_{i,t}[y_{j,t+h}]$  and the original revision in the CG regression. The results are reported in Table 6. I find that for both the original and text-instrumented revisions, the coefficient on the interaction with  $h^{herding}$  is positive and significant, indicating that analysts tend to underreact more when their reports explicitly reference consensus forecasts or street estimates. This supports the interpretation that herding is an important mechanism for underreaction in short-term forecasts. In contrast, there is limited evidence that analysts’ references to their own past forecasts makes them underreact more.

**Additional results on herding.** I further investigate whether herding contributes to the positive CG coefficients following the classification approach in Clement and Tse (2005). I define an EPS forecast as “herding” if the new forecast lies between the analyst’s previous forecast and the consensus forecast from the prior month; otherwise, it is classified as “bold.” I then estimate a CG regression that interacts the forecast revision with a herding indicator. Table A.14 shows that CG regressions for forecast revisions classified as “herding” exhibit more positive coefficients, indicating a greater degree of underreaction.

Forecast horizon: Short-term (1-2 years)						
Dependent variable: $y_{j,t+h} - F_{i,t}[y_{j,t+h}]$						
$\widehat{Rev}_{i,t}[y_{j,t+h}]$	0.086*	0.092*	0.085*			
	(7.04)	(10.96)	(6.91)			
$\widehat{Rev}_{i,t}[y_{j,t+h}] \times h^{herding}$	0.075*		0.065*			
	(7.30)		(5.87)			
$\widehat{Rev}_{i,t}[y_{j,t+h}] \times h^{sticky}$		0.061*	0.017			
		(4.60)	(1.12)			
<i>Revision</i>				0.167*	0.191*	0.190*
				(15.59)	(14.51)	(14.37)
<i>Revision</i> $\times h^{herding}$				0.023		0.075*
				(1.55)		(4.08)
<i>Revision</i> $\times h^{sticky}$					-0.026	-0.077*
					(-1.62)	(-3.84)
Fixed effects	✓	✓	✓	✓	✓	✓
Adj. $R^2$	0.004	0.004	0.004	0.012	0.012	0.012

Table 6: Impact of herding and sticky beliefs on underreaction

*Notes:* This table reports how  $h^{herding}$  and  $h^{sticky}$  affect underreaction. The first three columns show the results for instrumented revision, Eq. (22), and the last three columns show the results for original revision, Eq. (23).  $h^{herding}$  is a dummy that is equal to 1 if the analyst's report mentions consensus or street estimates, and  $h^{sticky}$  is a dummy that is equal to 1 if the analyst's report mentions their previous forecasts.  $t$ -statistics are reported in parentheses, with standard errors being two-way clustered at the analyst-firm and firm-quarter level; \* denotes statistical significance at 1% level.

A natural question arises: Why is the underreaction pronounced in short-term but not long-term forecasts? One explanation lies in the difference in the number of analysts issuing forecasts across horizons. Since a larger number of analysts issue short-term than long-term forecasts, the herding effect may be stronger at shorter horizons.<sup>51</sup> Evidence that the number of total forecasts may influence the extent of herding is provided in Table A.12, which shows that CG coefficients are more positive when there were more analysts issuing forecasts in the previous month. This suggests that the number of forecasters is indeed an important determinant of the degree of herding.<sup>52</sup>

### 6.3 A belief-formation model motivated by textual evidence

In Appendix B.5 I present a stylized model that is motivated by the empirical facts documented above. Specifically, the model features (1) an additional term in analysts' objective function

<sup>51</sup>On average, there are 6.62 analysts issuing 1-year-ahead EPS forecasts for a firm in a given month, compared to 6.35 for 2-year-ahead forecasts, 3.85 for 3-year, 2.13 for 4-year, and 1.91 for 5-year-ahead forecasts.

<sup>52</sup>See also Diebold et al. (2025) for a study of how forecast accuracy depends on the number of forecasters.

that penalizes their forecasts for being too different from the consensus, and (2) a higher level of perceived precision for qualitative signals, capturing analysts’ overconfidence in the interpretation of qualitative signals. The model solution, as stated in Proposition 4, demonstrates that, with certain parameter restrictions, it can simultaneously account for the more pronounced overreaction (i.e., more negative TICG coefficients) to qualitative topics relative to quantitative ones, the overall overreaction in long-term forecasts (negative original CG coefficient), and the overall underreaction in short-term forecasts (positive original CG coefficient) because there are more analysts issuing short-term forecasts, making the herding mechanism stronger.

## 7 Conclusion

In this paper, I study how financial analysts form beliefs by analyzing their words. I apply LLMs to extract interpretable information from the text of over 1.1 million analyst reports and link it to analysts’ earnings forecasts. This approach allows me to directly observe and identify the information on which analysts focus and how they react to this.

I find that analysts’ attention varies across time, firms, and forecast horizons in ways consistent with top-down attention allocation. I then develop a TICG to uncover over- or underreaction to specific types of information. I document evidence for the novel finding that analysts tend to overreact to qualitative, intangible information but not to quantitative, statistical information. I analyze the language used in the reports to understand these misreactions. I find that overreaction to qualitative topics is strongly linked to overconfidence, while underreaction in the short term is primarily driven by herding behavior.

This paper demonstrates that information from unstructured data can provide novel insights into the behavior of economic agents. Using generative AI, I establish a precedent for combining textual data with numerical survey responses to study important open questions in behavioral economics. Sophisticated information-extraction techniques offer a promising avenue for future research into the belief and decision-making processes of various economic agents. For example, future research could use the text from earnings-call transcripts to study the beliefs of corporate managers and mutual fund managers from their written annual reports (Jha et al., 2024), or to examine narratives in the speeches of FOMC members to understand their monetary policy decisions (Laarits et al., 2025). Future research could also use insights generated from these unstructured data to further discipline theories of belief formation, thereby assisting researchers in navigating the “wilderness” of alternative models of expectation formation (Angeletos et al., 2021).

## References

- Adam, K., Kuang, P., and Xie, S. (2025). Overconfidence in private information explains biases in professional forecasts. *Journal of Monetary Economics*.
- Adam, K. and Nagel, S. (2023). Expectations data in asset pricing. In *Handbook of economic expectations*, pages 477–506. Elsevier.
- Afrouzi, H., Kwon, S. Y., Landier, A., Ma, Y., and Thesmar, D. (2023). Overreaction in expectations: Evidence and theory. *The Quarterly Journal of Economics*, 138(3):1713–1764.
- Andre, P., Haaland, I., Roth, C., Wiederholt, M., and Wohlfart, J. (2024). Narratives about the macroeconomy. Technical report, SAFE Working Paper.
- Andre, P., Schirmer, P., and Wohlfart, J. (2023). Mental models of the stock market. *Univ. of Copenhagen Dept. of Economics Discussion Paper*, (07/23).
- Andrei, D., Cujean, J., and Wilson, M. (2023). The lost capital asset pricing model. *Review of Economic Studies*, 90(6):2703–2762.
- Angeletos, G.-M., Huo, Z., and Sastry, K. A. (2021). Imperfect macroeconomic expectations: Evidence and theory. *NBER Macroeconomics Annual*, 35(1):1–86.
- Angrist, J. D., Imbens, G. W., and Krueger, A. B. (1999). Jackknife instrumental variables estimation. *Journal of Applied Econometrics*, 14(1):57–67.
- Angrist, J. D. and Krueger, A. B. (1995). Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics*, 13(2):225–235.
- Asquith, P., Mikhail, M. B., and Au, A. S. (2005). Information content of equity analyst reports. *Journal of financial economics*, 75(2):245–282.
- Augenblick, N., Lazarus, E., and Thaler, M. (2025). Overinference from weak signals and underinference from strong signals. *The Quarterly Journal of Economics*, 140(1):335–401.
- Ba, C., Bohren, J. A., and Imas, A. (2024). Over-and underreaction to information. *Available at SSRN 4274617*.
- Bali, T. G., Kelly, B. T., Mörke, M., and Rahman, J. (2023). Machine forecast disagreement. Technical report, National Bureau of Economic Research.
- Banerjee, S., Davis, J., and Gondhi, N. (2025). Motivated beliefs in coordination games. *Available at SSRN 3662734*.

- Barber, B. M. and Odean, T. (2000). Trading is hazardous to your wealth: The common stock investment performance of individual investors. *The journal of Finance*, 55(2):773–806.
- Barberis, N. (2018). Psychology-based models of asset prices and trading volume. In *Handbook of behavioral economics: applications and foundations 1*, volume 1, pages 79–175. Elsevier.
- Barberis, N., Shleifer, A., and Vishny, R. (1998). A model of investor sentiment. *Journal of financial economics*, 49(3):307–343.
- Bastianello, F. (2022). Time-series and cross-section of risk premia expectations: A bottom-up approach. *Available at SSRN 4204968*.
- Bastianello, F., H Décaire, P., and Guenzel, M. (2024). Mental models and financial forecasts. *Marius, Mental Models and Financial Forecasts (October 30, 2024)*.
- Battaglia, L., Christensen, T., Hansen, S., and Sacher, S. (2025). Inference for regression with variables generated by ai or machine learning. *SSRN Working Paper*.
- Bellstam, G., Bhagat, S., and Cookson, J. A. (2021). A text-based analysis of corporate innovation. *Management Science*, 67(7):4004–4031.
- Ben-David, I. and Chincio, A. (2024). Expected eps x trailing p/e. *Working Paper*.
- Ben-David, I., Graham, J. R., and Harvey, C. R. (2013). Managerial miscalibration. *The Quarterly journal of economics*, 128(4):1547–1584.
- Bhandari, A., Borovička, J., and Ho, P. (2025). Survey data and subjective beliefs in business cycle models. *Review of Economic Studies*, 92(3):1375–1437.
- Bianchi, F., Ilut, C., and Saijo, H. (2024a). Diagnostic business cycles. *Review of Economic Studies*, 91(1):129–162.
- Bianchi, F., Ludvigson, S. C., and Ma, S. (2022). Belief distortions and macroeconomic fluctuations. *American Economic Review*, 112(7):2269–2315.
- Bianchi, F., Ludvigson, S. C., and Ma, S. (2024b). What hundreds of economic news events say about belief overreaction in the stock market. Technical report, National Bureau of Economic Research.
- Bikhchandani, S., Hirshleifer, D., and Welch, I. (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy*, 100(5):992–1026.
- Binetti, A., Nuzzi, F., and Stantcheva, S. (2024). People’s understanding of inflation. *Journal of Monetary Economics*, 148:103652.

- Bini, P., Cong, L. W., Huang, X., and Jin, L. J. (2025). Behavioral economics of ai: Llm biases and corrections. *Available at SSRN 5213130*.
- Bordalo, P., Coffman, K., Gennaioli, N., Schwerter, F., and Shleifer, A. (2021). Memory and representativeness. *Psychological Review*, 128(1):71.
- Bordalo, P., Conlon, J. J., Gennaioli, N., Kwon, S. Y., and Shleifer, A. (2023). Memory and probability. *The Quarterly Journal of Economics*, 138(1):265–311.
- Bordalo, P., Gennaioli, N., La Porta, R., and Shleifer, A. (2024a). Finance without exotic risk. Technical report, National Bureau of Economic Research.
- Bordalo, P., Gennaioli, N., Ma, Y., and Shleifer, A. (2020). Overreaction in macroeconomic expectations. *American Economic Review*, 110(9):2748–2782.
- Bordalo, P., Gennaioli, N., Porta, R. L., O’Brien, M., and Shleifer, A. (2024b). Long-term expectations and aggregate fluctuations. *NBER Macroeconomics Annual*, 38(1):311–347.
- Bordalo, P., Gennaioli, N., Porta, R. L., and Shleifer, A. (2019). Diagnostic expectations and stock returns. *The Journal of Finance*, 74(6):2839–2874.
- Bordalo, P., Gennaioli, N., Porta, R. L., and Shleifer, A. (2024c). Belief overreaction and stock market puzzles. *Journal of Political Economy*, 132(5):1450–1484.
- Bouchaud, J.-P., Krueger, P., Landier, A., and Thesmar, D. (2019). Sticky expectations and the profitability anomaly. *The Journal of Finance*, 74(2):639–674.
- Bradshaw, M. T. (2011). Analysts’ forecasts: what do we know after decades of work? *Available at SSRN 1880339*.
- Broer, T. and Kohlhas, A. N. (2024). Forecaster (mis-) behavior. *Review of Economics and Statistics*, 106(5):1334–1351.
- Bybee, J. L. (2023). The ghost in the machine: Generating beliefs with large language models. *arXiv preprint arXiv:2305.02823*.
- Bybee, L., Kelly, B., Manela, A., and Xiu, D. (2024). Business news and business cycles. *The Journal of Finance*, 79(5):3105–3147.
- Cao, S., Jiang, W., Wang, J., and Yang, B. (2024). From man vs. machine to man+ machine: The art and ai of stock analyses. *Journal of Financial Economics*, 160:103910.
- Cassella, S., Golez, B., Gulen, H., and Kelly, P. (2023). Horizon bias and the term structure of equity returns. *The Review of Financial Studies*, 36(3):1253–1288.



- Chao, J. C., Swanson, N. R., Hausman, J. A., Newey, W. K., and Woutersen, T. (2012). Asymptotic distribution of jive in a heteroskedastic iv regression with many instruments. *Econometric Theory*, 28(1):42–86.
- Charles, C. and Sui, P. (2024). Marketwide memory. *Available at SSRN 4872332*.
- Chaudhry, A. (2023). The impact of prices on analyst cash flow expectations: Reconciling subjective beliefs data with rational discount rate variation. *Available at SSRN 4443349*.
- Chen, H., Didisheim, A., and Somoza, L. (2025). Out of the (black) box: Ai as conditional probability. *Available at SSRN*.
- Chen, H., Hwang, B.-H., and Peng, Z. (2023a). Why do investors like short-leg securities? evidence from a textual analysis of buy recommendations. *Nanyang Business School Research Paper*, (23-02).
- Chen, L., Da, Z., and Zhao, X. (2013). What drives stock price movements? *The Review of Financial Studies*, 26(4):841–876.
- Chen, L., Zaharia, M., and Zou, J. (2023b). How is chatgpt’s behavior changing over time? *arXiv preprint arXiv:2307.09009*.
- Chen, Y., Kelly, B. T., and Xiu, D. (2022). Expected returns and large language models. *Available at SSRN 4416687*.
- Chen, Z. (2025). *Memory and beliefs in financial markets: A machine learning approach*. PhD thesis, University of Pennsylvania.
- Chi, F., Hwang, B.-H., and Zheng, Y. (2024). The use and usefulness of big data in finance: Evidence from financial analysts. *Management Science*.
- Chinco, A., Hartzmark, S. M., and Sussman, A. B. (2022). A new test of risk factor relevance. *The Journal of Finance*, 77(4):2183–2238.
- Choi, J. J. (2022). Popular personal financial advice versus the professors. *Journal of Economic Perspectives*, 36(4):167–192.
- Clement, M. B. and Tse, S. Y. (2005). Financial analyst characteristics and herding behavior in forecasting. *The Journal of finance*, 60(1):307–341.
- Coibion, O. and Gorodnichenko, Y. (2015). Information rigidity and the expectations formation process: A simple framework and new facts. *American Economic Review*, 105(8):2644–2678.
- Cookson, J. A., Lu, R., Mullins, W., and Niessner, M. (2024). The social signal. *Journal of Financial Economics*, 158:103870.

- Cookson, J. A. and Niessner, M. (2020). Why don't we agree? evidence from a social network of investors. *The Journal of Finance*, 75(1):173–228.
- Couts, S. J., Gonçalves, A. S., and Loudis, J. (2023). The subjective risk and return expectations of institutional investors. *Fisher College of Business Working Paper*, 14.
- Couts, S. J., S Gonçalves, A., Liu, Y., and Loudis, J. (2024). Institutional investors' subjective risk premia: Time variation and disagreement. *Fisher College of Business Working Paper*.
- Cui, T., Delao, R., and Myers, S. (2024). The subjective belief factor. *Available at SSRN 5205569*.
- Dahlquist, M. and Ibert, M. (2024). Equity return expectations and portfolios: Evidence from large asset managers. *The Review of Financial Studies*, 37(6):1887–1928.
- Daniel, K., Hirshleifer, D., and Subrahmanyam, A. (1998). Investor psychology and security market under-and overreactions. *the Journal of Finance*, 53(6):1839–1885.
- Daniel, K., Hirshleifer, D., and Sun, L. (2020). Short-and long-horizon behavioral factors. *The review of financial studies*, 33(4):1673–1736.
- Daniel, K. and Titman, S. (2006). Market reactions to tangible and intangible information. *The Journal of Finance*, 61(4):1605–1643.
- d'Arienzo, D. (2021). Maturity increasing overreaction and bond market puzzles. *Available at SSRN 3733056*.
- De Silva, T., Larsen-Hallock, E., Rej, A., and Thesmar, D. (2025). Expectations formation with fat-tailed processes: Evidence and theory. *SSRN Working Paper*.
- De Silva, T. and Thesmar, D. (2024). Noise in expectations: Evidence from analyst forecasts. *The Review of Financial Studies*, 37(5):1494–1537.
- Decaire, P. and Graham, J. (2024). Valuation fundamentals. *Working Paper*.
- Decaire, P. and Guenzel, M. (2023). What drives very long-run cash flow expectations? *Working Paper*.
- Delao, R. and Myers, S. (2021). Subjective cash flow and discount rate expectations. *The Journal of Finance*, 76(3):1339–1387.
- Delao, R. and Myers, S. (2024). Which subjective expectations explain asset prices? *The Review of Financial Studies*, 37(6):1929–1978.
- Derrien, F., Krueger, P., Landier, A., and Yao, T. (2025). Esg news, future cash flows, and firm value. *The Journal of finance*, (Forthcoming).

- Diebold, F. X., Mora, A., and Shin, M. (2025). On the wisdom of crowds (of economists). *arXiv preprint arXiv:2503.09287*.
- 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(5):2113–2141.
- Enke, B., Schwerter, F., and Zimmermann, F. (2024). Associative memory, beliefs and market interactions. *Journal of Financial Economics*, 157:103853.
- Faralli, M. (2024). What drives beliefs about climate risks? evidence from financial analysts. *Working Paper*.
- Farmer, L. E., Nakamura, E., and Steinsson, J. (2024). Learning about the long run. *Journal of Political Economy*, 132(10):3334–3377.
- Fedyk, A., Kakhbod, A., Li, P., and Malmendier, U. (2024). Ai and perception biases in investments: An experimental study. *Available at SSRN 4787249*.
- Filipovic, Z. M. and Wagner, A. F. (2024). The intangibles song in takeover announcements: Good tempo, hollow tune. *Swiss Finance Institute Research Paper*, (19-03).
- Flake, J. (2023). *Why Do Managers Interact with Unfavorable Analysts during Earnings Calls?* PhD thesis, Boston College. Carroll School of Management.
- Flynn, J. P. and Sastry, K. (2024). Attention cycles. Technical report, National Bureau of Economic Research.
- Freyaldenhoven, S., Ke, S., Li, D., and Montiel Olea, J. L. (2025). On the testability of the anchor-words assumption in topic models. *FRB of Philadelphia Working Paper*.
- Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–574.
- Gervais, S. and Odean, T. (2001). Learning to be overconfident. *The review of financial studies*, 14(1):1–27.
- Giglio, S., Maggiori, M., Rillo, J., Stroebel, J., Utkus, S., and Xu, X. (2025). Investor beliefs and expectation formation. *Annual Review of Financial Economics*.
- Giglio, S., Maggiori, M., Stroebel, J., and Utkus, S. (2021). Five facts about beliefs and portfolios. *American Economic Review*, 111(5):1481–1522.
- Glasserman, P. and Lin, C. (2023). Assessing look-ahead bias in stock return predictions generated by gpt sentiment analysis. *arXiv preprint arXiv:2309.17322*.

- Goetzmann, W. N., Kim, D., and Shiller, R. J. (2022). Crash narratives. Technical report, National Bureau of Economic Research.
- Gormsen, N. J. and Huber, K. (2024). Firms’ perceived cost of capital. Technical report, National Bureau of Economic Research.
- Gormsen, N. J. and Huber, K. (2025). Corporate discount rates. *American Economic Review*, 115(6):2001–2049.
- Graeber, T., Roth, C., and Zimmermann, F. (2024). Stories, statistics, and memory. *The Quarterly Journal of Economics*, 139(4):2181–2225.
- Greenwood, R. and Shleifer, A. (2014). Expectations of returns and expected returns. *The Review of Financial Studies*, 27(3):714–746.
- Grinblatt, M. and Keloharju, M. (2001). What makes investors trade? *The journal of Finance*, 56(2):589–616.
- Haaland, I. K., Roth, C., Stantcheva, S., and Wohlfart, J. (2024). Measuring what is top of mind. Technical report, National Bureau of Economic Research.
- Halperin, B. and Mazlish, J. Z. (2025). Overreaction and forecast horizon: longer-term expectations overreact more, shorter-term expectations drive fluctuations. *SSRN Working Paper*.
- Hanley, K. W. and Hoberg, G. (2019). Dynamic interpretation of emerging risks in the financial sector. *The Review of Financial Studies*, 32(12):4543–4603.
- Hansen, S., McMahon, M., and Prat, A. (2018). Transparency and deliberation within the fomc: A computational linguistics approach. *The Quarterly Journal of Economics*, 133(2):801–870.
- Hassan, T. A., Hollander, S., Van Lent, L., and Tahoun, A. (2019). Firm-level political risk: Measurement and effects. *The Quarterly Journal of Economics*, 134(4):2135–2202.
- He, S., Lv, L., Manela, A., and Wu, J. (2025). Chronologically consistent large language models. *arXiv preprint arXiv:2502.21206*.
- Hoberg, G. and Manela, A. (2025). The natural language of finance. *Foundations and Trends in Finance*.
- Hong, H. and Stein, J. C. (1999). A unified theory of underreaction, momentum trading, and overreaction in asset markets. *The Journal of finance*, 54(6):2143–2184.
- Hou, K., Xue, C., and Zhang, L. (2015). Digesting anomalies: An investment approach. *The Review of Financial Studies*, 28(3):650–705.

- Hribar, P. and McNinnis, J. (2012). Investor sentiment and analysts' earnings forecast errors. *Management science*, 58(2):293–307.
- Huang, A. H., Lehavy, R., Zang, A. Y., and Zheng, R. (2018). Analyst information discovery and interpretation roles: A topic modeling approach. *Management science*, 64(6):2833–2855.
- Huang, A. H., Zang, A. Y., and Zheng, R. (2014). Evidence on the information content of text in analyst reports. *The Accounting Review*, 89(6):2151–2180.
- Jegadeesh, N. and Kim, W. (2010). Do analysts herd? an analysis of recommendations and market reactions. *The Review of Financial Studies*, 23(2):901–937.
- Jensen, T. (2024). Subjective risk and return. *Working Paper*.
- Jensen, T. I., Kelly, B., and Pedersen, L. H. (2023). Is there a replication crisis in finance? *The Journal of Finance*, 78(5):2465–2518.
- Jha, M., Liu, H., and Manela, A. (2025). Does finance benefit society? a language embedding approach. *The Review of Financial Studies*, page hhaf012.
- Jha, M., Qian, J., Weber, M., and Yang, B. (2024). Chatgpt and corporate policies. Technical report, National Bureau of Economic Research.
- Jiang, H. (2010). Institutional investors, intangible information, and the book-to-market effect. *Journal of Financial Economics*, 96(1):98–126.
- Jiang, Z., Liu, H., Peng, C., and Yan, H. (2024). Investor memory and biased beliefs: Evidence from the field. Technical report, National Bureau of Economic Research.
- Jin, L. J. and Peng, C. (2023). The law of small numbers in financial markets: Theory and evidence. *Available at SSRN 3066369*.
- Juodis, A. and Kučinskas, S. (2023). Quantifying noise in survey expectations. *Quantitative Economics*, 14(2):609–650.
- Kacperczyk, M., Van Nieuwerburgh, S., and Veldkamp, L. (2016). A rational theory of mutual funds' attention allocation. *Econometrica*, 84(2):571–626.
- Kahneman, D. and Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive psychology*, 3(3):430–454.
- Kahneman, D. and Tversky, A. (1973). On the psychology of prediction. *Psychological review*, 80(4):237.

- Ke, S., Montiel Olea, J. L., and Nesbit, J. (2024). Robust machine learning algorithms for text analysis. *Quantitative Economics, Forthcoming*.
- Kelly, B., Xiu, D., et al. (2023). Financial machine learning. *Foundations and Trends® in Finance*, 13(3-4):205–363.
- Kelly, B. T., Malamud, S., Siriwardane, E., and Wu, H. (2024). Behavioral impulse responses. *Swiss Finance Institute Research Paper*, (25-04).
- Khaw, M. W., Li, Z., and Woodford, M. (2021). Cognitive imprecision and small-stakes risk aversion. *The review of economic studies*, 88(4):1979–2013.
- Kim, A. G. and Nikolaev, V. V. (2024). Contextualizing profitability. *Chicago Booth Research Paper*, (23-11):2023–76.
- Kohlhas, A. N. and Walther, A. (2021). Asymmetric attention. *American Economic Review*, 111(9):2879–2925.
- Kučinskas, S. and Peters, F. S. (2024). Measuring under-and overreaction in expectation formation. *Review of Economics and Statistics*, 106(6):1620–1637.
- Kwon, S. Y. and Tang, J. (2025). Extreme categories and overreaction to news. *Review of Economic Studies*, page rdaf037.
- La Porta, R. (1996). Expectations and the cross-section of stock returns. *The Journal of Finance*, 51(5):1715–1742.
- Laarits, T., Matthies, B., Vasudevan, K., and Yang, W. (2025). Policy by committee.
- Law, K. K. (2023). Good-bye i/b/e/s (or not?). *Journal of Financial Reporting*, 8(1):41–61.
- Li, K., Mai, F., Shen, R., Yang, C., and Zhang, T. (2023). Dissecting corporate culture using generative ai—insights from analyst reports. *Available at SSRN 4558295*.
- Li, Z., Van Nieuwerburgh, S., and Renxuan, W. (2025). Understanding rationality and disagreement in house price expectations. *The Review of Financial Studies*, page hhaf073.
- Lin, H.-w. and McNichols, M. F. (1998). Underwriting relationships, analysts’ earnings forecasts and investment recommendations. *Journal of accounting and economics*, 25(1):101–127.
- Liu, H., Peng, C., Xiong, W. A., and Xiong, W. (2022). Taming the bias zoo. *Journal of Financial Economics*, 143(2):716–741.

- Ljungqvist, A., Marston, F., Starks, L. T., Wei, K. D., and Yan, H. (2007). Conflicts of interest in sell-side research and the moderating role of institutional investors. *Journal of Financial Economics*, 85(2):420–456.
- Ljungqvist, A., Marston, F., and Wilhelm Jr, W. J. (2006). Competing for securities underwriting mandates: Banking relationships and analyst recommendations. *The Journal of Finance*, 61(1):301–340.
- Lochstoer, L. A. and Muir, T. (2022). Volatility expectations and returns. *The Journal of Finance*, 77(2):1055–1096.
- Loughran, T. and McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65.
- Ludwig, J., Mullainathan, S., and Rambachan, A. (2025). Large language models: An applied econometric framework. Technical report, National Bureau of Economic Research.
- Ma, S. (2025). Technological obsolescence. *Review of Financial Studies*.
- Malmendier, U. and Nagel, S. (2011). Depression babies: do macroeconomic experiences affect risk taking? *The quarterly journal of economics*, 126(1):373–416.
- Malmendier, U. and Nagel, S. (2016). Learning from inflation experiences. *The Quarterly Journal of Economics*, 131(1):53–87.
- Malmendier, U. and Shanthikumar, D. (2014). Do security analysts speak in two tongues? *The Review of Financial Studies*, 27(5):1287–1322.
- Mazumder, R., Pruitt, S., and Ross, L. (2023). Risk exposures from risk disclosures: What they said and how they said it. *Available at SSRN 4637697*.
- Michaely, R. and Womack, K. L. (1999). Conflict of interest and the credibility of underwriter analyst recommendations. *The review of financial studies*, 12(4):653–686.
- Mikusheva, A. and Sun, L. (2022). Inference with many weak instruments. *The Review of Economic Studies*, 89(5):2663–2686.
- Mikusheva, A. and Sun, L. (2024). Weak identification with many instruments. *The Econometrics Journal*, 27(2):C1–C28.
- Nagel, S. and Xu, Z. (2022). Asset pricing with fading memory. *The Review of Financial Studies*, 35(5):2190–2245.
- Nagel, S. and Xu, Z. (2023). Dynamics of subjective risk premia. *Journal of Financial Economics*, 150(2):103713.



- Odean, T. (1998). Volume, volatility, price, and profit when all traders are above average. *The journal of finance*, 53(6):1887–1934.
- Pouget, S., Sauvagnat, J., and Villeneuve, S. (2017). A mind is a terrible thing to change: confirmatory bias in financial markets. *The Review of Financial Studies*, 30(6):2066–2109.
- Rabin, M. (2002). Inference by believers in the law of small numbers. *The Quarterly Journal of Economics*, 117(3):775–816.
- Rabin, M. and Schrag, J. L. (1999). First impressions matter: A model of confirmatory bias. *The quarterly journal of economics*, 114(1):37–82.
- Sarkar, S. K. (2025). Economic representations.
- Sarkar, S. K. and Vafa, K. (2024). Lookahead bias in pretrained language models. *Available at SSRN*.
- Scharfstein, D. S. and Stein, J. C. (1990). Herd behavior and investment. *The American economic review*, pages 465–479.
- Sims, C. A. (2003). Implications of rational inattention. *Journal of monetary Economics*, 50(3):665–690.
- So, E. C. (2013). A new approach to predicting analyst forecast errors: Do investors overweight analyst forecasts? *Journal of Financial Economics*, 108(3):615–640.
- Song, W. and Stern, S. (2021). Firm inattention and the transmission of monetary policy: A text-based approach. Technical report, Mimeo, University of Michigan.
- Stambaugh, R. F. and Yuan, Y. (2017). Mispricing factors. *The review of financial studies*, 30(4):1270–1315.
- Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, 62(3):1139–1168.
- Thesmar, D. and Verner, E. (2025). Beliefs and stock market fluctuations: New evidence from the past seven decades. *Available at SSRN*.
- Trueman, B. (1994). Analyst forecasts and herding behavior. *The review of financial studies*, 7(1):97–124.
- Valchev, R. and Gemmi, L. (2023). Biased surveys. *NBER Working Paper*, (w31607).

- van Binsbergen, J. H., Bryzgalova, S., Mukhopadhyay, M., and Sharma, V. (2024). (almost) 200 years of news-based economic sentiment. Technical report, National Bureau of Economic Research.
- Van Binsbergen, J. H., 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(6):2361–2396.
- Wachter, J. A. and Kahana, M. J. (2024). A retrieved-context theory of financial decisions. *The Quarterly Journal of Economics*, 139(2):1095–1147.
- Walther, B. R. and Willis, R. H. (2013). Do investor expectations affect sell-side analysts’ forecast bias and forecast accuracy? *Review of Accounting Studies*, 18:207–227.
- Wang, C. (2021). Under-and overreaction in yield curve expectations. *Available at SSRN 3487602*.
- Welch, I. (2000). Herding among security analysts. *Journal of Financial economics*, 58(3):369–396.
- Xu, N. R., Yang, Y., and You, Y. (2025). Fiscal risk perception: Evidence from analyst forecasts. *Available at SSRN 5099927*.

# A Data collection and Preprocessing

## A.1 Investext

I obtain analyst research reports and the relevant identifying information from Investext, provided by Mergent Online. Each publication is identified with a unique ID and contains metadata that includes the title, publisher, authors, companies mentioned (with their tickers), as well as tags provided by Investext such as subject, report style, and categories. I focus on the reports published between 01/01/1998 and 09/30/2023 that satisfy the following filters: Language=English, Region=North America, Country=United States, Report Style=Equity or Company (Equity) Reports, and Category=Equity. I then download the PDF (Portable Document Format) of the analyst research report using the file link associated with each report ID.

## A.2 Parsing PDF and preparing content chunks

Optical Character Recognition (OCR)

Layout Recognition

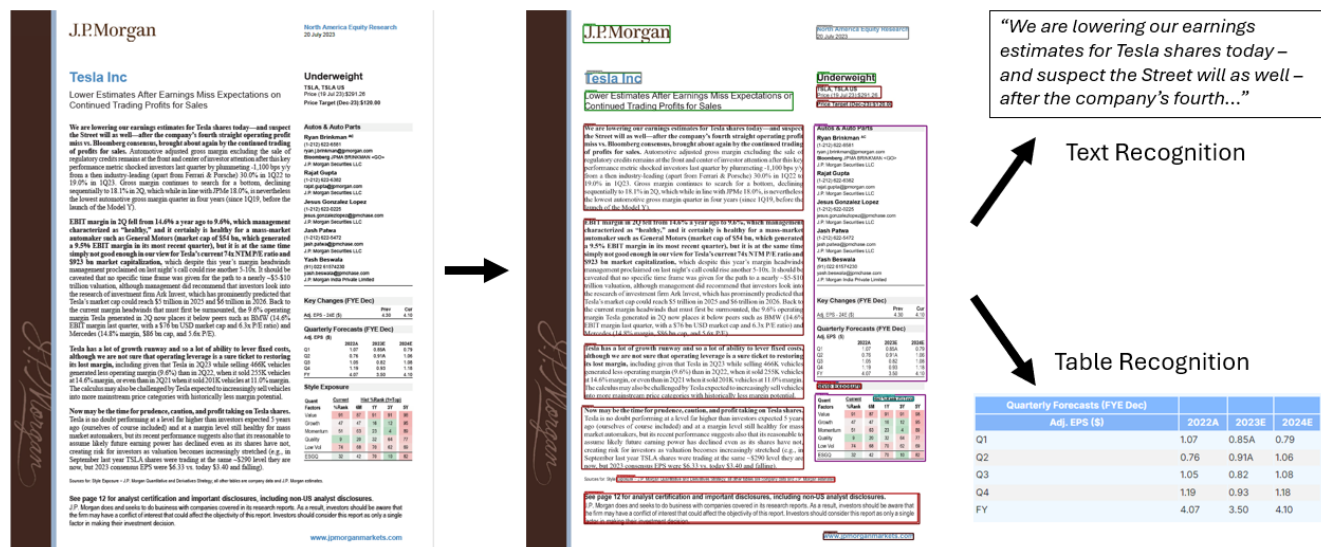


Figure A.1: Illustration of the parsing of analyst reports

A PDF file is an unstructured data format that can contain different layouts and forms of data such as text, pictures, or tables. Because of these versatile features sophisticated parsing techniques are required to represent the content of a PDF into a workable format.

I implement a parsing procedure adopted from the Deep Document module in RAGFlow, developed by InfiniFlow<sup>53</sup>. The implementation combines optical character recognition, layout recognition, and content recognition that distinguishes between textual and tabular content. Specifically, I first convert each PDF page into an image, and then use a layout recognition model pre-trained

<sup>53</sup>Further details can be found at <https://github.com/infiniflow/ragflow/blob/main/deepdoc/README.md>

using computer vision techniques to identify components on a PDF page. For each component, another pre-trained model is used to identify if it is textual, that is comprising mainly text, or tabular. I focus on the textual components and, in the final step, extract the texts from the part of the image that corresponds to those components. Figure A.1 illustrates the parsing process.

### A.3 Matching Investext companies

I match Investext companies with their CUSIP. To match the companies mentioned in the analysts' research reports, I first remove all reports that contain no references to companies identified by Investext. I merge the CRSP identity file (`stocknames` table) and the IBES identity file (`id` table) obtained from WRDS to create a `ticker-company-name-CUSIP` linking table. For the company identifiers provided by Investext and already in the CUSIP format, I use these directly as the identifier. If the identifier is in ticker format, I match it with the ticker and obtain the corresponding CUSIP. If none of the matches succeed, I try to match the company name using the FuzzyWuzzy package in Python, with a match cutoff of 90. All CUSIPs identified in this process are retained, as there could be multiple companies mentioned in a single research report.

### A.4 Matching Investext contributors and authors

I first match the authors of the research reports with IBES analysts identified by a unique code that contains their last name and sometimes their first initial (`amaskcd`). I download the analyst code from the detailed recommendation and price target file from IBES, and clean up the last name by removing the suffix (such as CPA, CFA, Jr). I then clean it using the HumanName package in Python. I try to match each name that appears as an author of an analyst report in Investext in a given quarter with analysts that provide announcements, activations or reviews of recommendations or price targets to IBES in the 12 months before or after the quarter, by selecting the `anndats`, `actdats`, or `revdats` variable. If there are multiple matches to the last name, I first check if the company CUSIPs for the report overlaps with the CUSIPs for which the analyst provided recommendations or price targets for the previous or following 12 months, removing those matches with no overlapping CUSIPs. For the remaining multiple matches, if multiple authors are associated with the report, I require all identified `amaskcd` codes to be associated with the same brokerage firm (identified by `estimid`) for the previous and following 12 months.

I then match contributors in Investext to brokerage firms in IBES. I focus on brokerage firms that, on average, publish more than 10 reports in a quarter. I obtain `estimid` from the detail `recommendation` file from IBES. I first match the abbreviated broker names in `estimid` to full broker names (or their subsidiaries) in Investext by resemblance, and check that the analysts associated with the brokerage firm in Investext has `amaskcd` codes that are associated with the same `estimid` in the same quarter. For the remaining brokerages that do not have a clear abbreviation

resemblance, I gather the analyst `amaskcd` code associated with the brokerage firm and find the `estimid` that contains the most overlapping `amaskcd` codes associated with it. As a sanity check, I obtain the brokerage firm matching table from Jared Flake, who constructed the matching of IBES `estimid` codes with brokerage firms in the Capital IQ Transcript dataset (Flake, 2023). I confirm the quality of the match and resolve any discrepancies manually. To match with the `estimator` code in the detailed EPS forecast file in IBES, I obtain the linking table from Kelvin Law and manually check the matching quality of the brokers in Investtext (Law, 2023). At the end of the matching process, I am able to match 115 brokers in the Investtext database.

## A.5 Preprocessing IBES data

I process the IBES earnings forecast data following Kelly et al. (2024). I download analyst EPS forecast from the IBES unadjusted detailed history table in WRDS (`ibes.detu_epsus`). The table contains raw EPS forecasts, identifiers for each stock (CUSIP), broker identifiers, analyst identifiers, forecast announcement dates, and forecast period indicators (FPI), which I use to identify forecast horizons. I gather annual FY1 to FY5 EPS forecasts, which correspond to FPI from 1 to 5. Realized EPS values and announcement dates are taken from the IBES unadjusted detail actuals table (`ibes.actu_epsus`). The actual and forecast tables are merged based on the CUSIP and forecast period end date (`fpedats`).

It has been noted in previous literature that EPS forecasts in IBES may differ mechanically from realizations if stock splits occur between the forecast date and the realization date (e.g. Diether et al., 2002). Following the literature, I covert all forecasts and realizations to the share basis at the time of the forecast date. Specifically, I obtain cumulative adjustment factors (CFACSHR) from CRSP’s daily stock file (`crsp_a_stock.dsf`). I merge it with IBES table using WRDS’s IBES-CRSP linking table (`wrdsapps_link_crsp_ibes.ibcrsphist`). I only retain the observations for which the linkage holds when the forecast is announced. When the IBES forecast announcement falls outside the trading dates in CRSP, I use the adjustment factor on the closest preceding trading date in CRSP.

I also retain only stocks with share codes 10 or 11 and with exchange codes 1, 2, or 3. I also drop the observations for which the forecast’s announcement date (`anndats`) is after the forecast period end date (`fpedats`), and drop the observations for which the implied forecast horizon from the forecast announcement date and the forecast’s target date align with the FPI.

## A.6 Matching reports with IBES records

In the final step, I match IBES records with Investtext analyst reports by matching on the `estimator-cusip-anndats` combinations. Each identifies an announcement of a company from a broker. For each `estimator-cusip-anndats` combination, I match it with reports that are pub-

lished by the identified broker in the period from 1 business day before to 1 business day after the announcement date that contains the identified CUSIP.

### A.6.1 Summary statistics

Table OA.2 reports summary statistics for the matched IBES sample with analyst reports, as well as for the sample of 115 selected brokers and the full IBES sample. The matched data set contains more than 1.1 million reports covering 818,780 unique EPS forecast announcements in IBES, which represents 49.8% of the total announcements made by the 115 brokers and 20.6% of all IBES announcements made by 1,201 brokers. Each report contains about nine pages. In terms of analysts' research activities, each broker has close to 200 reports every quarter that are mapped to the IBES records, and each firm, on average, gets five reports per quarter.

Figure OA.1 shows how the summary statistics with respect to analyst reports evolve over time. Panel (a) shows that the total number of reports and the number of stocks covered in the matched sample steadily increase over time, and Panel (b) shows the analysts' reports get longer over time. Both pattern indicates the growing importance of information production by sell-side analysts in the financial markets.

Table OA.3 reports the summary statistics of realized earnings yield, as well as the sample mean of analysts' forecast errors and forecast revisions. The earnings yield in the matched sample is similar to the full IBES sample. The forecast error in the matched sample is, on average, less negative than the full sample, and the revisions are, on average, larger, indicating the institutions in the sample are less prone to optimism bias and incorporate new information in their forecasts more actively.

## B Proofs and additional theoretical results

### B.1 Measurement error in embeddings

The main identification assumption under the text-instrumented Coibion-Gorodnichenko (TICG) regression (Assumption 2) is that the text embedding related to topic  $k$  provides sufficient statistical support for the information related to component  $k$  that drives the analyst’s belief revision. In this section, I discuss what happens, in a theoretical sense, when such an identification assumption fails. In Appendix B.2 I study  $\beta_{TICG}$  in simulations.

Assumption 2 is that  $X_k = \tilde{X}_k$  and  $X_k \perp \epsilon$ . When such conditions are violated, we can think of  $X_k$  containing measurement errors with general covariance structures

$$X_k = \tilde{X}_k + E_k, \quad (25)$$

where  $E_k$  is the measurement error in  $X_k$  that can be correlated with  $\tilde{X}_k$ , with other  $E_{k'}$  for topic  $k'$ , and with the non-informational driver of belief revision  $\epsilon$ . Below I discuss several examples of such measurement errors.

**Noise in textual embeddings.** While I use cutting-edge large language models (LLMs) to embed the text in analyst reports and generate a representation of the information, such embedding representation may not capture the true meaning of the text. While modern embedding techniques usually produce high-dimensional embedding vectors, they will not capture all the information in the text, which is inherently an ultra-high-dimension object. In this case,  $X_k$  will be different from  $\tilde{X}_k$ , even when analysts truthfully and completely describe all the information they observe.

In Eq. (25), noises induced by the LLM’s embedding function can be thought of as introducing some random  $E_k$ . When the noise in the representation is independent of the information content itself,  $E_k$  is the classic measurement error and is independent of  $\tilde{X}_k$  and  $\epsilon$ .

**Misclassification of topics** The misclassification of topics will induce a measurement error in  $X_k$ . If some information tuples of topic  $k$  were to be misclassified to be in topic  $k'$ , this would induce nonzero  $E_k$  and  $E_{k'}$  that are correlated. Specifically, suppose the misclassified information has embedding representation  $X_k^{miss}$ , then

$$X_k = \tilde{X}_k - X_k^{miss}; \quad X_{k'} = \tilde{X}_{k'} + X_k^{miss} \quad (26)$$

so that  $E_k = -X_k^{miss}$  and  $E_{k'} = X_k^{miss}$ .

**Information selection.** Analysts might selectively withhold information related to some topics, for example, due to constraints on the length of analyst reports. This case can be thought of as

akin to some informational content in topic  $k$  being missed, and the embedding representation of the missing content will show up as  $E_k$  in Eq. (25).

**Ex post justification of non-informational beliefs.** Analysts might make forecast revisions driven by non-informational factors, for example institutional reasons, and they ex post justify their forecasts by describing some information related to certain topics in the reports. Suppose they chooses to change the text related to topic  $k$ . This ex post justification will show up as  $E_k$  and be correlated with  $\epsilon$ .

I then derive the formula for  $\beta_{TICG}^k$  given the general measurement error, Eq. (25). Denote  $\tilde{\mathbf{X}} = [\tilde{X}^1, \tilde{X}^2, \dots, \tilde{X}^K]$  the concatenated  $\tilde{X}^k$ ,  $\mathbf{E} = [E^1, E^2, \dots, E^K]$  the concatenated  $E^k$ . The concatenated  $X^k$  is, therefore,  $\mathbf{X} = \tilde{\mathbf{X}} + \mathbf{E}$ , and the observed forecast revision is given by  $Rev[y] = \tilde{\mathbf{X}}'\delta + \epsilon$ , where  $\delta = [\delta_1, \delta_2, \dots, \delta_K]$ .

Denote  $\Sigma_X := Var(\tilde{\mathbf{X}})$ ,  $\Sigma_E := Var(\mathbf{E})$ ,  $\Sigma_{XE} := Cov(\tilde{\mathbf{X}}, \mathbf{E})$ , and  $\Sigma_{E\epsilon} := Cov(\mathbf{E}, \epsilon)$ . These are the covariances responsible for the violation of Assumption 2 when they are nonzero. Given this general setup, denote  $\lambda = [\lambda_1, \lambda_2, \dots, \lambda_K]$  the stacked  $\lambda$  coefficients in the first stage regression. It is given by

$$\lambda = (\Sigma_X + \Sigma_E + 2\Sigma_{XE})^{-1}((\Sigma_X + \Sigma_{XE})\delta + \Sigma_{E\epsilon}) \quad (27)$$

Note that  $\lambda = \delta$  only when  $\Sigma_E = \Sigma_{XE} = 0$  and  $\Sigma_{E\epsilon} = 0$ , which corresponds to Assumption 2. Denote  $\Lambda = diag(\lambda_k)$  and  $\Delta = diag(\delta_k)$ , the stacked TICG coefficients  $\beta_{TICG} = [\beta_{TICG}^1, \beta_{TICG}^2, \dots, \beta_{TICG}^K]$  equal to

$$\beta_{TICG} = [\Lambda'(\Sigma_X + \Sigma_E + \Sigma_{XE})\Lambda]^{-1}\Lambda(\Sigma_X + \Sigma_{XE})\Delta\mathbf{b}, \quad (28)$$

where  $\mathbf{b} = [b_1, b_2, \dots, b_K]$  is the vector of true misreaction parameters.

### B.1.1 Identification in the block-independence case

A special case is when  $\tilde{\mathbf{X}}$  and  $\mathbf{E}$  are block independent, that is,  $\tilde{X}^k$  and  $E^k$  are independent across  $k$ . This is more likely to be satisfied when the signals  $s_k$  are independent across  $k$ , or when signals are residualized. When such block independence is satisfied, all relevant matrices are block diagonal, so Eq. (28) decouples and each coefficient is a scalar multiple of  $b_k$ :

$$\beta_{TICG}^k = \frac{\lambda'_k(\Sigma_{X,kk} + \Sigma_{XE,kk})\delta_k}{\lambda'_k(\Sigma_{X,kk} + \Sigma_{E,kk} + 2\Sigma_{XE,kk})\lambda_k} b_k \quad (29)$$

Notice that the first-stage regressions give

$$\lambda_k = (\Sigma_{X,kk} + \Sigma_{E,kk} + 2\Sigma_{XE,kk})^{-1}(\Sigma_{X,kk} + \Sigma_{XE,kk})(\delta_k + \Sigma_{E\epsilon,k}).$$



Plugged into Eq. (29), we obtain

$$\beta_{TICG}^k = \frac{\lambda'_k(\Sigma_{X,kk} + \Sigma_{XE,kk})\delta_k}{\lambda'_k(\Sigma_{X,kk} + \Sigma_{XE,kk})\delta_k + \lambda'_k\Sigma_{E\epsilon,k}} b_k. \quad (30)$$

We conclude then that  $\beta_{TICG}^k$  still identifies  $b_k$  when embeddings and measurement errors are block independent and that  $\Sigma_{E\epsilon,k} = 0$ . The block independence can be ensured by residualizing embedding vectors with respect to each other, which is conducted when implementing the TICG regression (see Appendix B.7.2). After residualization, classic measurement errors and exogenous selection of information (for example due to page length) within topic can be allowed.

### B.1.2 Sign identification

Without any restrictions on  $\tilde{\mathbf{X}}$  and  $\mathbf{E}$ , Eq. (28) implies that  $\beta_{TICG}^k$  will be a biased estimate of  $b_k$ . However, if we are only interested in the *sign* of misreaction, that is, we only want to know if the analyst is over- or underreacting to information related to topic  $k$ , then the sign of  $\beta_{TICG}^k$  still identifies the sign of  $b_k$ , provided that the violation of Assumption 2 is not large.

Specifically, denote

$$v_{kj} := e'_k(\Lambda'(\Sigma_X + \Sigma_E + \Sigma_{XE})\Lambda)^{-1}\Lambda'(\Sigma_x + \Sigma_{XE})\delta_j,$$

where  $e_k$  is the  $k$ -th standard basis vector, we have

$$\beta_{TICG}^k = v_{kk}b_k + v'_{k,-k}b_{-k} \quad (31)$$

Assume, without loss of generality, that  $\Delta$  has all positive entries,  $v_{kk}$  is guaranteed to be positive, so the condition for  $\beta_{TICG}^k$  to have the same sign as  $b_k$  is

$$v_{kk}|b_k| > |v'_{k,-k}b_{-k}|. \quad (32)$$

By Holder's inequality, a sufficient condition for it is

$$\frac{|b_k|}{\max_{j \neq k} |\beta_j|} > \frac{\sum_{j \neq k} |v_{kj}|}{v_{kk}}. \quad (33)$$

Such condition guarantees the sign identification of  $\beta_{TICG}^k$ . It is more likely to be satisfied when the cross-topic correlation of the measurement error is small ( $\sum_{j \neq k} |v_{kj}|$  is small), the violation of the exclusion restriction is small ( $v_{kk}$  is large), and when the analyst has a greater misreaction to topic  $k$  ( $|\beta_k|$  is large).

## B.2 Simulation

I investigate the effectiveness of the TIGC regression in uncovering belief misreaction using simulations calibrated to match the empirical setting. I first generate 27 200-dimensional random vectors according to standard joint-normal distribution  $W_i^k$  for  $k = 1, 2, \dots, 27$ , and set the true embedding representation of the 27 topics as follows:

$$\tilde{X}_i^{k,j} = L'_x W_i^j,$$

where  $\tilde{X}_i^{k,j}$  is the  $j$ -th column of  $\tilde{X}_i^k$  and  $W_i^j = [W_i^{1,j}, W_i^{2,j}, \dots, W_i^{27,j}]$  is the vector of the  $j$ -th column of all 27  $W$  vectors. I set  $L$  to be the Cholesky decomposition of the correlation matrix, which I assume to have an equicorrelation structure

$$L_x L'_x = \begin{pmatrix} 1 & \rho_x & \cdots & \rho_x \\ \rho_x & 1 & \cdots & \rho_x \\ \vdots & \vdots & \ddots & \vdots \\ \rho_x & \rho_x & \cdots & 1 \end{pmatrix}$$

In this setup,  $\rho_x$  captures the correlation of cross-topic information. If  $\rho_x = 0$ , all topics contain uncorrelated information (which is equivalent to uncorrelated  $s_k$  in Eq. (2)).

I simulate analysts' subjective reactions to information related to topic  $\delta_k$  as drawn from the joint-normal distribution,

$$\delta_k \sim N(0, \frac{1}{5400}I).$$

I define analysts' true belief reaction about topic  $k$  as  $\hat{w}_k s_i^k = \delta'_k \tilde{X}_i^k$  (for simplicity I assume  $\mu_k = 0$ ), following the definition of the operational formulation in Definition 1. The observed belief revision is generated as follows:

$$Rev_i = \sum_{k=1}^{29} \delta'_k \tilde{X}_i^k + \epsilon_i \quad (34)$$

where I generate  $\epsilon_i$  according to  $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ . Because the variance of  $\sum_{k=1}^{27} \delta'_k \tilde{X}_i^k$  is equal to 1, I choose  $\sigma_\epsilon^2 = 5$  so that the explainable part of the forecast revision is 17%, which aligns empirically with the total explained variance ratio in the first stage.

For the bias parameter  $\mathbf{b}$ , I set  $b_1 = w_1/\hat{w}_1 - 1$  to be either 0.1 or -0.1, which corresponds to the case of under- and overreaction to the first topic, and the magnitude aligns with the typical magnitude in the TIGC results in Figure 5. I generate the bias parameters  $b_k$ ,  $k = 2, 3, \dots, 27$  using one of the following three schemes.

- Uniform Random:  $b_k \sim Uniform([-0.9, 0.9])$

- Same misreaction:  $b_k = b_1$
- Opposite misreaction:  $b_k = -b_1$

The rational belief revision is therefore given by

$$Rev_i^{rational} = \sum_{k=1}^{29} (1 + b_k) \delta'_k \tilde{X}_i^k.$$

I generate the empirically observed embedding vector  $X_i^k$  using

$$X_i^k = \tilde{X}_i^k + E_i^k + \rho_e \epsilon_i,$$

where  $E_i^{k,j} = L'_e V_i^j$  and where  $V_i^j$  is generated using standard joint-normal distribution  $V_i^{k,j} \sim N(0, \sigma_e^2)$  and  $L_e$  is the Cholesky decomposition of the correlation matrix

$$L_e L'_e = \begin{pmatrix} 1 & \rho_e & \cdots & \rho_e \\ \rho_e & 1 & \cdots & \rho_e \\ \vdots & \vdots & \ddots & \vdots \\ \rho_e & \rho_e & \cdots & 1 \end{pmatrix}$$

In other words,  $\sigma_e^2$  captures the amount of noise in the embedding vectors and  $\rho_e$  captures the correlation of noise across topics and topic misclassification.  $\rho_e$  is the correlation between text embeddings and the non-informational driver  $\epsilon_i$ , and captures analysts' ex post rationalizations or strategic communication incentives.

### B.2.1 Simulation results

I simulate  $\delta_k$ ,  $\tilde{X}_i^k$  and  $X_i^k$  for  $i = 1, 2, \dots, N$  with  $N = 100,000$ , which is slightly fewer than the number of observations in the long-term earnings forecast revisions in the sample. Using the simulated data, I perform the following second-stage regression:

$$Rev_i^{rational} - Rev_i = \sum_{k=1}^{27} \beta_k \cdot \widehat{Rev}_i^k + u_i, \quad (35)$$

where the predicted revision  $\widehat{Rev}_i^k$  is constructed in one of the three ways: true revision  $\widehat{Rev}_i^k = \delta'_k \tilde{X}_i^k$ , fitting  $Rev_i$  on  $\{X_i^k\}_{k=1}^{27}$  using OLS, and fitting  $Rev_i$  on  $\{X_i^k\}_{k=1}^{27}$  using the jackknife estimator. I simulate 100 draws of data realizations and look at the distributions of  $\beta_k^{true}$ ,  $\beta_k^{OLS}$ , and  $\beta_k^{JIVE}$  for the first topic, without loss of generality.

Figure OA.13 shows the simulation results with different calibrations. First, Panel (A) shows the results when Assumption 2 holds, such that there is no measurement error ( $\sigma_e^2 = 0$ ). I allow

for correlation between topics by setting  $\rho_x = 0.4$ . From the plots in Panel (A), we can see that the JIVE estimates of the misreaction parameter  $b_1$  is as good as if the analysts' true reaction  $\widehat{w}_k s_k$  were known, and that it is an unbiased estimator of  $b_1$ . This result is robust for different setups in the misreaction of other topics  $b_k$ . In contrast, if we use OLS to estimate predicted revisions in the first stage, the estimated misreaction coefficient will be negatively biased. This is because, with high-dimensional embedding vector  $X^k$ 's, OLS estimates are subject to overfitting on  $\epsilon_i$ , which results in a negative bias in the second-stage estimates in the same fashion as the negative bias in the CG regression.

Panel (B) introduces measurement errors  $E_i^k$  that are correlated across topics but are uncorrelated with  $\epsilon_i$ . We can see that  $\beta_k$  estimated using JIVE still aligns with the true  $b_1$  and is almost as good as knowing analysts' true reaction function. Such alignment is stronger for negative  $b_1$  (overreaction detection). For positive  $b_1$  (underreaction detection), there is a negative bias associated with the JIVE, which suggests that it will likely underestimate the degree of underreaction to a topic.

In Panel (C), I introduce certain violations of the exclusion restriction by setting  $\rho_\epsilon = 0.0001$ . Since  $\rho_\epsilon$  applies to every column of  $X_i^k$ , the extent of the total violation is meaningful even though the number 0.0001 seems small. The result shows, however, that the JIVE procedure can still recover the true misreaction, just as if the analysts' true reaction were observed, as long as  $\rho_\epsilon$  is not large enough. In Panel (D) I increase  $\rho_\epsilon = 0.005$ , and with this as the extent of violation, the JIVE estimate becomes severely biased and no longer identifies the true misreaction. However, a high degree of correlation between  $\epsilon$  and  $X^k$  also implies that the explained variance will be inflated, as shown in Table OA.11. This is somewhat inconsistent with the TICG results in Section 5, where the more negative TICG coefficients relate to the topics that explain less of the variation in forecast revisions.

In summary, the simulation exercises validate the efficacy of the TICG regression when Assumption 2 holds. When this assumption fails, the procedure can accommodate a certain degree of measurement error, and the correlation of measurement errors across topics is less concerning.

### B.3 Error-on-variable regression

**Proposition 3** (Regression of forecast error on researcher-constructed variable)

*Suppose that the analyst forms their belief according to Eq. (2), and a researcher constructs a signal  $x_{k,t}$  based on the information observable to them that aims to capture information about some latent component  $k$ . Assuming that  $x_{k,t}$  is an unbiased proxy for the true information the analyst observes, that is,  $x_k = s_{k,t} + \eta_t$ . When the researcher regresses forecast errors on the custom signal,*

$$y_{t+1} - F_t[y_{t+1}] = \alpha + \beta_{EV} x_{k,t} + e_t, \quad (36)$$

the coefficient is given by

$$\beta_{EV} = b_k \frac{\sigma_{\mu s_k}^2}{\sigma_s^2 + \sigma_\eta^2} + \sum_{k' \neq k} b_{k'} \frac{\sigma_{k, \mu s_{k'}}}{\sigma_s^2 + \sigma_\eta^2}, \quad (37)$$

where  $\sigma_{\mu s_k}^2 := \text{Var}(s_{k,t} - \mu_{t-1})$  and  $\sigma_{k, \mu s_{k'}} = \text{Cov}(s_{k,t}, s_{k',t} - \mu_{t-1})$ .

**Proof of Proposition 3** The error-on-variable regression has coefficient

$$\begin{aligned} \beta_{EV} &= \frac{\text{Cov}(\sum_{k=1}^K (w_k - \hat{w}_k)(s_{k,t} - \mu_{t-1}) - \epsilon_t, s_{k,t} + \eta_{k,t})}{\text{Var}(s_{k,t} + \eta_t)} \\ &= b_k \frac{\sigma_{\mu s_k}^2}{\sigma_s^2 + \sigma_\eta^2} + \sum_{k' \neq k} b_{k'} \frac{\sigma_{k, \mu s_{k'}}}{\sigma_s^2 + \sigma_\eta^2}, \end{aligned} \quad (38)$$

where  $\sigma_{\mu s_k}^2 := \text{Var}(s_k - \mu_k)$  and  $\sigma_{k, \mu s_{k'}} = \text{Cov}(s_k, s_{k'} - \mu_{k'})$ .  $\square$

## B.4 Proofs for results in main text

**Proof of Proposition 1** The baseline CG coefficient follows

$$\begin{aligned} \beta_{CG} &= \frac{\text{Cov}(y_{t+1} - F_t[y_{t+1}], F_{t-1}[y_{t+1}] - F_t[y_{t+1}])}{\text{Var}(F_t[y_{t+1}] - F_{t-1}[y_{t+1}])} \\ &= \frac{\text{Cov}(y_{t+1} - E_t[y_{t+1}](E_t[y_{t+1}] - F_t[y_{t+1}]), F_t[y_{t+1}] - F_t[y_{t+1}])}{\text{Var}(F_t[y_{t+1}] - F_t[y_{t+1}])} \\ &= \frac{\text{Cov}(E_t[y_{t+1}] - F_t[y_{t+1}], F_t[y_{t+1}] - F_{t-1}[y_{t+1}])}{\text{Var}(F_t[y_{t+1}] - F_{t-1}[y_{t+1}])} \\ &= \frac{\text{Cov}(\sum_{k=1}^K (w_k - \hat{w}_k)(s_{k,t} - \mu_t) - \epsilon_t, \sum_{k=1}^K \hat{w}_k(s_{k,t} - \mu_t) + \epsilon_t)}{\text{Var}(\sum_{k=1}^K \hat{w}_k(s_{k,t} - \mu_t) + \epsilon_t)} \\ &= \mathbf{b}' \frac{\Sigma_{\mu s} \hat{\mathbf{w}}}{\hat{\mathbf{w}}' \Sigma_{\mu s} \hat{\mathbf{w}} + \sigma_\epsilon^2} - \frac{\sigma_\epsilon^2}{\hat{\mathbf{w}}' \Sigma_{\mu s} \hat{\mathbf{w}} + \sigma_\epsilon^2}, \end{aligned} \quad (39)$$

where  $\Sigma_{\mu s} := \text{Var}(s_{k,t} - \mu_t)$  denotes the covariance matrix of the signals,  $\hat{\mathbf{w}} := [\hat{w}_1, \hat{w}_2, \dots, \hat{w}_K]$  and  $\mathbf{b} = [b_1, b_2, \dots, b_K]$ . The third line follows from the fact that  $F_t[y_{t+1}] - F_{t-1}[y_{t+1}]$  is a function of  $s_{k,t}$ , which means it is not correlated with the forecast error of rational forecasts, i.e.  $\text{Cov}(y_{t+1} - E_t[y_{t+1}], F_t[y_{t+1}] - F_{t-1}[y_{t+1}]) = 0$ .  $\square$

**Proof of Proposition 2** When  $X_i^k = \tilde{X}_i^k$  and  $X_i^k \perp \epsilon_i$ , given the formulation

$$\text{Rev}_t[y_{t+1}] = \sum_{k=1}^K \lambda'_k \tilde{X}_t^k + \epsilon_t,$$

the first-stage regression of  $Rev_t[y_{t+1}]$  onto  $\mathbf{X}_t = [X_t^1, X_t^2, \dots, X_t^K]$  has coefficients  $\widehat{\lambda}_k$  as a consistent estimator of  $\lambda_k$ , which means

$$\widehat{Rev}_t[y_{t+1}] = \sum_{k=1}^K \widehat{\lambda}'_k X_t^k = \sum_{k=1}^K \lambda'_k X_t^k$$

is a consistent estimator of  $\sum_{k=1}^K \widehat{w}_k(s_{k,t} - \mu_t)$ . Since  $\widehat{w}$  and  $w_k$  are not changing with  $t$ , we can define  $a_k := w_k/\widehat{w}_k$ , which means the rational forecast revision can be written as

$$E_t[y_{t+1}] - E_{t-1}[y_{t+1}] = \sum_{k=1}^K a_k \widehat{w}_k(s_{k,t} - \mu_t)$$

and is also a function of  $X_t^k$ . Therefore, the second-stage regression for an individual topic  $k$

$$Error_t[y_{t+1}] = \beta_{TICG}^k \widehat{Rev}_t[y_{t+1}] + \sum_{k' \neq k} \xi'_{k'} X_t^{k'} + u_t \quad (40)$$

is equivalent to the “residualized” regression where  $Error_t[y_{t+1}]$  and  $\widehat{Rev}_t$  are residualized with respect to all embedding vectors of  $X_t^{k'}$  for  $k' \neq k$  by the Frisch–Waugh–Lovell (FWL) theorem,

$$Error_t[y_{t+1}]^\perp = \beta_{TICG}^k \widehat{Rev}_t^\perp[y_{t+1}] + u_t. \quad (41)$$

where  $^\perp$  denotes the residualized quantity. Notice that

$$\begin{aligned} Cov(y_{t+1}^\perp - F_t[y_{t+1}]^\perp, Rev_t[y_{t+1}]^\perp) &= Cov(y_{t+1}^\perp - E_t[y_{t+1}]^\perp + (E_t[y_{t+1}]^\perp - F_t[y_{t+1}]^\perp), \widehat{Rev}_t^\perp[y_{t+1}]) \\ &= Cov(E_t[y_{t+1}]^\perp - F_t[y_{t+1}]^\perp, \widehat{Rev}_t^\perp[y_{t+1}]), \end{aligned} \quad (42)$$

since  $\widehat{Rev}_t[y_{t+1}]$  is a function of  $\{X_t^k\}_{k=1}^K$  and thus uncorrelated with the forecast error of rational beliefs. Using the assumption that  $E_{t-1}[y_{t+1}] = F_{t-1}[y_{t+1}]$ , we have

$$\begin{aligned} \beta_{TICG}^k &= \frac{Cov(E_t[y_{t+1}]^\perp - F_t[y_{t+1}]^\perp, \widehat{Rev}_t^\perp[y_{t+1}])}{Var(\widehat{Rev}_t^\perp[y_{t+1}])} \\ &= \frac{Cov(E_t[y_{t+1}]^\perp - E_{t-1}[y_{t+1}]^\perp - \widehat{Rev}_t^\perp[y_{t+1}] - \epsilon_t, \widehat{Rev}_t^\perp[y_{t+1}])}{Var(\widehat{Rev}_t^\perp[y_{t+1}])} \\ &= \frac{Cov(a_k \widehat{w}_k(s_{k,t} - \mu_t) - \widehat{w}_k(s_{k,t} - \mu_t) - \epsilon_t, \widehat{w}_k(s_{k,t} - \mu_t))}{Var(\widehat{w}_k(s_{k,t} - \mu_t))} \\ &= a_k - 1 = \frac{w_k}{\widehat{w}_k} - 1 \end{aligned} \quad (43)$$

where the third equality follows from the assumption that  $\widehat{w}_k(s_{k,t} - \mu_t) = \lambda'_k X_t^k$  for all  $k$ , and the fourth equality follows from the assumption that  $X_t \perp \epsilon_t$   $\square$

## B.5 A belief formation model motivated by textual evidence

Given the set of empirical findings regarding potential behavioral mechanisms, I propose a stylized model of belief formation that seeks to jointly explain these observations. The model builds on the general framework introduced in Section 2 and provides microfoundations for two key empirical patterns: (1) analysts exhibit greater overconfidence when interpreting qualitative information, and (2) analysts have a tendency to herd in their forecasts.

In the model, there are  $N$  analysts indexed by  $i$ , each receiving a set of noisy signal  $s_{k,t}$  for  $k = 1$  to  $K$  about the firm's future earnings  $y_{t+1}$ , that is,

$$s_{i,k,t} = y_{t+1} + v_{i,k,t}; \quad v_{k,t} \sim N(0, \tau^{-1}). \quad (44)$$

For simplicity, I assume that all signals have the same precision. Before receiving signals at time  $t$ , each analyst is endowed with the common and correct prior belief about  $y_{t+1}$ , which I assume to be normal:

$$y_{t+1} | \mathcal{F}_{t-1} \sim N(\mu_{t-1}, \tau_0) \quad (45)$$

The first deviation from rationality is that the analyst is more overconfident when interpreting qualitative than quantitative information. I denote the analyst's subjective estimate of the precision of signal  $k$  by  $\widehat{\tau}_k$  and assume that

$$\widehat{\tau}_k > \tau \text{ for } k \in \mathcal{K}_{qual}; \quad \widehat{\tau}_k = \tau \text{ for } k \in \mathcal{K}_{quant}. \quad (46)$$

This means that analysts overestimate the precision of qualitative signals, while their perception of the precision of quantitative signals is accurate. This is supported by the finding that when analysts face more qualitative information, they exhibit greater overconfidence and use more assertive language (Table 4).

The second deviation from rationality is the analysts' tendency to herd. Specifically, the analyst's objective function includes the expected squared forecast error and a penalty term for deviating from the consensus forecast. This formulation captures the herding behavior. In other words, analyst  $i$  chooses a forecast  $\widehat{y}$  that minimizes

$$\min_{\widehat{y}} F_{i,t}[(\widehat{y} - y_{t+1})^2 + \gamma(\widehat{y} - \bar{\widehat{y}})^2 | \{s_{i,k,t}\}_{k=1}^K]. \quad (47)$$

Here,  $\bar{\widehat{y}} = \frac{1}{N} \sum_i \widehat{y}_i$  denotes the consensus forecast, and  $\gamma > 0$  captures the strength of the herding

tendency. This herding behavior may arise from reputational or career concerns, in which case  $\gamma$  reflects the magnitude of such incentives. Alternatively, herding may result from a misperception that the consensus is highly informative. In this interpretation,  $\gamma$  represents the excess weight that the analyst places on the consensus forecast relative to their private information.

The next proposition characterizes the original CG and text-instrumented CG-regression coefficients when there are a total of  $N$  forecasters issuing forecasts.

**Proposition 4** (A belief-formation model supported by textual evidence)

*Suppose there are a total of  $N$  analysts forecasting the firm earnings, and each has subjective precision given in Eq. (46), with their objective function given in Eq. (47) and the rational weight  $w_k$  and subjective weight  $\hat{k}$  given by*

$$w_k = \frac{\tau}{\tau + \tau_0}; \quad \hat{w}_k = \frac{\frac{\hat{\tau}_k}{\hat{\tau}_k + \tau_0}}{(1 + \gamma) - \frac{1}{\hat{\tau}_k + \tau_0}(\gamma\hat{\tau}_k + \frac{\gamma}{N}\tau_0)}. \quad (48)$$

*For certain parameter values of  $\hat{\tau}$ , for small  $N$ , the text-instrumented CG regression finds  $\beta_{TICG}^k < 0$  for  $k \in \mathcal{K}_{qual}$ . The original  $\beta_{CG}$  is negative for small  $N$  but positive for large  $N$ .*

**Proof of Proposition 4** The rational updating weight  $w_k$  follows directly from Bayesian updating. To derive the subjective weight  $\hat{w}_k$ , notice that the first-order condition of an analyst's objective function implies

$$\hat{y} = \frac{1}{1 + \lambda} F_{i,t}[y_{t+1}] + \frac{\lambda}{1 + \lambda} F_{i,t}[\hat{y}], \quad (49)$$

where  $F_{i,t}$  denotes analyst  $i$ 's subjective expectation. Given subjective precision  $\hat{\tau}_k$ , their subjective forecast for  $y$  is

$$F_{i,t}[y] = \mu_{t-1} + \sum_{k=1}^K g_k(s_{i,k,t} - \mu_{t-1}), \quad (50)$$

where  $g_k = \frac{\hat{\tau}_k}{\hat{\tau}_k + \tau_0}$ . Now, solving for  $\hat{y}$  is effectively solving for a fixed point, and given  $i$ 's information, their forecast of analyst  $j$ 's forecast is

$$\begin{aligned} F_{i,t}[\hat{y}_j] &= \mu_{t-1} + \sum_{k=1}^K \hat{w}_k F_{i,t}[s_{j,k,t} - \mu_{t-1}] \\ &= \mu_{t-1} + \sum_{k=1}^K \hat{w}_k (F_{i,t}[y_{t+1} - \mu_{t-1}] + F_{i,t}[u_{j,k,t}]) \\ &= \mu_{t-1} + \sum_{k=1}^K \hat{w}_k g_k(s_{i,k,t} - \mu_{t-1}). \end{aligned} \quad (51)$$



Further, we have their forecast of the consensus as

$$\begin{aligned}
F_{i,t}[\widehat{y}] &= \frac{1}{N} \sum_{j \neq i} F_{i,t}[\widehat{y}_j] + \frac{1}{N} \widehat{y}_i \\
&= \mu_{t-1} + \sum_k^K \left( \widehat{w}_k g_k (s_{i,k,t} - \mu_{t-1}) + \frac{1}{N} F_i[u_{i,k,t}] \right) \\
&= \mu_{t-1} + \sum_k^K \widehat{w}_k \left( g_k + \frac{1}{N} (1 - g_k) \right) (s_{i,k,t} - \mu_{t-1}).
\end{aligned} \tag{52}$$

Combining equations and match coefficients, we know  $\widehat{w}_k$  is given by

$$\widehat{w}_k = \frac{1}{1 + \lambda} g_k + \frac{\lambda}{1 + \lambda} \widehat{w}_k (g_k + \frac{1}{N} (1 - g_k)). \tag{53}$$

Solving for  $\widehat{w}_k$  gives the desired expression.  $\square$

## B.6 Top-down attention allocation

In this section, I provide a rational benchmark for attention choice. The model is similar to the one in Section 2.1, but now the firm's future earnings are driven by  $K$  structural components  $\theta_{k,t+1}$ ,

$$y_{t+1} = \theta_{1,t+1} + \theta_{2,t+1} + \cdots + \theta_{K,t+1}. \tag{54}$$

These structural components represent factors that matter for the firm's future earnings. One way to interpret these components is to recognize that, when the analyst makes a forecast about the firm's earnings, they do not treat these as a single variable. Instead, they first build a financial model to make projections about components of the income statement, and then aggregate these. In the analyst's financial model, the firm's earnings (net income) can thus be (roughly) represented as Earnings = Revenue – COGS – Operating Expense – Interest Expense –  $\cdots$ , which resembles the formulation in Eq. (54).

The analyst observes a set of  $K$  signals, each with respect to one structural component,

$$s_{k,t} = \theta_{k,t+1} + v_{k,t}; \quad v_{k,t} \sim N(0, \tau_k)$$

I assume now that the analyst has rational expectations,  $F_t[\cdot] = E_t[\cdot]$  and makes a one-time attention-allocation choice in the initial period  $t = 0$  to choose the precision  $\tau_k$ , in order to maximize their rationally expected lifetime utility, which is the discounted squared error minus an attention cost,

$$\min_{\tau} E_0[(y_{t+1} - E_t[y_{t+1}])^2] + C(\tau), \tag{55}$$

where  $C(\tau)$  denotes the cost of acquiring greater signal precision, which I assume is positive,

increasing in all  $\tau_k$ , and convex. I define the *relative* attention as

$$m_k := \frac{\tau_k}{\sum_{k'=1}^K \tau_{k'}}$$

and it can be mapped to the computation of topic weights in (12). The following proposition characterizes the analyst's rational attention choice.

**Proposition 5** (Optimal attention allocation for a rational analyst)

*Assuming all components are independent, i.e.  $\theta_k \sim N(\mu_k, \text{Var}(\theta_k))$ . The optimal precision  $\tau_k^*$  is characterized by*

$$\frac{1}{\text{Var}(\theta_k)^{-1} + \tau_k^*} = \frac{\partial C(\tau)}{\partial \tau_k^*}. \quad (56)$$

*Furthermore, the optimal attention is increasing in the importance of a structural component*

$$\frac{\partial m_k^*}{\partial \text{Var}(\theta_k)} > 0 \quad (57)$$

Intuitively, the effect of higher levels of attention increases the precision of signals about component  $k$ , and the rational agent will make a more accurate forecast if the analyst pays more attention to the component that matters more for future earnings.

**Proof of Proposition 5** Because the prior and the signal noises are independent Gaussian, the posterior variance of each component is  $\text{Var}(\theta_k | s_k) = (\text{Var}(\theta_k) + \tau_k)^{-1}$ , and the MSE of the forecast is

$$E[(y - E[y])^2] = \sum_{k=1}^K (\text{Var}(\theta_k)^{-1} + \tau_k)^{-1}. \quad (58)$$

Taking the first-order condition, we have the optimal precision  $\tau_k^*$  given by

$$\frac{1}{\text{Var}(\theta_k)^{-1} + \tau_k^*} = \frac{\partial C(\tau)}{\partial \tau_k^*}. \quad (59)$$

To study the relationship between optimal relative attention and component importance, notice that

$$m_k^* = \frac{\tau_k^*}{\sum_{k'} \tau_{k'}^*} = \frac{1}{1 + \tau_k^{*-1} \sum_{k' \neq k} \tau_{k'}^*}. \quad (60)$$

Since  $\tau_{k'}^*$  does not depend on  $\text{Var}(\theta_k)$  for  $k' \neq k$ , it is equivalent to prove  $\frac{\partial \tau_k^*}{\partial \text{Var}(\theta_k)} > 0$ . By the implicit function theorem, we have

$$\frac{\partial \tau_k^*}{\partial \text{Var}(\theta_k)} = -\frac{\partial F}{\partial \text{Var}(\theta_k)} / \frac{\partial F}{\partial \tau}, \quad (61)$$

where  $F = \frac{\partial C(\tau)}{\partial \tau_k^*} - \frac{1}{(Var(\theta_k)^{-1} + \tau_k)^2}$ . We have

$$\begin{aligned}\frac{\partial F}{\partial \tau_k^*} &= \frac{\partial^2 C(\tau)}{\partial \tau_k^*} + 2(Var(\theta_k) + \tau_k^*)^{-3} > 0 \\ \frac{\partial F}{\partial Var(\theta_k)} &= -\frac{2}{Var(\theta_k)^2}(Var(\theta_k) + \tau_k^*)^{-3} < 0,\end{aligned}\tag{62}$$

where the first inequality follows from the convexity assumption of  $C(\cdot)$ . Therefore,  $\frac{\partial m_k^*}{\partial Var(\theta_k)} > 0$ .  $\square$

## B.7 Details of TIGG regression estimation

### B.7.1 Dealing with high-dimensional embedding vectors

In the real data, for analyst  $i$  issuing forecast for firm  $j$  at time  $t$  for  $h$  horizon, I estimate the two-stage regression for each topic  $k$  with embedding vector  $X_{i,j,t}^k$  using the following specification

$$\begin{aligned}Rev_{i,t}[y_{j,t+h}] &= \lambda_k' X_{i,j,t}^k + \sum_{k' \neq k} \lambda_{k'}' X_{i,j,t}^{k'} + \eta_i + \eta_j + \eta_t + u_{i,j,t} \\ Error_{i,t}[y_{j,t+h}] &= \beta_{TIGG}^k \cdot \widehat{Rev}_{i,t}[\theta_{j,t+h}] + \sum_{k' \neq k} \xi_{k'}' X_{i,j,t}^{k'} + \eta_i + \eta_j + \eta_t + e_{i,j,t}\end{aligned}\tag{63}$$

where  $\widehat{Rev}_{i,t}[\theta_{j,t+h}] = \widehat{\lambda}_k' X_{i,j,t}^k$  and  $\eta_i, \eta_j, \eta_t$  denote analyst-, firm-, and time- (quarter-) fixed effects.

Given a potentially high-dimensional embedding vector  $X^k$  for topic  $k$ , the estimation environment mimics the one in a “many-IV” setting. One issue is that when the number of instruments goes large relative to the sample size, the first-stage estimates fits on the noise in the endogenous variable ( $\epsilon$  in  $Rev[y_i]$ ), therefore threatens the identification assumption. The typical solution is use the JIVE or split-sample (SSIV) rather than the least-squares estimator in the first stage to prevent overfitting on noise (e.g. Angrist and Krueger, 1995; Angrist et al., 1999). In this paper, I use the jackknife estimator in the first stage, where I estimate observation  $i$ ’s predicted revision using observations excluding  $i$  as follows:

$$Rev_i = \widehat{\lambda}_k(i)' X_i^k \quad ; \quad \widehat{\Lambda}(i) = (\mathbf{X}_{-i}' \mathbf{X}_{-i})^{-1} \mathbf{X}_{-i}' S_{-i},\tag{64}$$

where  $\widehat{\Lambda}(i) = [\widehat{\lambda}_1(i), \widehat{\lambda}_2(i), \dots, \widehat{\lambda}_K(i)]$  and  $\mathbf{X}_{-i}$  is the concatenated embedding matrix for observations other than  $i$ ,  $\mathbf{X}_{-i} = [X_{-i}^1, X_{-i}^2, \dots, X_{-i}^K]$ .

### B.7.2 Pre-test of instrument strength and inference

When there is substantial noise or imperfection in the embedding vectors, the inference might suffer from the weak-instrument problem. To assess the instrument strength, I compute the  $F$ -statistic

proposed in Mikusheva and Sun (2022) as a pre-test for weak identification in the many-instrument setting. The statistic is computed as follows:

$$\tilde{F} = \frac{1}{\sqrt{DK}\sqrt{\hat{\Upsilon}}} \sum_{i=1}^N \sum_{j \neq i} P_{ij} S_i S_j, \quad (65)$$

where  $P$  is the projection matrix  $P = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$  of the concatenated embedding matrices,  $M = I - P$ ,  $\hat{\Upsilon} = \frac{2}{DK} \sum_i \sum_{j \neq i} \frac{P_{ij}^2}{M_{ii}M_{jj} + M_{ij}^2} S_i M_i S S_j M_j S$  and  $D$  is the dimensionality of the embedding vectors. Mikusheva and Sun (2022) propose cutoffs of 4.14 for an asymptotic size of 15% and 9.98 for an asymptotic size of 5% if the JIVE-Wald test is used. I compute this F-stat for each individual topic  $k$  and find that the  $\tilde{F}$  statistics are comfortably higher than these cutoffs, and, I thus proceed with the JIVE-Wald inference test. The JIVE-Wald test is constructed by estimating the cross-fit estimator of the JIVE variance derived in Chao et al. (2012) and analyzed in Mikusheva and Sun (2022), namely

$$\hat{V} = \frac{\sum_{i=1}^N \left( \sum_{j \neq i} P_{ij} X_j \right)^2 \frac{\hat{e}_i M_i \hat{e}}{M_{ii}} + \sum_{i=1}^N \sum_{j \neq i} \tilde{P}_{ij}^2 (M_i S \hat{e}_i) (M_j S \hat{e}_j)}{\left( \sum_{i=1}^N \sum_{j \neq i} P_{ij} X_i X_j \right)^2}, \quad (66)$$

where  $\hat{e}_i$  is the residual in the second stage and  $\tilde{P}_{ij}^2 = \frac{P_{ij}^2}{M_{ii}M_{jj} + M_{ij}^2}$ .

## C Additional results on topic-weight variation

**Cross-sectional variation.** I also find substantial cross-sectional variation in topic weights across firms. Figure OA.6 presents the total  $R^2$  from fixed-effect regressions with time-, industry-, and stock-fixed effects. On average, time and industry effects explain only 26% of the variation in topic weights, while including stock-fixed effects raises the  $R^2$  to 42%. This indicates that there is considerable within-industry heterogeneity in the information on which analysts choose to focus.

Second, I examine how information focus varies with firm characteristics. Figure OA.7 reports the regression coefficients of  $\overline{m}_{j,t}^k$  for each topic  $k$  on the firm’s book-to-market and debt-to-market ratios. I find that firms with higher book-to-market ratios are more likely to be associated with discussions of cost, debt, cash, the macroeconomy, and risk, and less likely to be linked with discussions of M&A, business strategies, R&D, and product-related topics. The greater focus on topics related to intangibles for growth firms, coupled with a relatively muted focus on profitability, suggests that their high valuations are primarily supported by forward-looking growth narratives rather than past earnings performance.

**Forecast horizons.** The literature documents that analysts exhibit different behavioral patterns when forecasting short-versus long-term targets (e.g. Bouchaud et al., 2019; Van Binsbergen et al., 2023; De Silva and Thesmar, 2024; Bordalo et al., 2024c). This naturally raises the question: Do analysts focus on different types of information when forming forecasts at different horizons?

Figure OA.8 shows the differences in topic weights when analysts issue long-term (more than three years) versus short-term (one to two years) EPS forecasts. I find that, when making long-term forecasts, analysts focus significantly more on a firm’s R&D information. They also place greater emphasis on legal, industry-landscape, product-development, and M&A topics. In contrast, when forecasting short-term EPS, analysts devote greater attention to profitability, cost, and financial conditions such as cash, debt, and asset management. Overall, these findings suggest that analysts focus on different types of information depending on the forecast horizon, and that these differences are consistent with the relevance of different value drivers at different horizons.

**Upward versus downward revision.** Figure OA.9 shows the differences in topic weights when analysts revise up or revise down their forecasts. I find that when analysts revise their forecasts upward, they mainly focus on profitability information. In contrast, when they revise their forecasts downward, they pay more attention to a firm’s intangible information, such as R&D, corporate leadership, and customer demand, as well as information related to financial conditions, including debt, cash, and funding. These differences are more pronounced for longer-term forecasts.

**Cross-analyst variation.** Finally, I explore the variation in attention across analysts. Figure OA.10 shows the total  $R^2$  from fixed-effect regressions including time-, stock-, and analyst-fixed effects. I find that within-firm attention variation is low: on average, time-stock-fixed effects explain about half of the variation in topic weights, while adding analyst-fixed effects increases the total  $R^2$  by only 3%. This suggests that analyst-specific attention choices are limited.

**Topic concentration and processing capacity constraint.** I find that analyst-level attributes, such as experience and stock coverage, help explain topic concentration, measured by the Herfindahl-Hirschman Index (HHI) of the 28 topic weights. Table OA.4 reports regressions of topic HHI on analyst experience and coverage. The results show that analysts with more experience (measured by longer tenure in EPS forecasting) tend to have more evenly distributed topic weights, whereas those covering more stocks exhibit higher topic concentration. These patterns suggest the presence of a constraint on processing capacity: as an analyst covers more stocks, it becomes more challenging to keep track of all relevant topics, leading to a more concentrated focus. In contrast, experienced analysts appear less constrained and are better able to analyze a broader range of topics in greater detail.

## D Evidence of top-down attention allocation.

The results so far demonstrate that there is significant variation in topic weights across analyst reports. This provides a useful setting to examine the attention choices made by analysts. In this section, I present evidence that the observed variation in topic weights is qualitatively consistent with the predictions of the rational inattention model described in Section B.6.

The key prediction of the rational inattention model is that analysts pay more attention to topics with greater predictive power for future earnings, that is,  $\frac{\partial m_k^*}{\partial \text{Var}(\theta_k)} > 0$ . I implement this idea by estimating the following regression:

$$\Delta y_{j,t+h} = \beta_1 Z_{j,t}^k + \beta_2 Z_{j,t}^k \times \mathbf{1}\{\text{High Attn}\}_{j,t} + \epsilon_{j,t+h}, \quad (67)$$

where  $\Delta y_{j,t+h}$  is the  $h$ -year growth in firm  $j$ 's earnings,  $Z_{j,t}^k$  is an observable that captures the current realization of component  $k$  (serving as a proxy for  $\theta_k$ ), and  $\mathbf{1}\{\text{High Attn}\}$  is an indicator for firm-quarter observations in which analysts devote above-median attention to topic  $k$ . The magnitude and significance of  $\beta_2$  thus provide a measure of how the predictive power of component  $k$  for future earnings growth differs between high- and low-attention subsamples.

**Attention to the macroeconomy.** I first study the variation in analysts' attention to the macroeconomy using the change in log industrial production as a measure of the current macroeconomic state. The results in Table OA.5 show that when analysts devote greater attention to

macroeconomic topics in their reports, changes in log industrial production in the current quarter more strongly predict future one-year EPS growth at the firm level. This difference in predictability is both statistically and economically significant: the predictive coefficients more than double in the high-macro-attention subsample. These results suggest a top-down allocation of attention consistent with the predictions of a rational inattention model—analysts actively pay more attention to macroeconomic topics when firms’ future earnings become more sensitive to current macroeconomic conditions.

**Attention to firm-level features.** I also explore whether firm-level characteristics exhibit differential predictability of future earnings growth when analysts shift their weighting of topics. Since it is difficult to identify a numerical variable for many intangible topics, such as legal, M&A, or marketing, I focus on topics where the underlying conditions can be better proxied by accounting-based stock characteristics. In particular, I examine the profitability, cash, debt, and R&D topics, and use a broad set of stock-level characteristics constructed by [Jensen et al. \(2023\)](#) to proxy for the unobserved fundamental state ( $\theta_k$ ) of these. Specifically, the proxies include 87 characteristics for profitability, 53 for debt, 38 for cash, and 3 for R&D.<sup>54</sup>

Table [OA.6](#) reports the results of earnings predictability conditional on analysts’ attention levels. In this table, I predict future 1-year or 5-year earnings using a “composite” characteristic for each topic, constructed as the average across all individual characteristics associated with that topic. In Panel(A), I find that when analysts increase their discussion of debt in their reports, realized debt-related stock characteristics (such as debt-to-market ratio and debt issuance) predict future 1-year and 5-year EPS growth more strongly, as evidenced by significantly larger predictive coefficients in the high-attention subsample. Similarly, Panels (B) and (C) show that when analysts increase their discussions of the cash or R&D topics, realized cash-related characteristics (such as cash-to-market) and R&D-related characteristics (such as R&D expenditure-to-market) become more predictive of future 5-year EPS growth. The stronger predictive power of R&D characteristics for 5-year rather than 1-year EPS growth, together with the earlier finding that analysts devote more attention to the R&D topic when making long-term forecasts (Figure [OA.8](#)), further suggests that analysts are purposefully changing their topic weights to focus on the more important aspects of a firm’s fundamentals.

As a robustness check, I run the prediction regression, Eq. (67), using individual stock characteristics. Figure [OA.11](#) presents the histograms of the individual predictive coefficients. Consistent with the results from the “composite” regression, I find that when analysts increase their attention to the cash and debt topics, the distribution of predictive coefficients from individual stock characteristics shifts to the left, indicating that cash and debt information becomes more relevant for forecasting future earnings.

---

<sup>54</sup>See Table [OA.14](#) for the list of characteristics used in the analysis.

However, for the profitability topic, I find the opposite effect: when analysts talk more about profitability in their reports, the realized profitability characteristics actually have lower predictive power for future earnings growth. This suggests that, in addition to top-down attention allocation, analysts' topic discussions are influenced by other forces.

In summary, I demonstrate that there is significant variation in the topic weights in analyst reports, and this is broadly consistent with rational allocation of inattention. However, a definitive test of rational inattention models, such as the one in Section D, requires knowledge of the true data-generating process as well as the cognitive costs faced by the agent, which remains a challenging task. Moreover, there is some evidence suggesting that topic allocations may also be influenced by other non-rational forces yet to be explored.

## E Variables for institutional incentives

I compute the following variables that are related to institutional distortions in the sell-side analyst industry identified in the literature.

**Issuance affiliation.** Following prior work (e.g. Lin and McNichols, 1998; Michaely and Womack, 1999; Malmendier and Shanthikumar, 2014), I define analysts to be affiliated with a public firm through issuance if their brokerage was the lead or co-underwriter of an initial public offering of the firm on which the analyst is reporting during the past five years, or of a seasoned equity offering of the firm during the past two years, using the SDC New Issues database.

**M&A affiliation.** Similar to the issuance affiliation, I define analysts as affiliated with a public firm through issuance if their brokerage participates in an M&A deal involving the public firm in the past 5 years. I include both advisors for the acquirer and the target.

**Investment-banking pressure.** Following Ljungqvist et al. (2006), Ljungqvist et al. (2007) and Malmendier and Shanthikumar (2014), I calculate investment-banking pressure of analysts in brokerage firm  $j$  covering public firm  $k$  in year  $t$  as follows: I use SDC New Issues data and determine whether  $k$  extended an underwriting mandate to broker  $j$ . I then accumulate the principal amounts from the deals that broker  $j$  managed for company  $k$  in the preceding five years and divide that by the total file amount of  $k$ 's deals during the same period.

**Broker reputation capital.** Following Ljungqvist et al. (2006), Ljungqvist et al. (2007) and Malmendier and Shanthikumar (2014) I compute a broker's reputational capital as the underwriter quality (the broker's share in the IPO market) to a broader set of securities. I use the SDC New Issues data to calculate a broker's market share as the amount of equity it raised as the lead



underwriter for its clients in the preceding calendar year divided by the total equity raised by all issuers in that year. In the case of more than one lead underwriter, we assign one  $n$ th of the amount raised to each of the  $n$  underwriters.

**Broker loyalty index.** Following [Ljungqvist et al. \(2006\)](#), [Ljungqvist et al. \(2007\)](#) and [Malmendier and Shanthikumar \(2014\)](#), I compute the broker loyalty index for broker  $j$  in year  $t$  as the ratio of the number of firms that used broker  $j$  both in their penultimate and in their most recent deals to the number of firms that used broker  $j$  in their penultimate deal.

**Institutional ownership ratio.** I compute a stock's institutional ownership ratio as the fraction of shares outstanding owned by institutional investors as recorded in the 13F SEC filings in each quarter.

**Brokerage size.** Brokerage size is the number of analysts employed by the brokerage in a given year.

# F Additional Figures

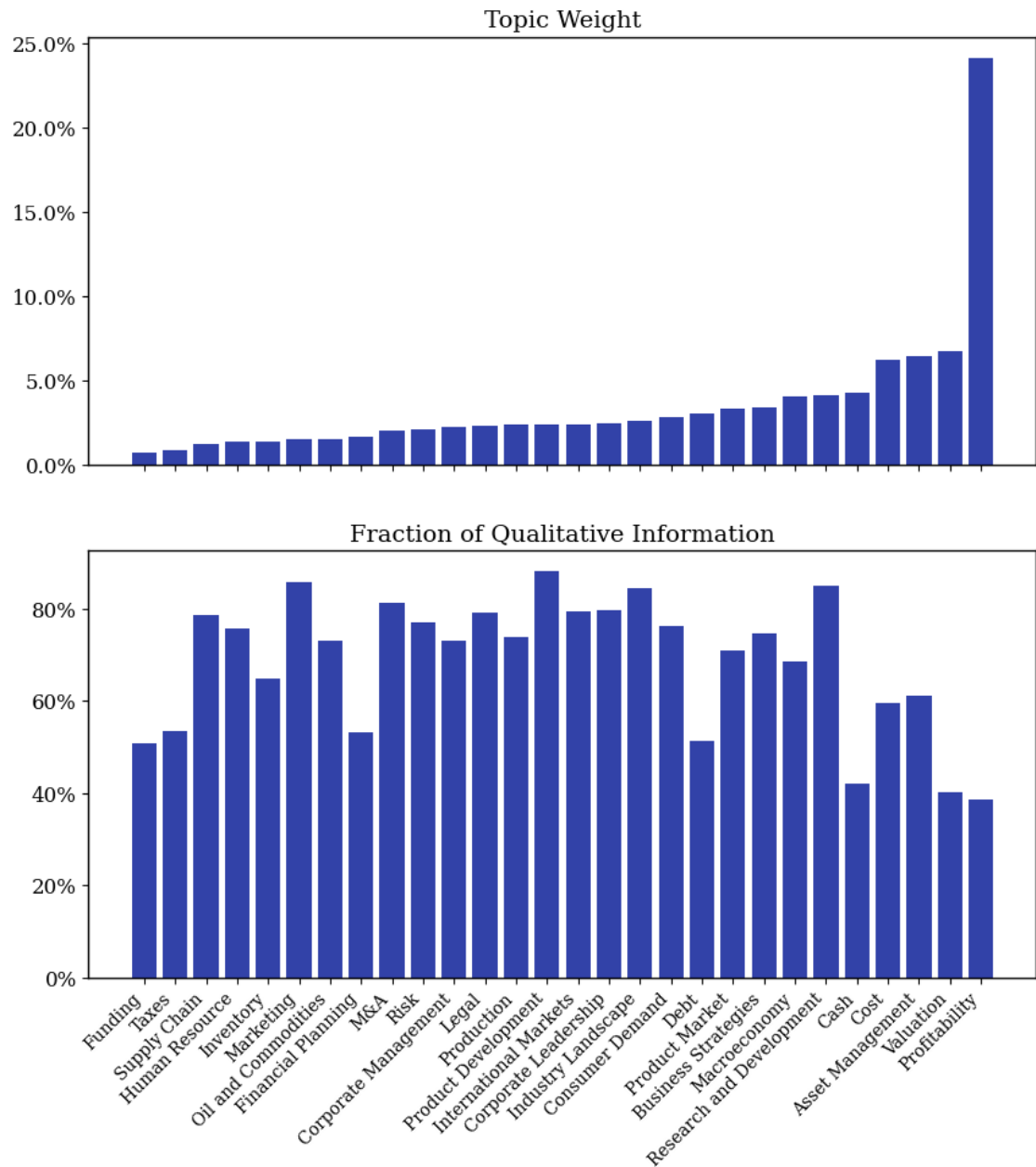


Figure A.2: Aggregate topic weight and fraction of qualitative information

*Notes:* The top panel shows the full-sample average topic weight, defined as the share of all *Nouns* that are classified as belonging to each topic. The bottom panel shows the fraction of information tuples that are classified as “qualitative” in the full sample.

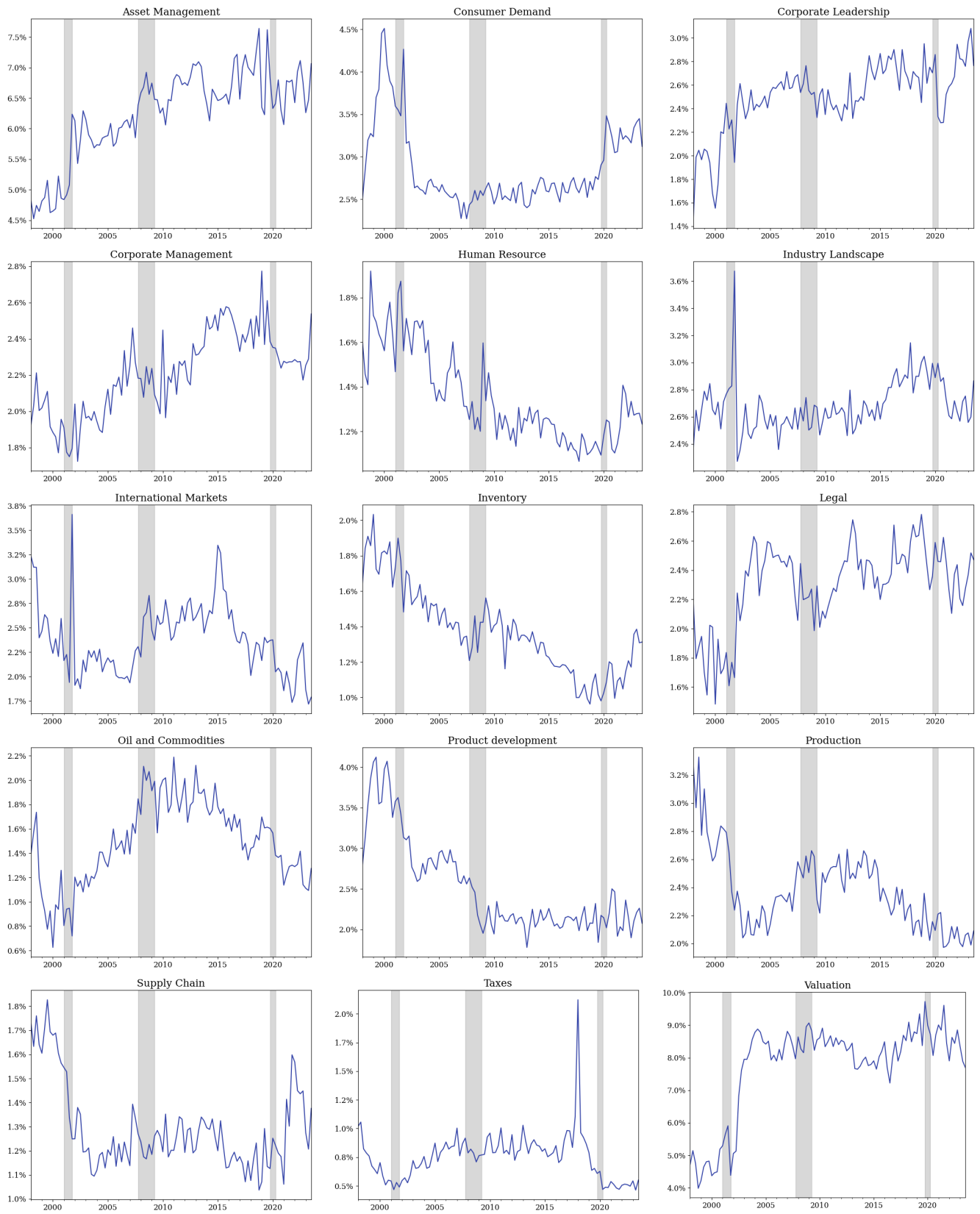


Figure A.3: Additional times series variation in topic focus

*Notes:* The shaded region denotes recession periods identified by the NBER.

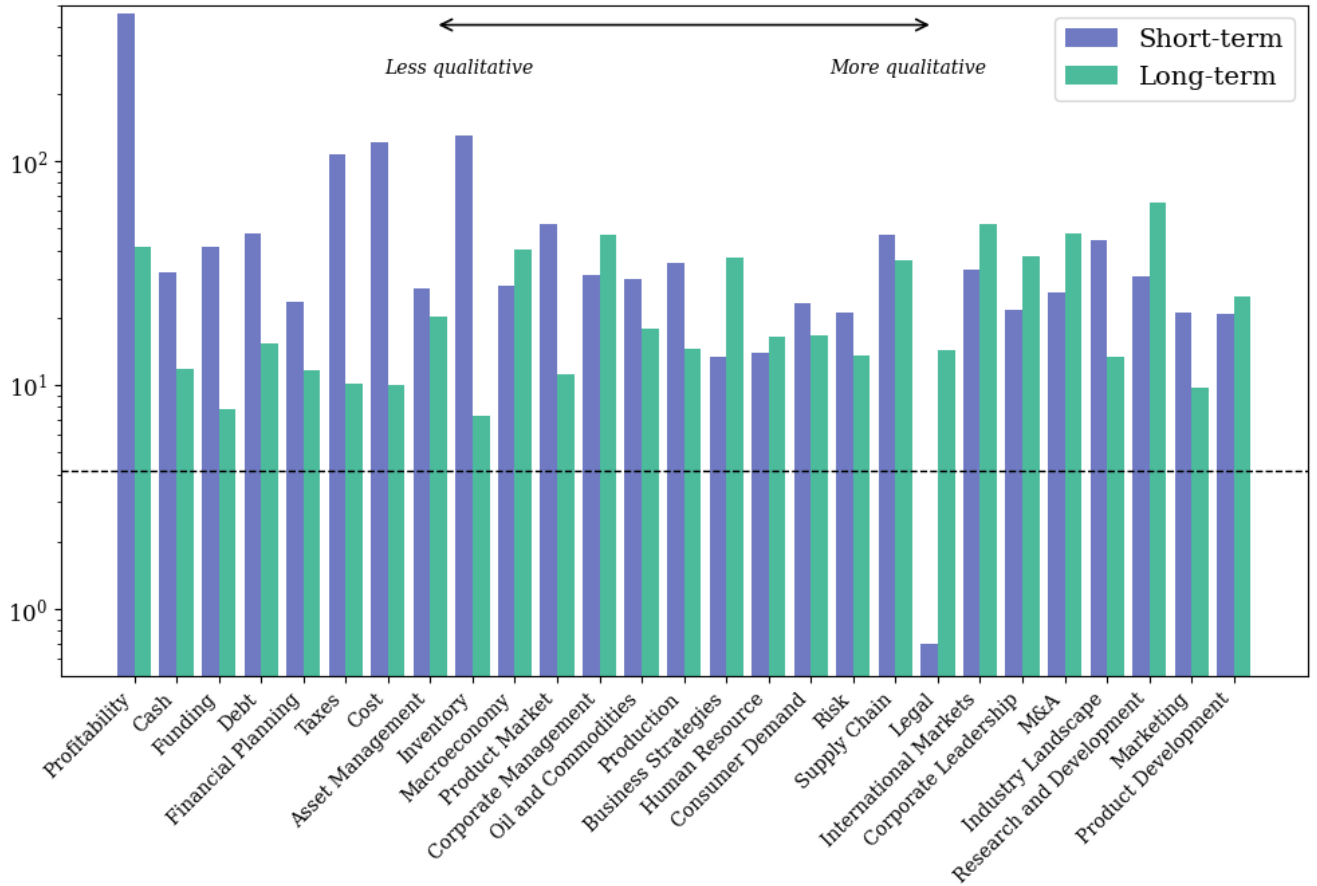


Figure A.4: First-stage F-statistics of individual topics in TICG regression

*Notes:* This figure reports the first-stage F-statistics for each individual topic following Mikusheva and Sun (2022). For each topic  $k$ , I estimate the first-stage regression

$$Rev_{i,t}[y_{j,t+h}] = \lambda'_k X_{i,j,t}^k + \sum_{k' \neq k} \lambda'_{k'} X_{i,j,t}^{k'} + \eta_i + \eta_j + \eta_t + u_{i,j,t}$$

where  $X_{i,j,t}^k$  denotes the topic-level text embedding for topic  $k$  in report written by analyst  $i$  at time  $t$  covering firm  $j$ , and  $\eta_i, \eta_j, \eta_t$  denote analyst-, firm-, and time (quarter)-fixed effects. The exact formula for Mikusheva and Sun (2022) adjusted F-statistics is given in Appendix B.7.2. The dashed line represents the weak many instrument threshold (4.14) proposed in Mikusheva and Sun (2022).

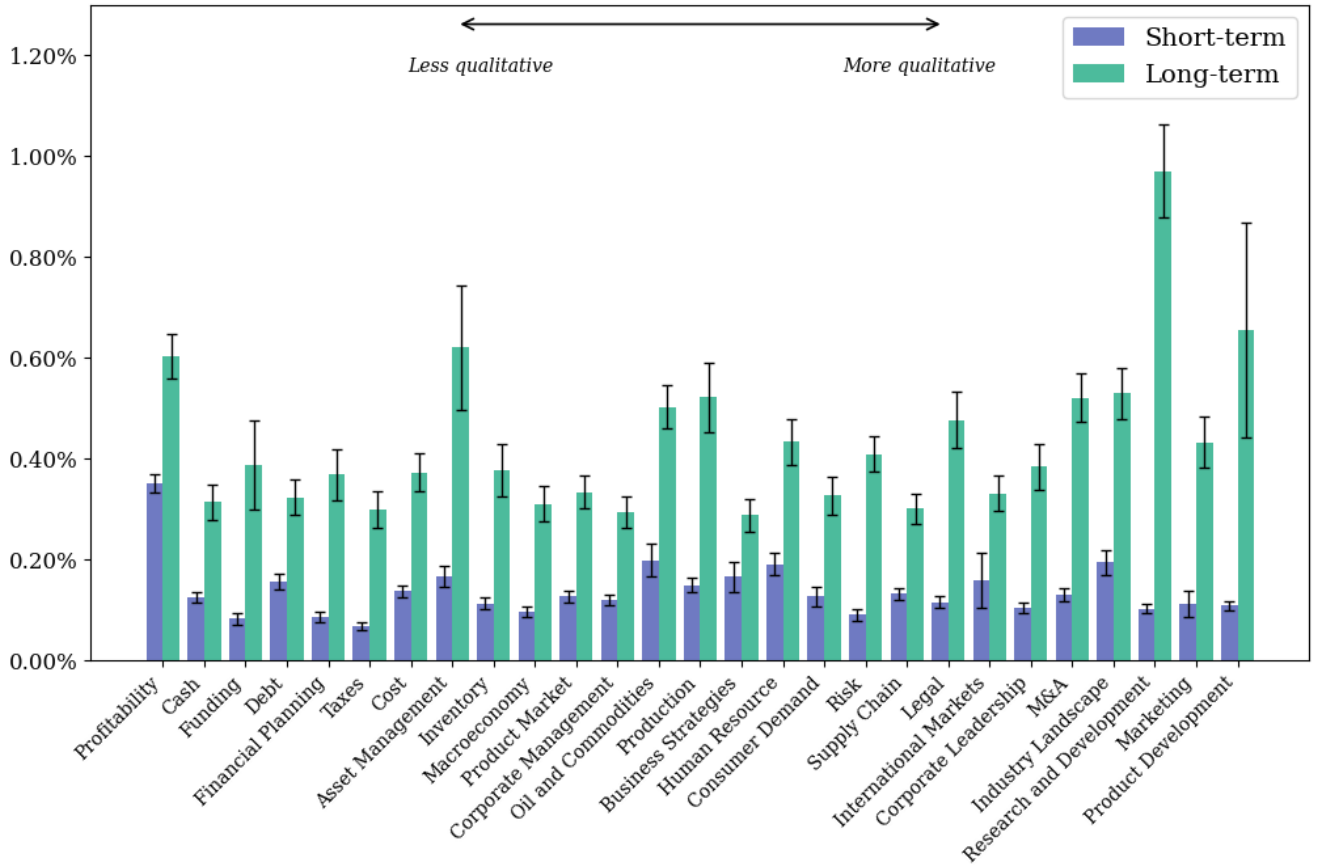


Figure A.5: Variance decomposition of forecast errors into different topics

*Notes:* This figure reports the variance decomposition of forecast errors using embedding vectors of different topics. Specifically, I estimate

$$Error_{i,t}[y_{j,t+h}] = \sum_{k=1}^{27} \xi_k' X_{i,j,t}^k + \eta_i + \eta_j + \eta_t + \epsilon_{i,j,t}$$

where  $X_{i,j,t}^k$  denotes the topic-level text embedding for topic  $k$  in report written by analyst  $i$  at time  $t$  covering firm  $j$ , and  $\eta_i, \eta_j, \eta_t$  denote analyst-, firm-, and time (quarter)-fixed effects. The variance contribution is then computed as  $\frac{Cov(\hat{\xi}_k' X_{i,j,t}^k, Error_{i,t}[y_{j,t+h}])}{Var(Error_{i,t}[y_{j,t+h}])}$  for topic  $k$  by regressing  $\hat{\xi}_k' X_{i,j,t}^k$  onto  $Error_{i,t}[y_{j,t+h}]$ . Error bars represent 95% confidence intervals, with standard errors two-way clustered at the analyst-firm and firm-quarter levels.

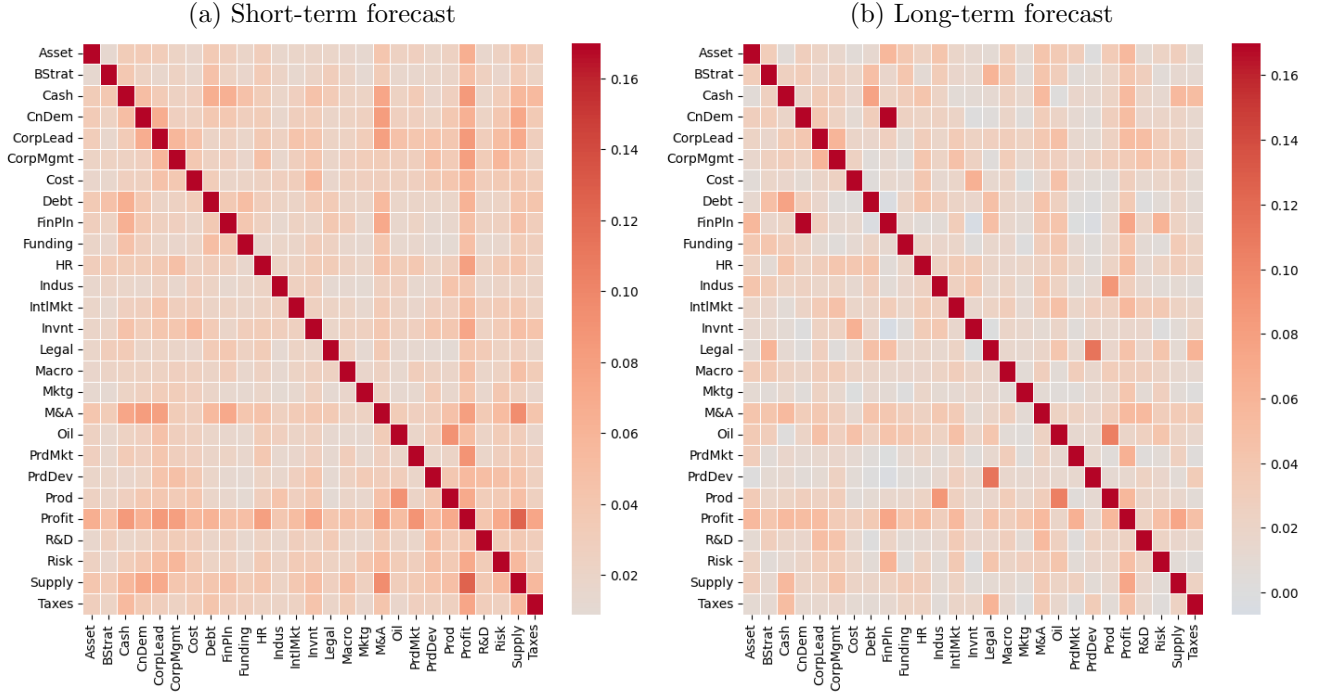


Figure A.6: Correlation matrix of topic-level revisions

*Notes:* This figure shows the correlation matrix of topic-level revisions. For each topic, I estimate

$$Rev_{i,t}[y_{j,t+h}] = \lambda'_k X_{i,j,t}^k + \eta_i + \eta_j + \eta_t + e_{i,j,t}$$

where  $X_{i,j,t}^k$  denotes the topic-level text embedding for topic  $k$  in report written by analyst  $i$  at time  $t$  covering firm  $j$ , and  $\eta_i, \eta_j, \eta_t$  denotes analyst-, firm-, and time (quarter)-fixed effects. The correlation matrix denotes the correlation of  $\hat{\lambda}_k' X_{i,j,t}^k$ 's. Predictions are computed using jackknife estimator.

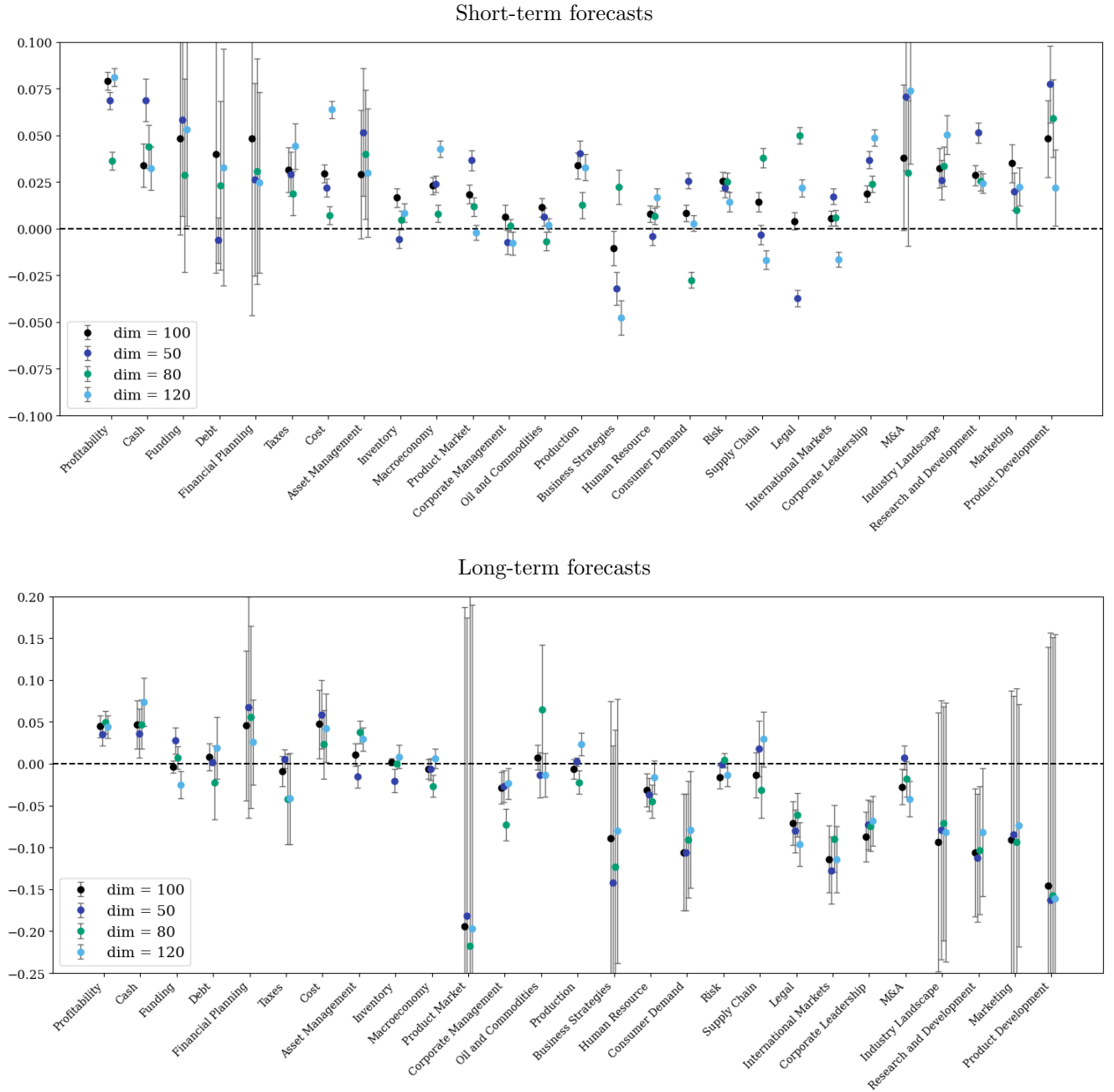


Figure A.7: TICG coefficients: Robustness to embedding dimensions

*Notes:* This figure shows the TICG-regression coefficient estimates of each topic using different number of top principal components of embedding vectors. See the note to Figure 5 for additional details.

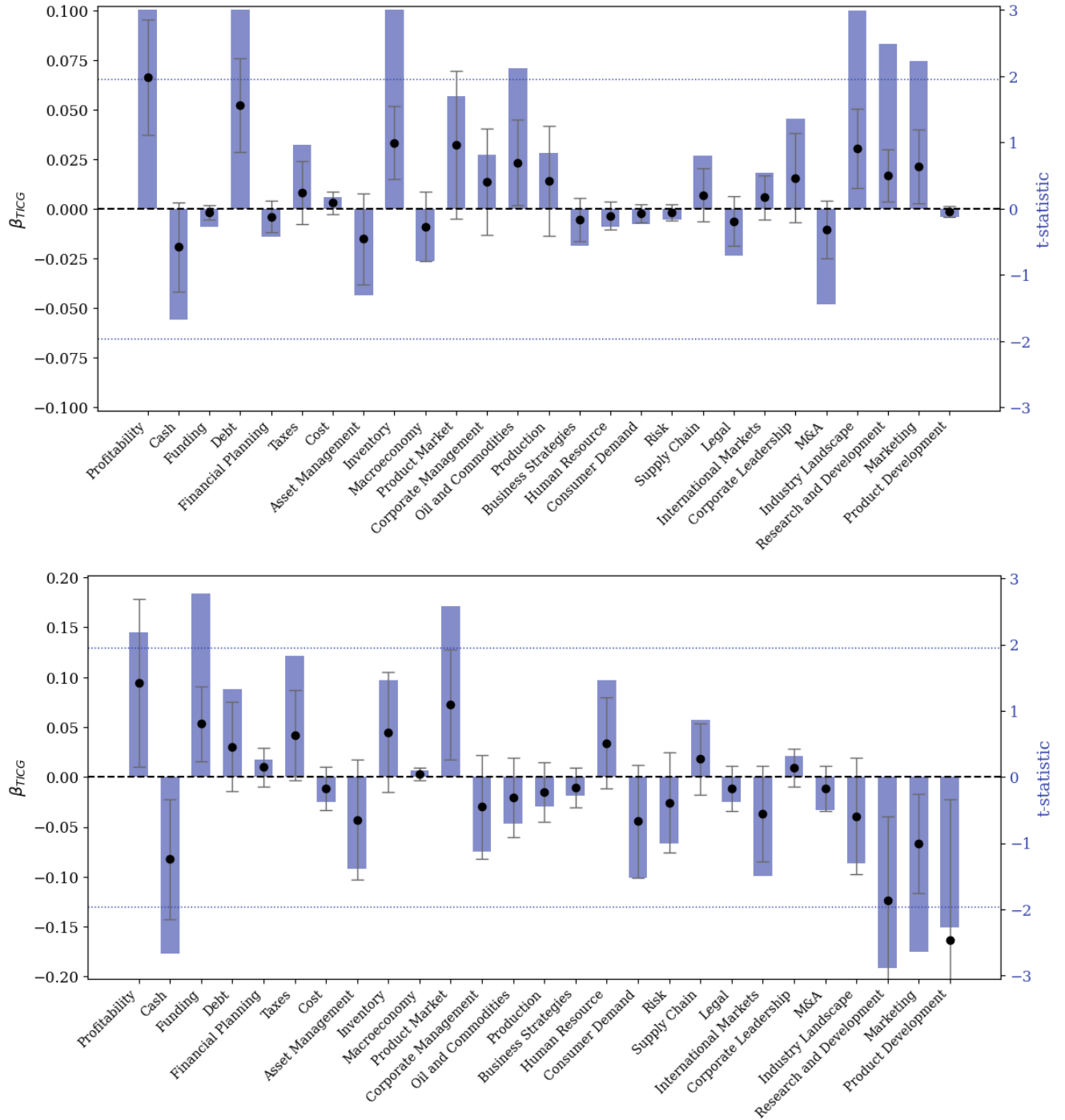


Figure A.8: Topic weight-interacted CG-regression coefficients

*Notes:* This figure shows the topic weight-interacted CG-regression coefficient estimates for each topic.

$$y_{j,t+h} - F_{i,t}[y_{j,t+h}] = \sum_k \beta_k \times m_{i,j,t}^k \times Rev_{i,t}[y_{j,t+h}] + \epsilon_{i,j,t},$$

where  $m_{i,j,t}^k$  is the weight of topic  $k$  in reports from analyst  $i$  covering stock  $j$  at time  $t$  as defined in Eq. 12. Standard errors are clustered at analyst-firm and firm-time (quarter) levels. Analyst-, firm- and time-fixed effects are included. Gray error bars indicate 95% confidence intervals. The blue bars indicate the  $t$ -statistics (right axis) and the dashed lines indicate the 5% significance level.



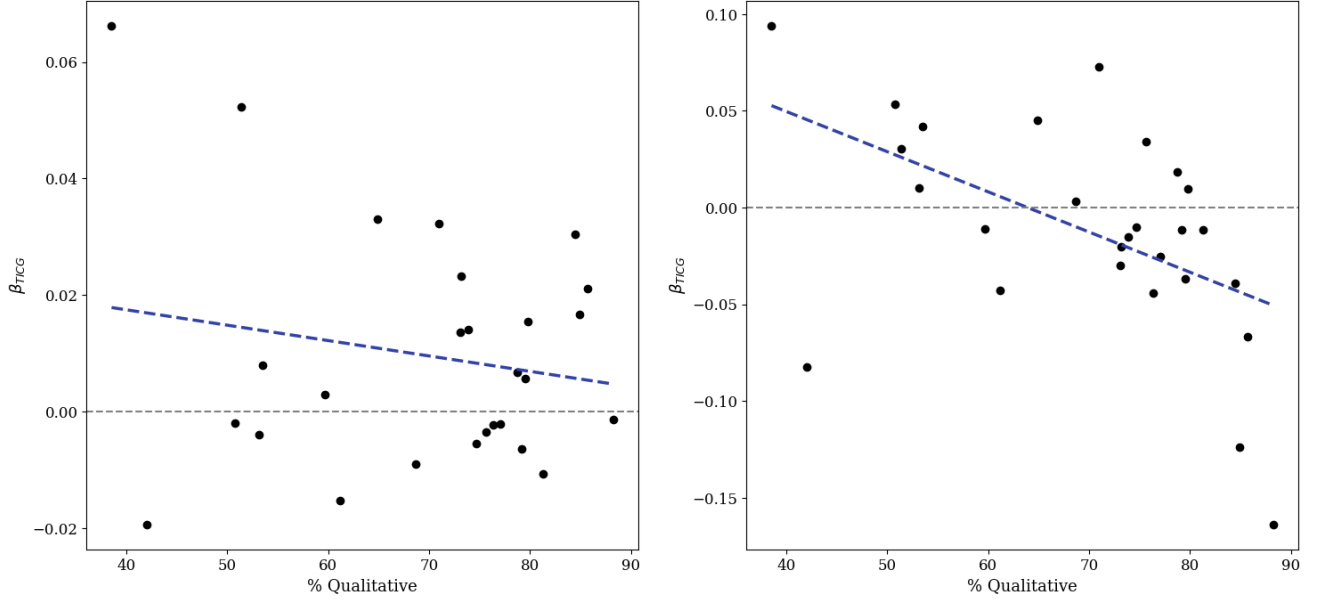


Figure A.9: Topic weight-interacted CG coefficients versus fraction of qualitative information tuples

*Notes:* The scatter plots show the relationship between attention-interacted CG regression coefficients and the share of qualitative information tuples in a topic.

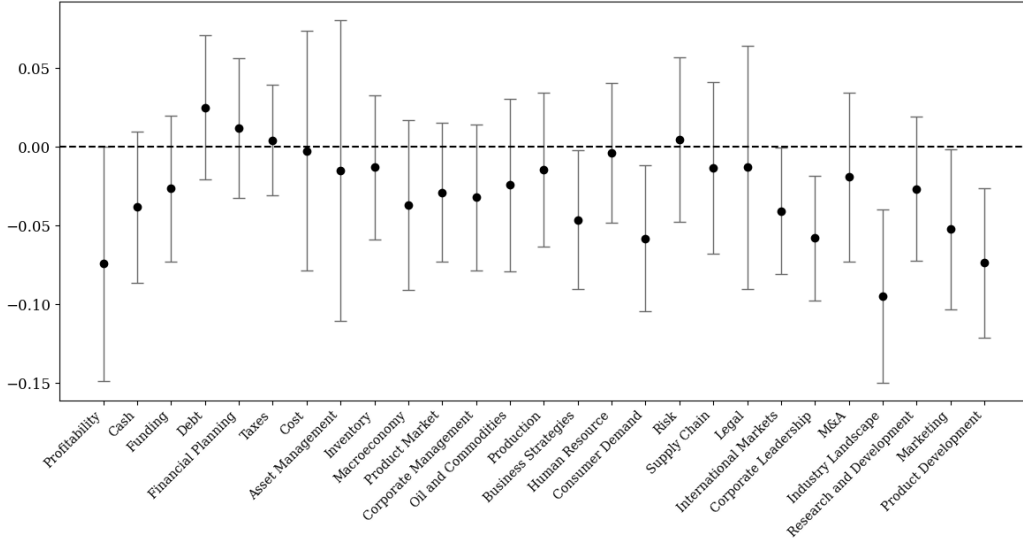


Figure A.10: TICG regression interacted with topic-level fraction of qualitative information tuples

*Notes:* This plot shows the  $\beta_k^{qual}$  in the second stage of the TICG regression for different topics,

$$y_{j,t+h} - F_{i,t}[y_{j,t+h}] = \sum_k \beta_k \widehat{Rev}_{i,t}^k[y_{j,t+h}] + \sum_k \beta_k^{qual} \widehat{Rev}_{i,t}^k[y_{j,t+h}] \times Qual\_Frac_{i,j,t}^k + \epsilon_{i,j,t},$$

where  $\widehat{Rev}_{i,t}^k[y_{j,t+h}]$  is the same predicted long-term forecast revisions with respect to topic  $k$  information that goes into the baseline TICG regression in Section 5. Error bars show the 90% confidence intervals.

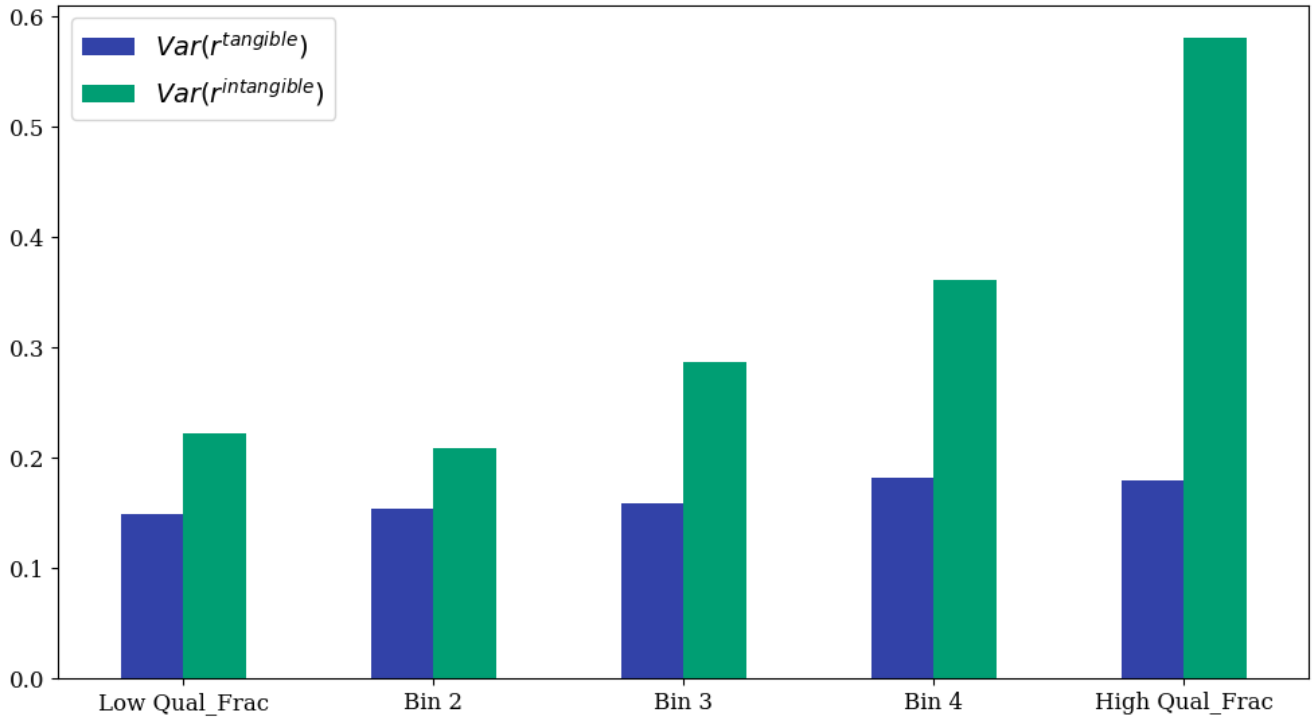


Figure A.11: Variance of tangible and intangible returns with information qualitiveness

*Notes:* This plot shows how variances in tangible and intangible returns relate to the information qualitiveness in analyst reports. I compute tangible returns as the book return (changes in log book equity) from year  $t - 1$  to year  $t$ ,  $r_{j,t}^{tangible} = \log(BE_{j,t}) - \log(BE_{j,t-1})$ . The intangible return is computed as the residual in the cross-sectional regression  $r_{j,t} = \gamma_0 + \gamma_1 \log(BE\_ME_{j,t-1}) + \gamma_2 r_{i,t}^{tangible} + v_{j,t}$  and  $r_{j,t}^{intangible} = \hat{v}_{j,t}$ . I then compute the variance of  $r_{j,t}^{tangible}$  and  $r_{j,t}^{intangible}$  in five groups of stock-year observations based on the stock's  $Qual\_Frac$  in that year. Variances are computed after controlling for stock- and year-fixed effects.

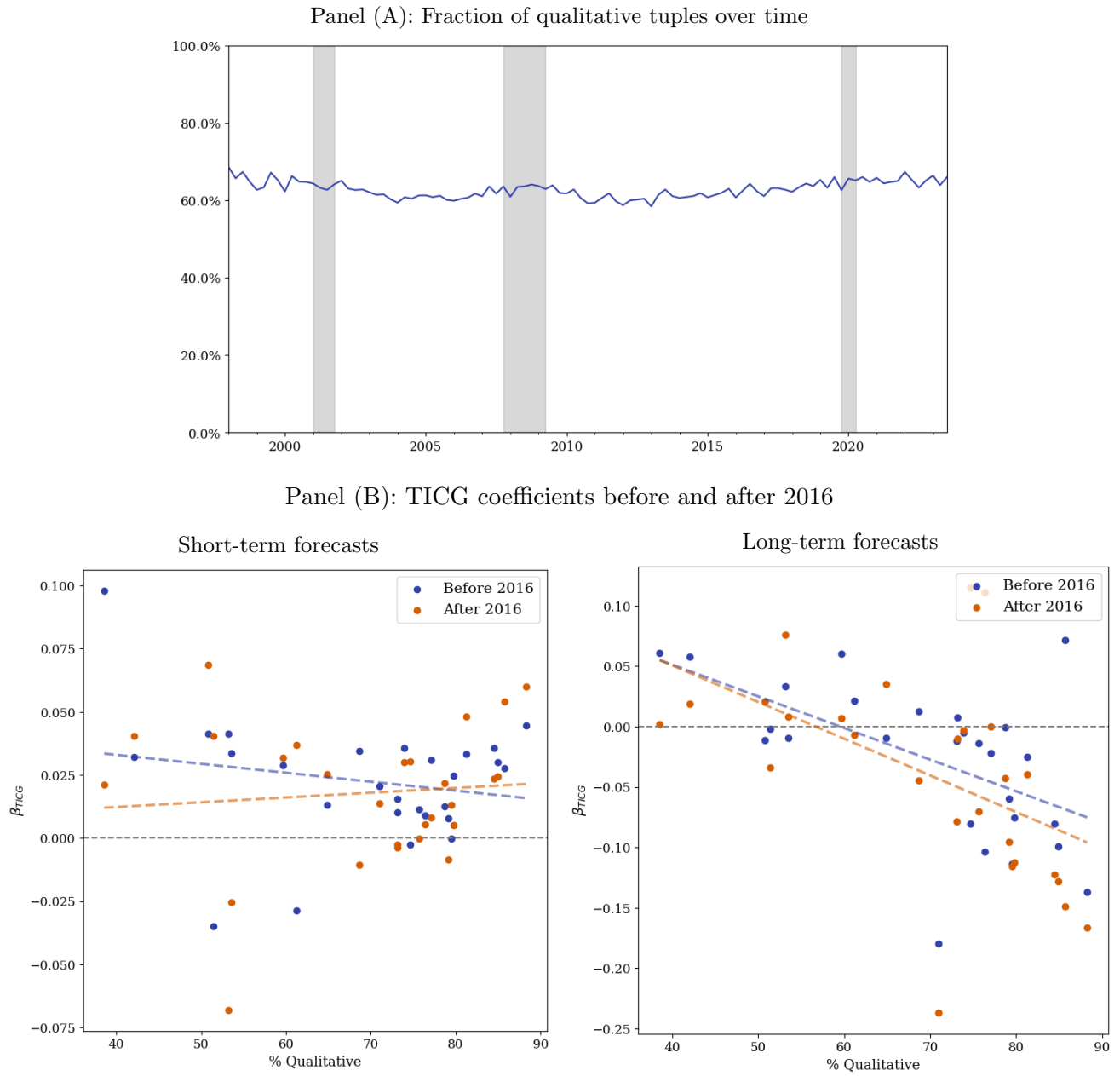
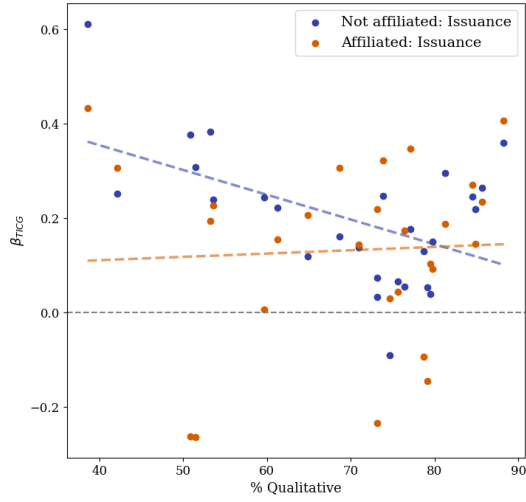


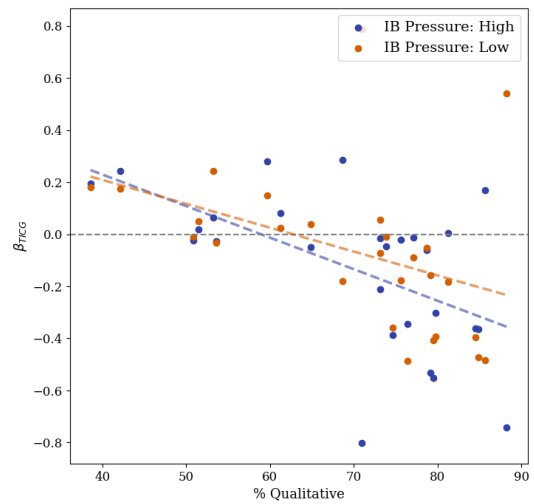
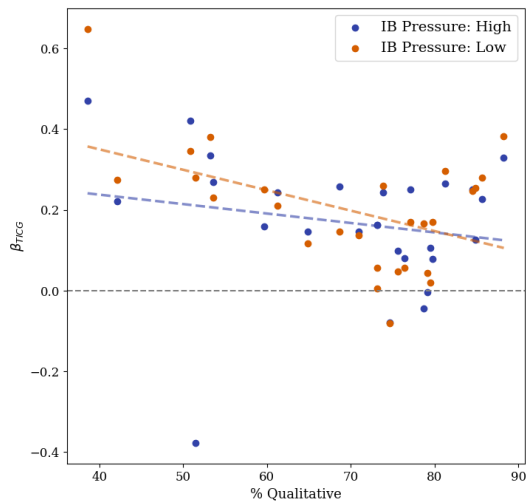
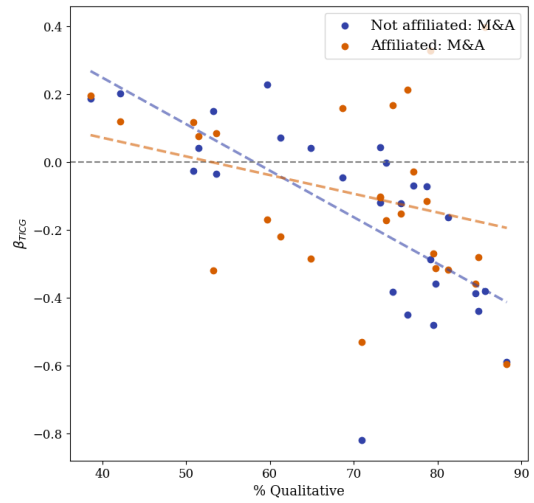
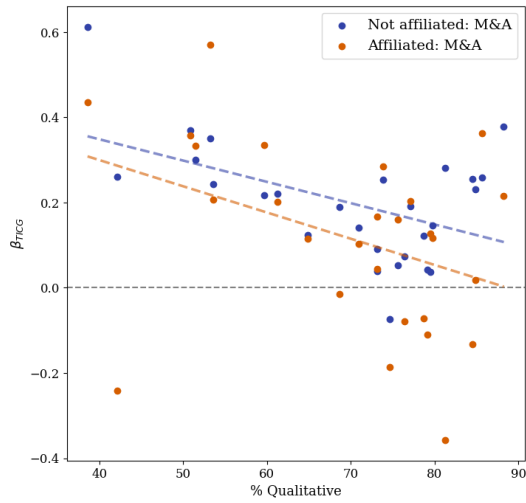
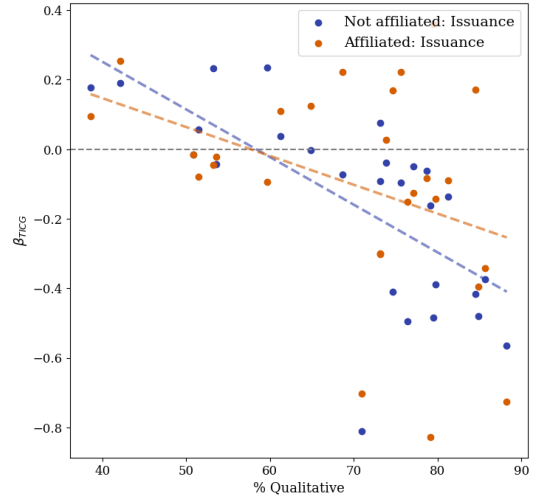
Figure A.12: TIGC coefficients versus share of qualitative-information tuples: Before and after 2016

*Notes:* Panel (A) shows the average fraction of qualitative information tuples in analysts' reports over time. Panel (B) reproduces Figure 6 but uses samples before and after 2016.

Short-term forecasts



Long-term forecasts



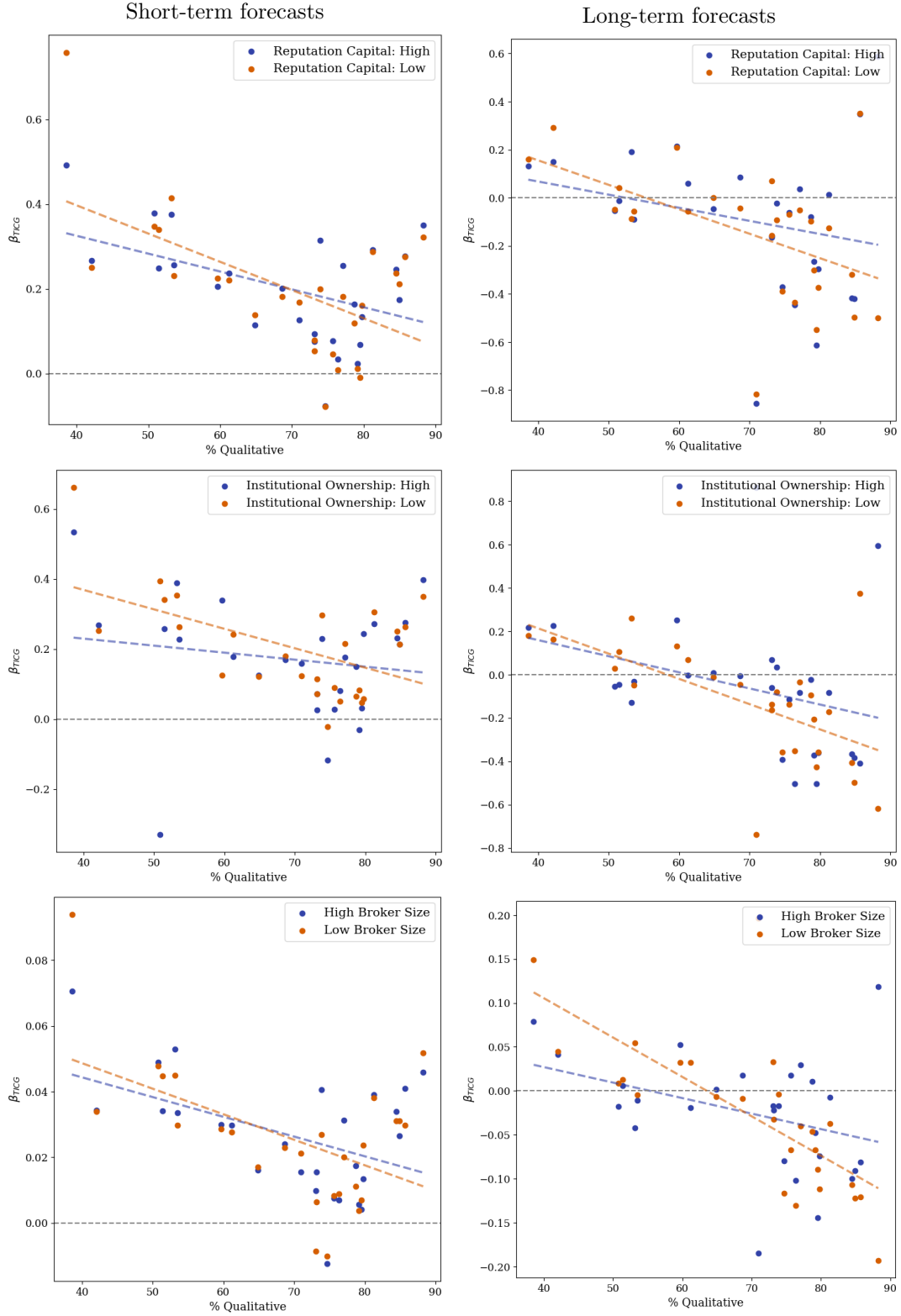


Figure A.14: TICG coefficients versus fraction of qualitative information tuples: Brokerage features

*Notes:* The scatter plots show the relationship between TICG-regression coefficients and the share of qualitative information tuples in a topic, in subsamples split based on institutional-incentive variables computed in Appendix E.

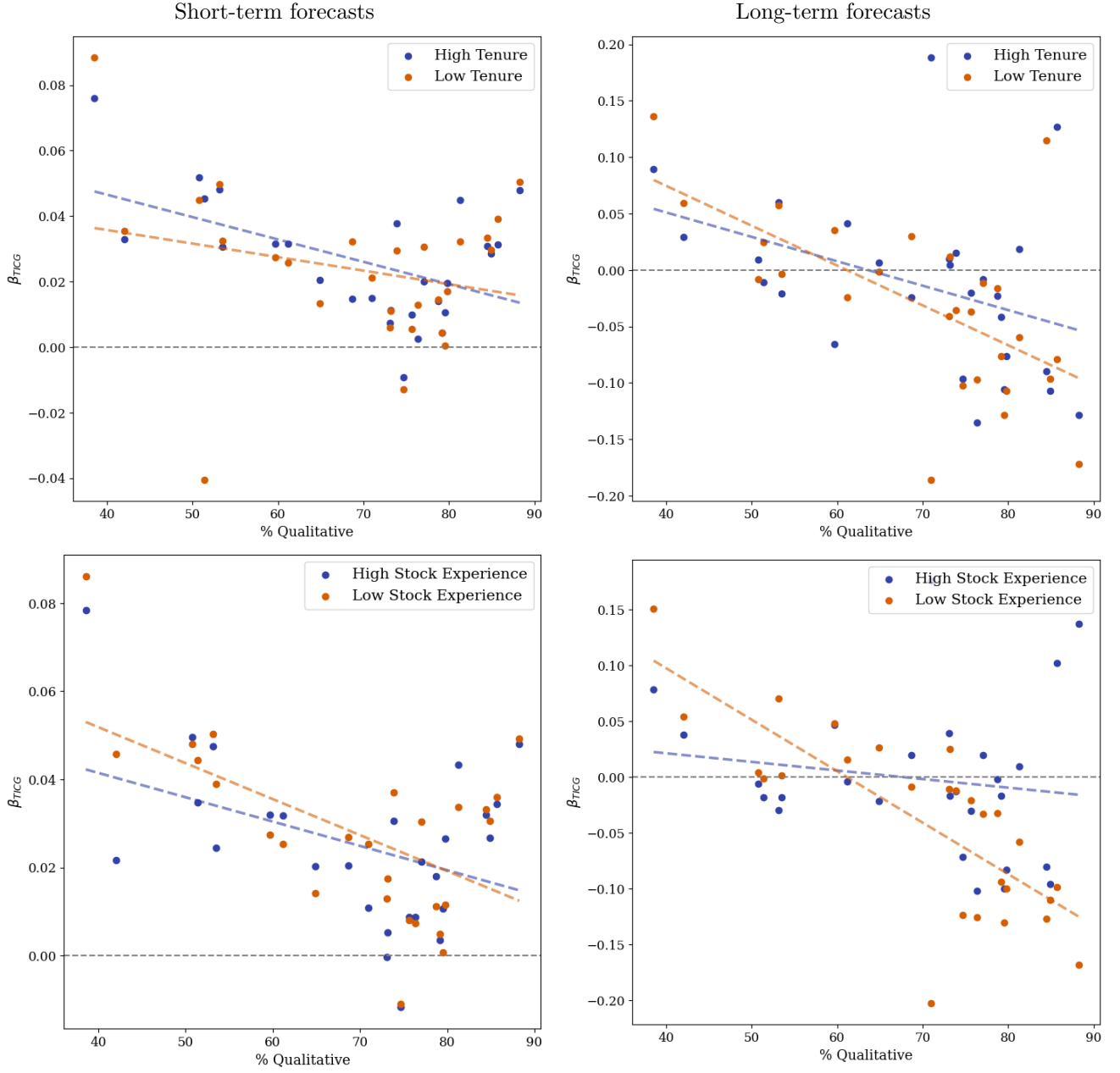


Figure A.15: TICG coefficients versus fraction of qualitative information tuples: Analyst features

*Notes:* The scatter plots show the relationship between attention-interacted CG-regression coefficients and the fraction of qualitative information tuples in a topic. *Tenure* is the number of quarters between time  $t$  and the first time analyst  $i$  issues an EPS forecast. *Analyst Stock Experience* is the number of quarters between time  $t$  and the first time analyst  $i$  issues an EPS forecast for firm  $j$ .

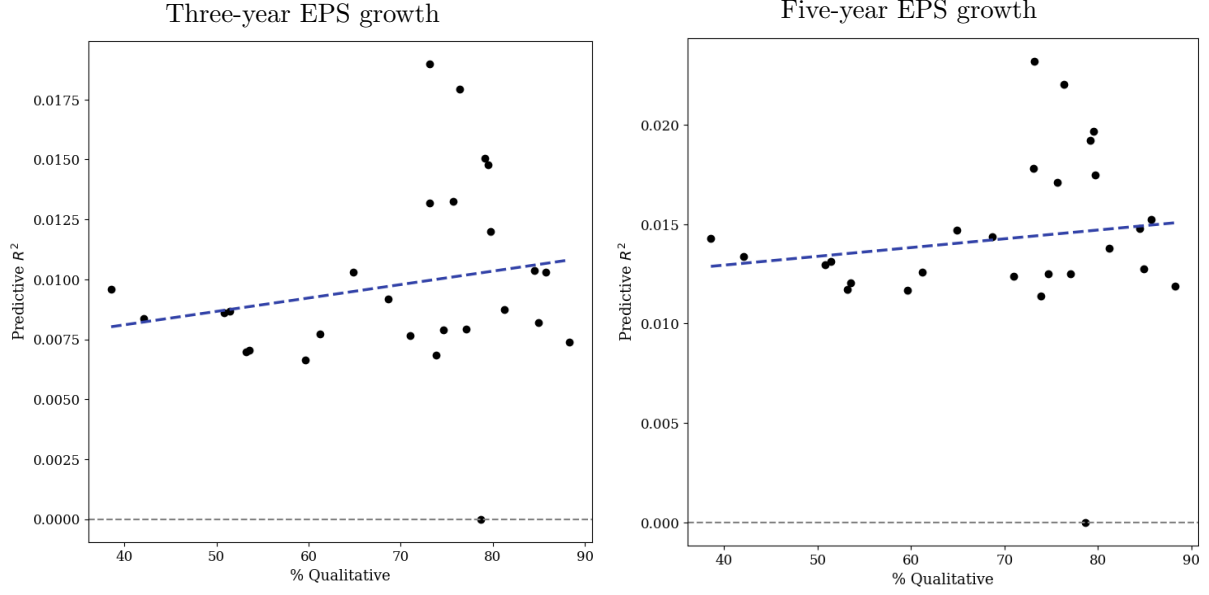


Figure A.16: Earnings growth predictive  $R^2$  versus share of qualitative information Tuples

*Notes:* This plot reports the total  $R^2$  in predictive regression  $\Delta y_{j,t+h} = \beta'_k X_{j,t}^k + \epsilon_{j,t+h}$  against the fraction of qualitative tuples in topic  $k$ , where  $\Delta y_{j,t+h}$  denotes the  $h$ -year EPS growth defined as  $(EPS_{j,t+h} - EPS_{j,t})/P_{j,t}$  and  $P_{j,t}$  is the price of stock  $j$  (in the most recent month before revision).  $X_{j,t}^k$  denotes the average embedding vector of information tuple related to topic  $k$  for stock  $j$  at time  $t$ , taken as the average across analyst reports published in that quarter.

## G Additional Tables

	Dependent variable: $y_{j,t+h} - F_{i,t}[y_{j,t+h}]$					
	Short-term (1-2 years)			Long-term (> 3 years)		
Text Instrumented	0.290*		0.271*	-0.145*		-0.168*
	(39.41)		(12.33)	(-6.34)		(-2.66)
Residual		0.107*	0.101*		-0.147*	-0.152*
		(9.94)	(9.60)		(-5.48)	(-5.47)
Fixed effects	✓	✓	✓	✓	✓	✓
Num Obs.	793462	793462	793462	176285	176285	176285

Table A.1: Forecast error predictability from text-instrumented and residual revisions

*Notes:* This table reports forecast error predictability from text-instrumented revisions and its residuals. The text-instrumented revision is estimated from

$$Rev_{i,t}[y_{j,t+h}] = \sum_{k=1}^{27} \lambda'_k X_{i,j,t}^k + \eta_i + \eta_j + \eta_t + \epsilon_{i,j,t}$$

and computed as  $\widehat{Rev}_{i,t}[y_{j,t+h}] = \sum_{k=1}^{27} \widehat{\lambda}'_k X_{i,j,t}^k$ , while the residual is  $\epsilon_{i,j,t}$ . Jackknife (leave-one-out) estimator is used for estimating the text-instrumented revisions. Analyst-, firm-, and time (quarter)-fixed effects are controlled for.  $t$ -statistics are reported in parentheses. \* denotes statistical significance at 1% level.

	Forecast horizon: Long-term (> 3 Years)								
	Dependent variable: $y_{j,t+h} - F_{i,t}[y_{j,t+h}]$								
$\widehat{Rev}_{i,t}^{qual}[y_{j,t+h}]$	-0.308*		-0.308*	-0.135*		-0.135*	-0.192*		-0.192*
	(-3.41)		(-3.41)	(-2.41)		(-2.41)	(-2.75)		(-2.73)
$\widehat{Rev}_{i,t}^{quant}[y_{j,t+h}]$		0.019	0.019		0.006	0.006		0.078	0.075
		(0.23)	(0.23)		(0.09)	(0.09)		(0.90)	(0.87)
Fixed effects	✓	✓	✓	✓	✓	✓	✓	✓	✓
Control: Lagged Forecast	✓	✓	✓				✓	✓	✓
Control: Past Returns				✓	✓	✓	✓	✓	✓
Adj. $R^2$	0.007	0.001	0.007	0.004	0.004	0.004	0.008	0.008	0.008
Num obs.	176285	176285	176285	151185	151185	151185	151185	151185	151185

Table A.2: Impact of qualitative information on overreaction: controlling for lagged forecasts and past returns

*Notes:* This table reports the results of forecast-error prediction regression, Eqs. (18) and (19) for long-term forecasts. The instrumented forecast revision  $\widehat{Rev}_{i,t}[y_{j,t+h}]$  is defined in Eq. (15), and  $\widehat{Rev}_{i,t}^{qual}[y_{j,t+h}]$  ( $\widehat{Rev}_{i,t}^{quant}[y_{j,t+h}]$ ) are instrumented forecast revisions using qualitative (quantitative) topics defined in Eq. (17). Analyst-, firm-, and quarter-fixed effects are included. Past 1-, 3-, and 5-year returns are used as past return controls. Standard errors are two-way clustered at the analyst-firm and firm-quarter levels,  $t$ -statistics are reported in parentheses, and \* denote statistical significance at the 1% levels.



Dependent variable: Future stock returns									
	$r_{j,t+1}$			$r_{j,t+2}$			$r_{j,t+3}$		
$Rev_t[y_{j,LT}]$	0.016 (0.66)	0.020 (0.63)	0.014 (0.43)	0.005 (0.19)	0.005 (0.16)	0.007 (0.13)	0.012 (0.78)	0.010 (0.52)	0.006 (0.31)
$\widehat{Rev}_t[y_{j,LT}]$		-0.038** (-2.19)			0.007 (0.16)			-0.002 (-0.05)	
$\widehat{Rev}_t^{qual}[y_{j,LT}]$			-0.089** (-2.11)			-0.062* (-1.81)			-0.045* (-1.78)
$\widehat{Rev}_t^{quant}[y_{j,LT}]$			-0.006 (-0.38)			0.033 (0.97)			0.019 (0.70)
Additional Return Predictors	✓	✓	✓	✓	✓	✓	✓	✓	✓
Avg. $R^2$	0.008	0.014	0.018	0.012	0.018	0.023	0.007	0.012	0.018
Nobs	11406	11406	11406	11405	11405	11405	11348	11348	11348

Table A.3: Fama-MacBeth regression of future stock returns on forecast revisions

*Notes:* This table reports results of the Fama-MacBeth regression of future 1- to 3-year stock return on long-term forecast revisions in (20), controlling for known stock return predictors including market capitalization, book-to-market ratio, past 12-month return, asset growth, and gross profitability growth.  $t$ -statistics are reported in parentheses with Driscoll–Kraay standard errors using optimal bandwidth and \* and \*\* denote statistical significance at 10% and 5% level.

Dependent variable: $r_{j,t}$			
$Rev_t[y_{j,LT}]$	0.330* (6.54)	0.302* (6.20)	0.302* (6.18)
$\widehat{Rev}_t[y_{j,LT}]$		0.224* (8.87)	
$\widehat{Rev}_t^{qual}[y_{j,LT}]$			0.197* (7.07)
$\widehat{Rev}_t^{quant}[y_{j,LT}]$			0.098* (3.13)
Avg. $R^2$	0.024	0.030	0.031
Nobs	16831	16831	16831

Table A.4: Panel regression of contemporaneous stock returns on forecast revisions

*Notes:* This table reports results of the panel regression of contemporaneous annual stock return on long-term forecast revisions.

$$r_{j,t \rightarrow t+h} = \beta_1 Rev_t[y_{j,LT}] + \beta_2 \widehat{Rev}_t^{qual}[y_{j,LT}] + \beta_3 \widehat{Rev}_t^{quant}[y_{j,LT}] + \epsilon_{j,t},$$

All original and text-instrumented forecast revisions are scaled to have the same variance.  $t$ -statistics are reported in parentheses with Driscoll–Kraay standard errors using optimal bandwidth and \* denote statistical significance at 1% level.

	Qualitative	Quantitative	Numerical	
Sharpe ratio	0.854	0.122	-0.059	
Qualitative $\alpha$		0.720%*	0.719%*	
		(4.07)	(3.76)	
Qualitative $\beta$		-0.085	0.157*	
		(-1.09)	(2.61)	
CAPM	FF5F + MOM	SY	HXZ	DHS
0.702%*	0.544%*	0.691%*	0.685%*	0.630%*
(3.74)	(2.77)	(3.50)	(3.66)	(2.83)

Table A.5: Summary statistics of long-short portfolios

*Notes:* This table reports the summary statistics of long-short quintile portfolios based on long-term forecast revisions in response to qualitative information, quantitative information, and raw IBES numerical forecasts. The top panel shows the out-of-sample Sharpe ratios of the three portfolios, as well as the monthly alphas and betas from the qualitative-revision portfolio. The bottom panel shows the monthly alphas of the portfolio based on  $\widehat{Rev}_t^{qual}[y_{j,LT}]$  on various factor models, including CAPM, Fama-French 5-factor models plus Momentum (FF5F + MOM), the Stambaugh-Yuan mispricing model (Stambaugh and Yuan, 2017), the q-factor model (Hou et al., 2015), and the 3-factor behavioral asset pricing model in (Daniel et al., 2020).  $t$ -statistics are reported in parentheses, with standard errors computed using Newey-West with 12 lags. \* denotes statistical significance at 1% level.

	$r_t^{FIN}$		$r_t^{PEAD}$	
$\alpha$	0.053*	0.031	0.035*	0.041*
	(2.70)	(1.51)	(2.14)	(2.36)
$r_t^{mkt}$	-0.275*	-0.260*	-0.131*	-0.135*
	(-5.55)	(-5.52)	(-3.01)	(-3.20)
$r_t^{qual}$		0.239*		-0.058
		(2.32)		(-1.22)
$R^2$	0.166	0.231	0.072	0.079
Nobs	180	180	180	180

Table A.6: Spanning behavioral factors in Daniel et al. (2020)

*Notes:* This table reports the spanning regression results of the long-term behavioral factor  $r_t^{FIN}$  and the short-term behavioral factor  $r_t^{PEAD}$  using the long-short quintile portfolio based on  $\widehat{Rev}_t^{qual}[y_{j,LT}]$ .  $t$ -statistics are reported in parentheses, with standard errors computed using Newey-West with 12 lags. \* denotes statistical significance at 1% level.

Forecast horizon: Long-term ( $> 3$ years)			
Dependent variable: $y_{j,t+h} - F_{i,t}[y_{j,t+h}]$			
$\widehat{Rev}_{i,t}^{qual}[y_{j,t+h}]$	-0.111** (-6.55)		-0.119** (-6.87)
$\widehat{Rev}_{i,t}^{quant}[y_{j,t+h}]$		0.039* (2.33)	0.055** (3.24)
Fixed effects	✓	✓	✓
Adj. $R^2$	0.002	0.000	0.002
Num obs.	176285	176285	176285

Table A.7: Impact of qualitative information on overreaction: Alternative construction

*Notes:* This table reports the results of Eq. (19) using alternative definitions of  $\widehat{Rev}_{i,t}^{qual}[y_{j,t+h}]$  and  $\widehat{Rev}_{i,t}^{quant}[y_{j,t+h}]$  as

$$\begin{aligned}\widehat{Rev}_{i,t}^{qual}[y_{j,t+h}] &:= \sum_{k=1}^K \widehat{Rev}_{i,t}[y_{j,t+h}]^k \times Qual\_Frac_{i,j,t}^k; \\ \widehat{Rev}_{i,t}^{quant}[y_{j,t+h}] &:= \sum_{k=1}^K \widehat{Rev}_{i,t}[y_{j,t+h}]^k \times (1 - Qual\_Frac_{i,j,t}^k).\end{aligned}\tag{68}$$

Analyst-, firm-, and quarter-fixed effects are included,  $t$ -statistics are reported in parentheses, with standard errors two-way clustered at the analyst-firm and firm-quarter levels, and \*, \*\* denotes statistical significance at the 5% and 1% level.

$r_{j,t}^{tangible}$				$r_{j,t}^{intangible}$		
$\widehat{Rev}_{j,t}^{qual}[y_{j,LT}]$	0.017* (1.78)	0.018* (1.92)		0.059** (3.86)	0.060** (3.94)	
$\widehat{Rev}_{j,t}^{quant}[y_{j,LT}]$		0.016** (3.35)	0.017** (3.60)		0.015** (2.48)	0.019** (3.06)
Fixed effects	✓	✓	✓	✓	✓	✓
Within $R^2$	0.000	0.000	0.001	0.004	0.000	0.004
Nobs	15878	15878	15878	15878	15878	15878

Table A.8: Predictability of tangible and intangible returns by forecast revisions

*Notes:* This table shows the contemporaneous predictability of tangible and intangible returns by long-term forecast revisions due to qualitative and quantitative information. I compute tangible returns as the book return (changes in log book equity) from year  $t - 1$  to year  $t$ ,  $r_{j,t}^{tangible} = \log(BE_{j,t}) - \log(BE_{j,t-1})$ . The intangible return is computed as the residual in the cross-sectional regression  $r_{j,t} = \gamma_0 + \gamma_1 \log(BE\_ME_{j,t-1}) + \gamma_2 r_{i,t}^{tangible} + v_{j,t}$  and  $r_{j,t}^{intangible} = \widehat{v}_{j,t}$ . Analyst-, firm-, and quarter-fixed effects are included,  $t$ -statistics are reported in parentheses and standard errors are two-way clustered at the analyst-firm and firm-quarter level; \* and \*\* denote statistical significance at the 10% and 1% level.

	Forecast horizon: Long-term (>3 years)		
	Dependent variable: $y_{j,t+h} - F_{i,t}[y_{j,t+h}]$		
$\widehat{Rev}_{i,t}^{quant}[y_{j,t+h}]$	-0.014 (-1.58)	0.001 (0.08)	-0.009*** (-0.92)
$\widehat{Rev}_{i,t}^{quant}[y_{j,t+h}] \times h^{overconfidence}$	0.053*** (1.57)		0.062* (1.78)
$\widehat{Rev}_{i,t}^{quant}[y_{j,t+h}] \times h^{memory}$		-0.040 (-0.88)	-0.056 (-1.21)
Fixed effects	✓	✓	✓
Adj. $R^2$	0.003	0.003	0.003

Table A.9: Impact of overconfidence and memory on overreaction: Revision due to quantitative information

*Notes:* This table reports the analogue of Eq. (22) but using revision due to quantitative information  $\widehat{Rev}_{i,t}^{quant}[y_{j,t+h}]$  and  $h^{overconfidence}$  is the share of chunks of subjective statements that contains assertive language identified by the LLM, and  $h^{memory}$  is the share of factual chunks that describes historical episodes as identified by the LLM. Analyst-, firm- and quarter-fixed effects are included and  $t$ -statistics are reported in parentheses, with standard errors being two-way clustered at the analyst-firm and firm-quarter level; \*\*\* denotes statistical significance at the 1% level.

	Dependent variable: $h^{overconfidence}$	
	Short-term (1-2 Years)	Long-term (> 3 Years)
Abs. revision	-0.047 (-1.45)	0.072 (1.49)
Fixed effects	✓	✓
Adj. $R^2$	0.396	0.388

Table A.10: Impact of absolute revision on  $h^{overconfidence}$

*Notes:* This table shows the regression result of  $h_{i,j,t}^{overconfidence} = \beta Abs.Revision_{i,j,t} + \epsilon_{i,j,t}$ , where  $h_{i,j,t}^{overconfidence}$  is the share of chunks in analyst reports that contain assertive language identified by the LLM, and  $Abs.Revision_{i,j,t}$  is the absolute value of the forecast revision. Analyst-, firm-, and time-fixed effects are included and  $t$ -statistics are reported in parentheses, with standard errors being two-way clustered at the analyst-firm and firm-quarter levels.

Forecast horizon: Long-term (>3 years)						
Dependent variable: $y_{j,t+h} - F_{i,t}[y_{j,t+h}]$						
$\widehat{Rev}_{i,t}^{qual}[y_{j,t+h}]$	-0.036*** (-7.68)	-0.050*** (-11.42)	-0.033*** (-5.82)			
$\widehat{Rev}_{i,t}^{qual}[y_{j,t+h}] \times h_{t-1}^{overconfidence}$	-0.074*** (-3.36)		-0.087*** (-3.25)			
$\widehat{Rev}_{i,t}^{qual}[y_{j,t+h}] \times \Delta h^{overconfidence}$		0.006 (0.51)	-0.039** (-2.04)			
<i>Revision</i>				-0.025 (-1.19)	-0.063*** (-3.36)	0.010 (0.43)
<i>Revision</i> $\times h_{t-1}^{overconfidence}$				-0.340*** (-3.71)		-0.550*** (-4.97)
<i>Revision</i> $\times \Delta h^{overconfidence}$					-0.258*** (-3.62)	-0.515*** (-5.56)
Fixed effects	✓	✓	✓	✓	✓	✓
Adj. $R^2$	0.009	0.008	0.009	0.025	0.025	0.025

Table A.11: Impact of overconfidence on overreaction: Lag and change in  $h^{overconfidence}$

Notes: The first three columns of this table report the regression results of

$$y_{j,t+h} - F_{i,t}[y_{j,t+h}] = \beta_1 \cdot \widehat{Rev}_{i,t}^{qual}[y_{j,t+h}] + \beta_2 \cdot \widehat{Rev}_{i,t}^{qual}[y_{j,t+h}] \times h_{i,j,t-1}^{overconfidence} + \beta_3 \cdot \widehat{Rev}_{i,t}^{qual}[y_{j,t+h}] \times \Delta h_{i,j,t}^{overconfidence} + \epsilon_{i,j,t},$$

where  $h_{i,j,t-1}^{overconfidence}$  is the overconfidence measure from analyst  $i$ 's previous report covering stock  $j$ , and  $\Delta h_{i,j,t}^{overconfidence} = h_{i,j,t}^{overconfidence} - h_{i,j,t-1}^{overconfidence}$  is the change in the overconfidence score. The last three columns report the results of

$$y_{j,t+h} - F_{i,t}[y_{j,t+h}] = \beta_1 \cdot Rev_{i,t}[y_{j,t+h}] + \beta_2 \cdot Rev_{i,t}[y_{j,t+h}] \times h_{i,j,t-1}^{overconfidence} + \beta_3 \cdot Rev_{i,t}[y_{j,t+h}] \times \Delta h_{i,j,t}^{overconfidence} + \epsilon_{i,j,t}.$$

$t$ -statistics are reported in parentheses, with standard errors being two-way clustered at the analyst-firm and firm-quarter level; \*\*, \*\*\* denotes statistical significance at the 5% and 1% level.

	Dependent variable: Forecast errors	
	Forecast horizon	
	Short-term (1-2 Years)	Long-term (> 3 Years)
<i>Revision</i>	0.140*** (10.75)	-0.115*** (-3.99)
<i>Revision</i> $\times$ Num_Est	0.003*** (2.53)	0.006* (1.72)
Fixed effects	✓	✓
Adj. $R^2$	0.013	0.025
Num obs.	883686	173902

Table A.12: Impact of the number of analysts on herding behavior

*Notes:* This table reports the original CG regression, Eq. (13) with an additional interaction term of forecast revision and *Num\_Est*, the number of estimators in the previous month that issued forecasts for the same horizon. *t*-statistics are reported in parentheses, with standard errors being two-way clustered at the analyst-firm and firm-quarter level; \*\*\* denotes statistical significance at 1% level.

	Dependent Variable			
	$h^{overconfidence}$	$h^{memory}$	$h^{herding}$	$h^{sticky}$
Experience	-0.00023*** (-8.53)	0.00007*** (3.68)	0.00002* (1.71)	-0.00001 (-1.61)
Fixed effects	✓	✓	✓	✓
Adj. $R^2$	0.532	0.553	0.886	0.988
Num obs.	503044	503044	503044	503044

Table A.13: Impact of experience on behavioral bias mechanisms

*Notes:* This table reports the impact of analysts' experience on measures of overconfidence, memory effects, herding, and belief stickiness. Qual\_Frac is the fraction of information tuples that are classified as qualitative in the analyst report,  $h^{overconfidence}$  is the share of chunks of subjective statements that contains assertive language identified by the LLM,  $h^{memory}$  is the share of factual chunks that describes historical episodes as identified by the LLM. The dummy  $h^{herd}$  takes the value of 1 if the LLM identifies a reference to consensus or street estimate in the report and  $h^{sticky}$  takes the value of 1 if the LLM identifies a reference to previous earnings forecast in the report. Analyst-, stock-, and time (quarter)-fixed effects are included, *t*-statistics are reported in parentheses, with standard errors two-way clustered at the analyst-firm and firm-quarter level; \*, \*\*, and \*\*\* denote statistical significance at the 10%, 5%, and 1% levels.

	Dependent variable: Forecast errors	
	Forecast horizon	
	Short-term (1-2 Years)	Long-term (> 3 Years)
Revision	0.111*** (13.09)	-0.227*** (-11.30)
Revision $\times$ Herding	0.379*** (22.62)	0.871*** (16.31)
Fixed effects	✓	✓
Adj. $R^2$	0.016	0.035
Num obs.	884244	178932

Table A.14: Impact of herding on CG coefficient

*Notes:* This table reports the original CG regression, Eq. (13), with an additional interaction term of forecast revision and a dummy, *Herding*, which takes the value of 1 when the revision is towards the consensus in the previous month, following Clement and Tse (2005). Firm-month consensus forecasts with fewer than 5 forecasters are dropped. Standard errors are two-way clustered at the analyst-firm and firm-quarter level; \*\*\* denotes statistical significance at the 1% level.

## Online Appendix



## OA.1 Details on Information Retrieval Steps

### OA.1.1 Removing boilerplate content

I use LLM to identify and remove boilerplate content, such as analyst certification and disclosures, using Prompt #0.

### OA.1.2 Relevant context retrieval

The first component for RAG implementation is to retrieve relevant context that will be fed into LLM. For each IBES announcement, I gather the report(s) that are matched with it, and split the text into chunks with at most 128 tokens. This step results in the context set, from which retrieval will be performed.

For each context chunk  $n$ , I generate a 768-dimensional embedding  $x_n$  using ChronoBERT.<sup>55</sup> I then generate another embedding  $y$  for the retrieval query “{TICKER} earnings”, where {TICKER} is the stock ticker being studied. I extract the relevant context chunks as the 20 most similar context chunks with the query, measured by cosine distance of the context embedding with the query embedding:

$$\text{Relevant Context Set} = \{n \text{ for } n \in \text{Top 20 of } \left\{ \frac{x_n \cdot y}{||x_n|| ||y||} \right\}\} \quad (69)$$

By retrieving using similarity in the embeddings, the retrieved context will be more relevant for the earnings forecast of the intended stock. I use 20 chunks to balance context relevance and LLM generation quality. Including more chunks will include more text that is potentially irrelevant for the stock’s earning, and will increase the difficulty for the LLM to extract relevant information given the longer context.

### OA.1.3 Information extraction

After the relevant contexts are retrieved, I feed them into the LLM to extract relevant information for analysts’ EPS forecasts. To guarantee the quality of information extraction, I implement a multi-step prompting process. The steps are described below.

#### Step 1: Identify stock information

In the first step, the LLM is asked to identify the stock in the context based on Prompt #1. I will only proceed with the information extraction if the stock can be correctly identified. This step is important to guarantee that the information extracted is specifically about the target stock, especially as there could be multiple stocks mentioned in the report. I conduct a validation check

---

<sup>55</sup>I do not use Llama-3 embedding for runtime considerations. Manual inspection shows that the retrieval results using FinBERT embedding are similar to those when using Llama-3 embedding.

in Appendix OA.7.1 and find that the LLM achieves high accuracy in both identifying the presence of correct stocks and reporting null content when the stock information is missing.

### Step 2: Separate factual and subjective content

In step 2, I use LLM to summarize factual and subjective statements in analyst reports in natural language. As explained in the main text, this step is important because it allows for more precise measurement of the information set of analysts. In Prompt #3, the LLM is instructed to use the exact wording from the analyst reports. Once the first pass is finished, I repeated run Prompt #4 to ensure that there is no missing content and the LLM output is complete. This is done for each 128-token chunk.

### Step 3: Extracting information from factual and subjective content

In the next step, I extract information from both factual and subjective content. As introduced in Section 2.3, I define information as a tuple of (*Noun: Description*), where *Noun* is a noun or a noun phrase for a business or economic-related concept or object, and *Description* is an adjective or short phrase that describes the *Noun*. I use Prompt #5 for this task, and once the initial run is completed, I repeated run Prompt #6 to guarantee the completeness of extraction.

## OA.1.4 Topic classification

The information retrieval step extracts 70.5 million (*Noun: Description*) tuples with 9.5 million unique *Nouns*. To further distill the information and map to the idea of structural components in Section 2.3, I classify the nouns into topics. I do not impose a pre-defined topic list, but use LLM to generate topic list and assign each noun into topics. The steps are detailed below.

**Step 4: Assign labels to nouns** Before generating topic assignments, I use LLM to first assign a label to each noun. This assignment is independent of other nouns and serves as an intermediate step to introduce some standardization. I find that without this step, the LLM does not generate meaningful topics based and leads to lots of nouns being grouped in the “other” category. Prompt #6 is used in this step.

**Step 5: Generate topic list** I then use LLMs to generate topic lists based on the labels of nouns. As this is an important step, in addition to Llama3.1-8b, I also use other 11 LLM models including Llama3.2-3b, Phi3-8b, Phi3-14b, Phi3.5-3.8b, Phi4-14b, Deepseek-r1-1.5b, Deepseek-r1-7b, Deepseek-r1-8b, Deepseek-r1-14b, and Mistral-7b to run the prompt and generate a complete list of topics. Prompt #7 is used to generate topics, and once it is generated, I run Prompt #8 in each model to ensure the completeness of the topics.

**Step 6: Manual reading and construction of topic list** Each LLM generates 50 to 627 topics with some overlaps and distinctions, but these topic lists are small enough for human inspection. In this step, I manually read and cross-reference topic names that are important and meaningful. Based on manual reading, I conclude a final list of 153 topics. I further group these topics into 28 meta-topics. The topic and meta-topics are listed in Table OA.13.

**Step 7: Classify nouns by topic** Finally, I provide the 153 topic names to LLM and ask the model to classify each noun into one or several topics. I use Prompt #9 to complete the task. To ensure the validity of the topic assignment, I also use multiple LLMs to cross-reference, and determine topic assignment based on a majority rule across LLMs.

Model	Num. Topics	Min. Dist
Llama3.1-8b	294	0.95
Llama3.2-8b	329	0.95
Deepseek-r1-1b	55	0.94
Deepseek-r1-7b	57	0.94
Deepseek-r1-8b	145	0.95
Deepseek-r1-14	591	0.96
Phi3-3.8b	86	0.94
Phi3-14b	56	0.95
Phi4-14b	629	0.95
Mistral-7b	443	0.95

Table OA.1: Topic generation agreement

## OA.2 Additional Tables

	Matched	Selected brokers	Full IBES
Total # of reports	1144952		
# of IBES announcements	818780	1644069	3963955
# of unique stocks ( <i>cusip</i> )	6890	10646	13066
# of unique brokerage firms ( <i>estimid</i> )	115	115	1201
# of unique research analysts ( <i>analys</i> )	5479	7972	19730
Avg. # of pages per report	9.10		
Quarterly avg. # of stock coverage per broker	81.14	130.00	79.63
Quarterly avg. # of reports per broker	197.41		
Quarterly avg. # of brokers coverage per firm	2.99	3.72	6.82
Quarterly avg. # of reports per firm	5.09		

Table OA.2: Summary statistics of the full Investext sample.

	Matched	Selected brokers	Full IBES
<i>EPS/P</i>	0.0313	0.0277	0.0315
<i>Error_1Y/P</i>	-0.0014	-0.0033	-0.0046
<i>Error_2Y/P</i>	-0.0012	-0.0046	-0.0068
<i>Error_3Y/P</i>	-0.0009	-0.0029	-0.0049
<i>Error_4Y/P</i>	-0.0192	-0.0221	-0.0226
<i>Error_5Y/P</i>	-0.0432	-0.0446	-0.0454
<i>Revision_1Y/P</i>	0.0006	-0.0001	-0.0001
<i>Revision_2Y/P</i>	0.0011	0.0004	0.0003
<i>Revision_3Y/P</i>	0.0014	0.0003	-0.0001
<i>Revision_4Y/P</i>	0.0026	0.0018	0.0008
<i>Revision_5Y/P</i>	0.0043	0.0026	0.0014

Table OA.3: Summary statistics of analyst EPS forecasts.

Topic HHI				
Tenure	-0.0004 (-1.26)	-0.0007** (-2.09)	-0.0005 (-1.30)	-0.0007** (-2.04)
Coverage	0.0012** (2.22)	0.0011** (2.15)	0.0010** (1.97)	0.0010* (1.94)
Analyst FE	✓	✓	✓	✓
Stock FE		✓		✓
Time FE			✓	✓
Within $R^2$	0.00	0.00	0.00	0.00
Total $R^2$	0.00	0.07	0.00	0.07

Table OA.4: Determinants of topic concentration

*Notes:* This table shows the determinants of topic concentration by estimating the following regression

$$TopicHHI_{i,j,t} = \beta_1 \cdot Experience_{i,t} + \beta_2 \cdot Coverage_{i,t} + \epsilon_{i,j,t}$$

where  $TopicHHI_{i,j,t}$  is defined as the Herfindahl-Hirschman Index of topic weights in each analyst report, i.e.  $TopicHHI_{i,j,t} = \sum_k (m_{i,j,t}^k)^2$ .  $Tenure_{i,t}$  is analyst  $i$ 's tenure at time  $t$ , defined as the number of quarters between time  $t$  and the first time analyst  $i$  issues an EPS forecast.  $Coverage_{i,t}$  is the number of stocks the analyst  $i$  issues forecasts for in year  $t$ . Standard errors are two-way clustered at the analyst-firm and firm-quarter level. \*, \*\* denotes statistical significance at 10% and 5% level.

	Low-macro attn		High-macro attn		Full sample	
	$\Delta y_{j,t+1}$	$\Delta y_{j,t+1}$	$\Delta y_{j,t+1}$	$\Delta y_{j,t+1}$	$\Delta y_{j,t+1}$	$\Delta y_{j,t+1}$
$\Delta \log IP_t$	0.547 (1.26)	0.544 (1.62)	1.118*** (2.99)	1.165*** (2.56)	0.547 (1.26)	0.470 (1.31)
$\Delta \log IP_t \times \mathbf{1}\{\text{High Attn}\}$					0.571*** (2.08)	0.653*** (3.04)
Stock FE		✓		✓		✓
Num Obs.	72088	72088	72088	72088	144176	144176

Table OA.5: Earnings predictability in high- vs low-macro attention states

*Notes:* This table reports the regression results of 1-year ahead earnings growth predictability using changes in log industrial productions,  $\Delta y_{j,t} = \beta_1 \Delta \log IP_t + \beta_2 \Delta \log IP_t \times \mathbf{1}\{\text{High Macro Attn}\}_{j,t} + \epsilon_{j,t}$ , where  $\Delta y_{j,t}$  is changes in earnings in future one year, and  $\mathbf{1}\{\text{High Macro Attn}\}_{j,t}$  is an indicator that the topic weight of “Macroeconomy” (averaged across all analysts) for firm  $j$  in quarter  $t$  is above the full-sample median.  $\Delta \log IP_t$  is scaled within the high- and low-attention sample to have the same standard deviation. Standard errors are two-way clustered by quarter and stock. \*\*\* denotes statistical significance 1% level.

	1-year future EPS growth			5-year future EPS growth		
	Low $\overline{Attn}_{j,t}$	High $\overline{Attn}_{j,t}$	Full sample	Low $\overline{Attn}_{j,t}$	High $\overline{Attn}_{j,t}$	Full sample
Panel A: Debt						
$\overline{Z}_{j,t}^{Debt}$	0.202*** (8.48)	0.295*** (6.27)	0.202*** (8.48)	0.052*** (7.48)	0.111*** (5.79)	0.052*** (7.48)
$\overline{Z}_{j,t}^{Debt} \times \mathbf{1}\{\text{High Attn}\}$			0.092** (2.05)			0.058*** (3.32)
Panel B: R&D						
$\overline{Z}_{j,t}^{R\&D}$	0.079*** (2.36)	0.086** (2.30)	0.079*** (2.37)	0.046*** (5.04)	0.096*** (5.96)	0.046*** (5.04)
$\overline{Z}_{j,t}^{R\&D} \times \mathbf{1}\{\text{High Attn}\}$			0.006 (0.11)			0.049*** (3.15)
Panel C: Cash						
$\overline{Z}_{j,t}^{Cash}$	0.265*** (7.04)	0.331*** (6.52)	0.264*** (7.04)	0.064*** (10.53)	0.096*** (7.18)	0.064*** (10.53)
$\overline{Z}_{j,t}^{Cash} \times \mathbf{1}\{\text{High Attn}\}$			0.067 (1.45)			0.031** (2.33)
Panel D: Profitability						
$\overline{Z}_{j,t}^{profit}$	0.523*** (11.07)	0.355*** (10.13)	0.523*** (11.07)	0.205*** (12.04)	0.130*** (12.75)	0.205*** (8.68)
$\overline{Z}_{j,t}^{profit} \times \mathbf{1}\{\text{High Attn}\}$			-0.167*** (-3.21)			-0.075*** (-4.82)

Table OA.6: Earnings predictability in high- vs low-attention states: Stock characteristics

*Notes:* This table reports the regression results of 1-year or 5-year ahead earnings growth predictability using current realizations of stock characteristics, i.e.

$$\Delta y_{j,t} = \beta_1 \overline{Z}_{j,t}^k + \beta_2 \overline{Z}_{j,t}^k \times \mathbf{1}\{\text{High Attn}\}_{j,t}^k + \epsilon_{j,t}$$

where  $\Delta y_{j,t}$  is changes in earnings in future 1 year or 5 years, and  $\mathbf{1}\{\text{High Attn}\}_{j,t}^k$  is an indicator that the topic weight of  $k \in \{\text{Profitability, R\&D, Debt, Cash}\}$  (averaged across all analysts) for firm  $j$  in quarter  $t$  is above the full-sample median. Each stock characteristics  $z_{j,t}$  is standardized within each stock to have zero mean and unit standard deviation, and are signed to have positive predictive coefficient.  $\overline{Z}_{j,t}^k$  is the average over the topic  $k$  stock characteristics.  $\overline{Z}_{j,t}^k$  is further scaled within the high- and low-attention sample to have the same standard deviation. Standard errors are two-way clustered by quarter and stock. \*\*, \*\*\* denotes statistical significance at 5% and 1% level.

	Short-term (1-2 Years)					Long-term (> 3 Years)				
$\beta_{CG}$	0.178*** (22.80)	0.178*** (22.76)	0.176*** (22.55)	0.176*** (22.76)	0.176*** (22.55)	-0.114*** (-6.40)	-0.114*** (-6.40)	-0.112*** (-6.33)	-0.114*** (-6.41)	-0.112*** (-6.32)
Analyst FE		✓		✓	✓		✓		✓	✓
Time FE			✓	✓			✓	✓		
Firm FE			✓	✓			✓	✓		
Adj. $R^2$	0.008	0.009	0.012	0.009	0.012	0.023	0.023	0.027	0.024	0.027
Num Obs.	793462	793462	793462	793462	793462	176285	176285	176285	176285	176285

Table OA.7: Coibion-Gorodnichenko regression of EPS forecasts: Robustness of fixed effects

*Notes:* This table reports the original Coibion-Gorodnichenko regression, Eq. (13) with different levels of fixed-effect controls. Standard errors are two-way clustered at the analyst-firm and firm-quarter level and \*\*\* denotes statistical significance at the 1% level.

	Forecast Horizon				
	1 Year	2 Years	3 Years	4 Years	5 Years
$\beta_{CG}$	0.231*** (16.96)	0.137*** (20.39)	-0.035** (-1.99)	-0.262*** (-6.73)	-0.429*** (-7.60)
Fixed Effects	✓	✓	✓	✓	✓
Adj. $R^2$	0.015	0.012	0.013	0.063	0.105
Num Obs.	420205	464451	136122	26885	14230

Table OA.8: Coibion-Gorodnichenko regression of EPS forecasts: Robustness of forecast horizons

*Notes:* This table reports the original Coibion-Gorodnichenko regression, Eq. (13) with different forecast horizons as identified by the `fpi` column in IBES. Standard errors are two-way clustered at the analyst-firm and firm-quarter level.\*\*\* denotes statistical significance at 1% level.

	Forecast Horizon	
	Short-term (1-2 Years)	Long-term (> 3 Years)
$\beta_{CG}$	0.176*** (34.06)	-0.030*** (-2.74)
Fixed Effects	✓	✓
Adj. $R^2$	0.020	0.036
Num Obs.	4306431	820969

Table OA.9: Coibion-Gorodnichenko regression of EPS forecasts: Full IBES dataset

*Notes:* This table reports the original Coibion-Gorodnichenko regression, Eq. (13) using the full IBES dataset. Analyst-, firm-, and time (quarter)-fixed effects are included. Standard errors are two-way clustered at the analyst-firm and firm-quarter level, and \*\*\* denotes statistical significance at the 1% level.

	Forecast Horizon	
	Short-term (1-2 Years)	Long-term (> 3 Years)
Revision	0.183* (22.99)	-0.099* (-5.62)
Lagged Forecast	0.000 (0.38)	0.001* (4.29)
Fixed Effects	✓	✓
Adj. $R^2$	0.012	0.027
Num Obs.	884636	178832

Table OA.10: Coibion-Gorodnichenko regression of EPS forecasts: controlling for lagged forecast

*Notes:* This table reports the original Coibion-Gorodnichenko regression, Eq. (13) controlling for lagged forecast. Analyst-, firm-, and time (quarter)-fixed effects are included. Standard errors are two-way clustered at the analyst-firm and firm-quarter level, and \* denotes statistical significance at the 1% level.



## OA.3 Additional Figures

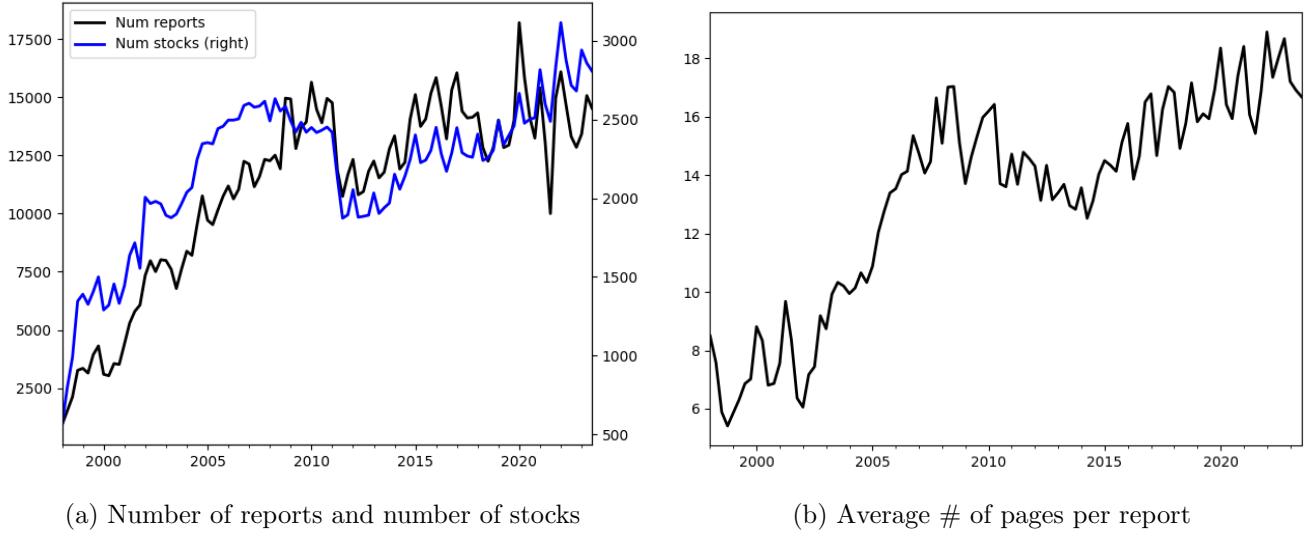


Figure OA.1: Quarterly summary statistics of analyst reports

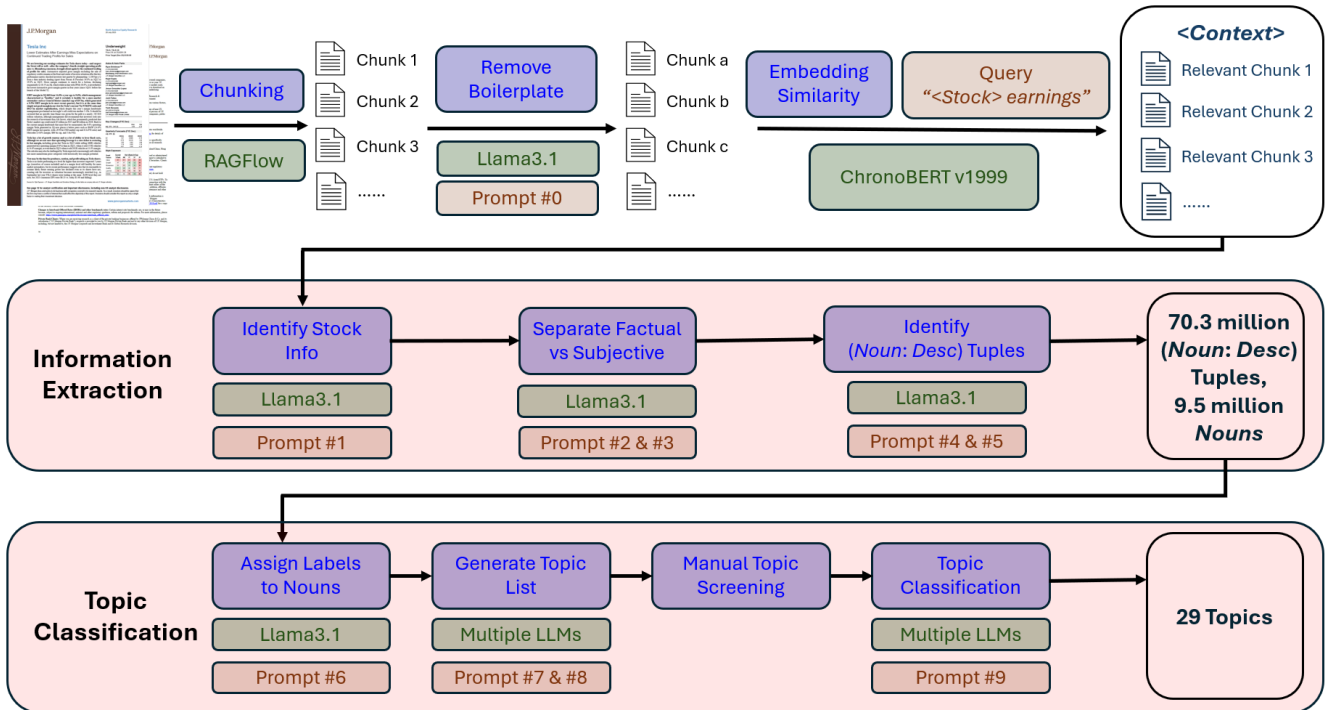


Figure OA.2: Illustration of Information Extraction from Analyst Reports.

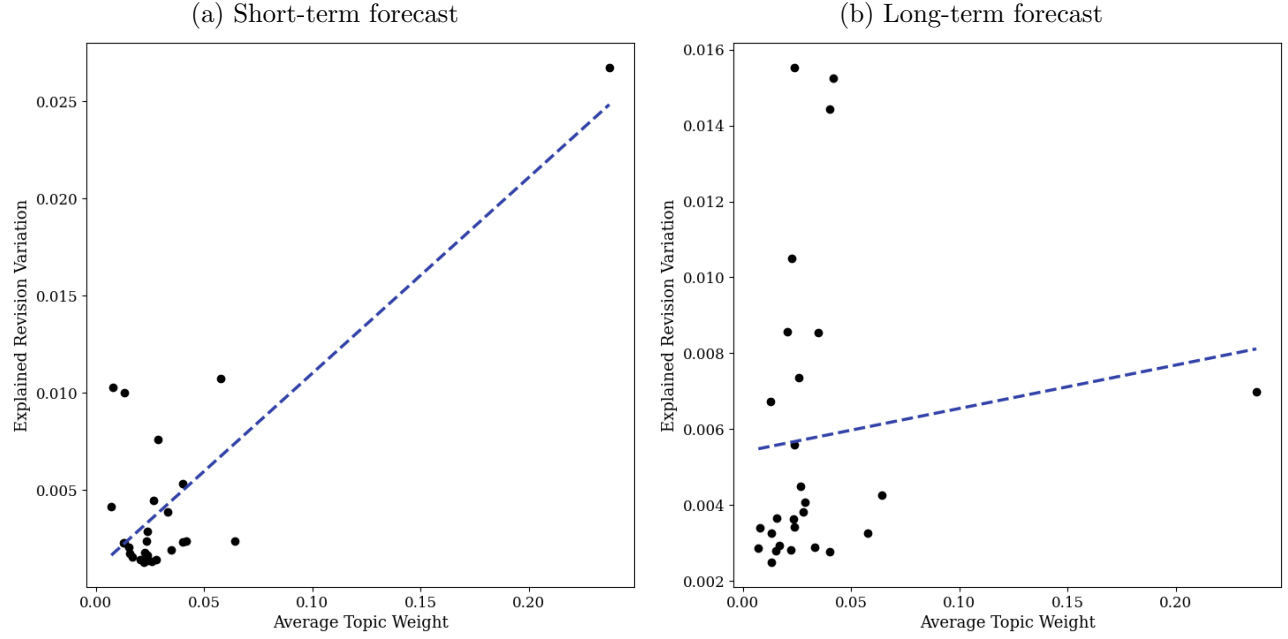


Figure OA.3: Relationship between first-stage variance decomposition and topic weight

*Notes:* This figure shows the relationship between forecast revisions' attributable to a topic and that topic's average weight in analysts' reports. The x-axis shows the average topic weight in analyst reports, and the y-axis shows the variance decomposition of different topics in the first-stage of the TICG regression, as in Figure 4.

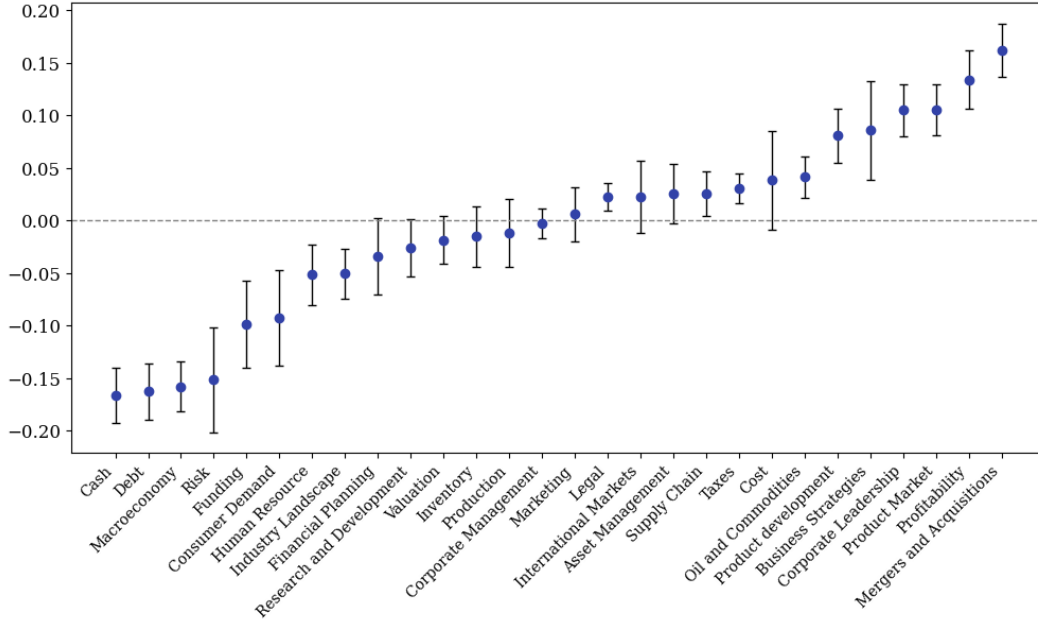


Figure OA.4: Heterogeneity in cyclicality of topic weight

*Notes:* This figure reports the  $\beta$  estimates of regression  $\Delta \bar{m}_t^k = \alpha + \beta \cdot \Delta \log IP_t + \epsilon_t$  for the 28 topics.  $\bar{m}_t^k$  is the average topic weight for topic  $k$  across all reports published in quarter  $t$ .  $\log IP_t$  is the log industrial production obtained from FRED. The error bars reports the 95% confidence intervals. Standard errors are estimated with Newey-West with 4 lags.

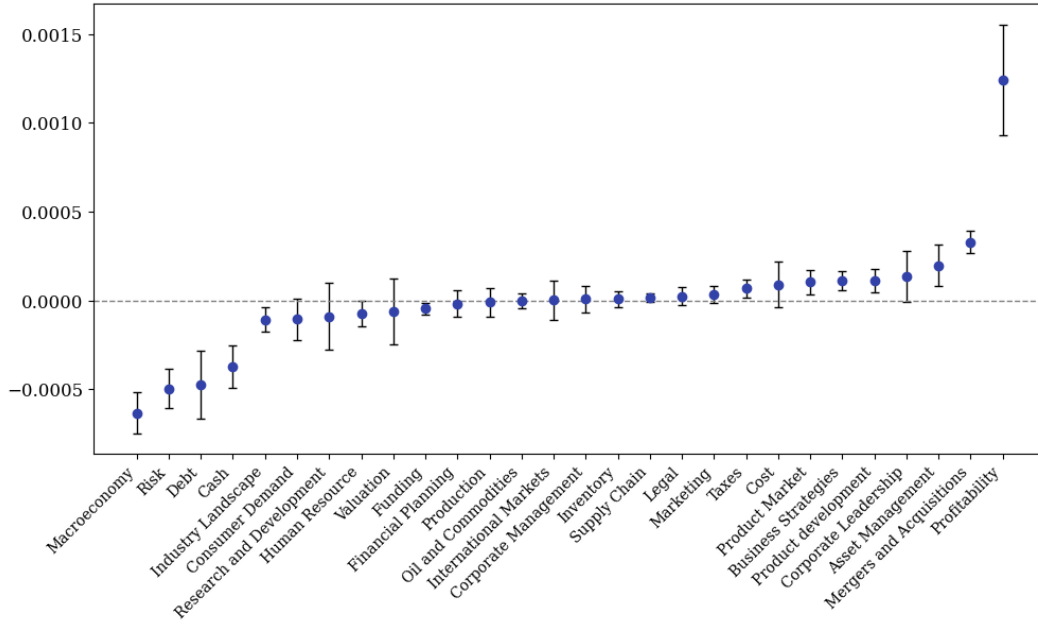


Figure OA.5: Topic weight cyclicity in panel regression

*Notes:* This plot shows the coefficient estimates of  $\beta_k$  from the panel regression  $\Delta \bar{m}_{j,t}^k = \alpha + \beta_j \cdot \Delta \log IP_t + \eta_j + \epsilon_{j,t}$  where  $\bar{m}_{j,t}^k = \frac{1}{N_i} \sum_i m_{i,j,t}^k$  is the average attention across analysts within a quarter  $t$  for firm  $j$ , and  $\eta_j$  captures firm-fixed effects. Standard errors are two-way clustered. The error bar reports the 95% confidence interval.

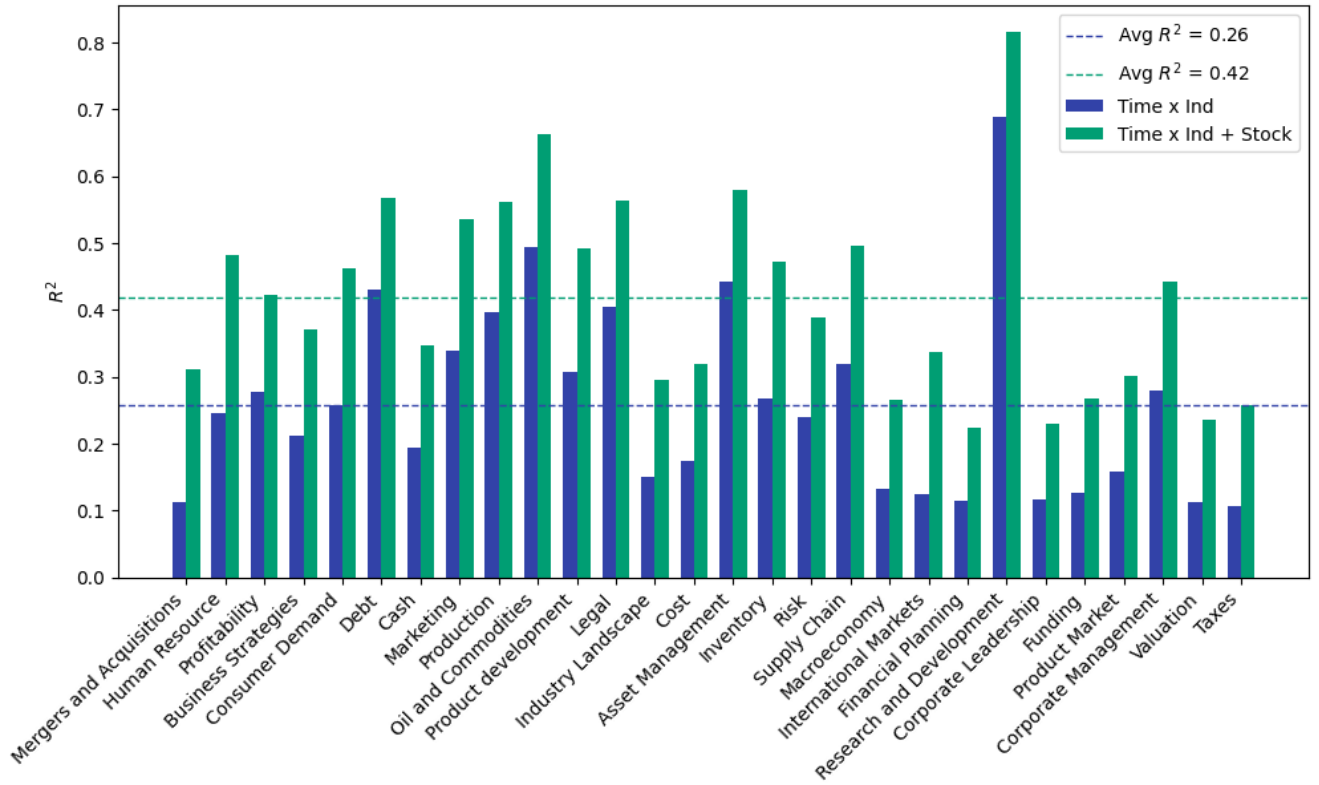


Figure OA.6:  $R^2$  in panel regressions: Time-, industry-, and stock-fixed effects

*Notes:* This plot reports the total  $R^2$  in the panel regression  $\Delta \bar{m}_{j,t}^k = \eta_{t,Ind} + \eta_j + \epsilon_{j,t}$  where  $\bar{m}_{j,t}^k = \frac{1}{N_i} \sum_i m_{i,j,t}^k$  is the average attention across analysts within a quarter  $t$  for firm  $j$ .  $\eta_{t,Ind}$  captures the time (quarter)-  $\times$  industry-fixed effect, where industry group is taken as the 74 GICS industry codes.  $\eta_j$  captures firm fixed effect. The blue (green) bar shows the total  $R^2$  without (with) firm-fixed effects.

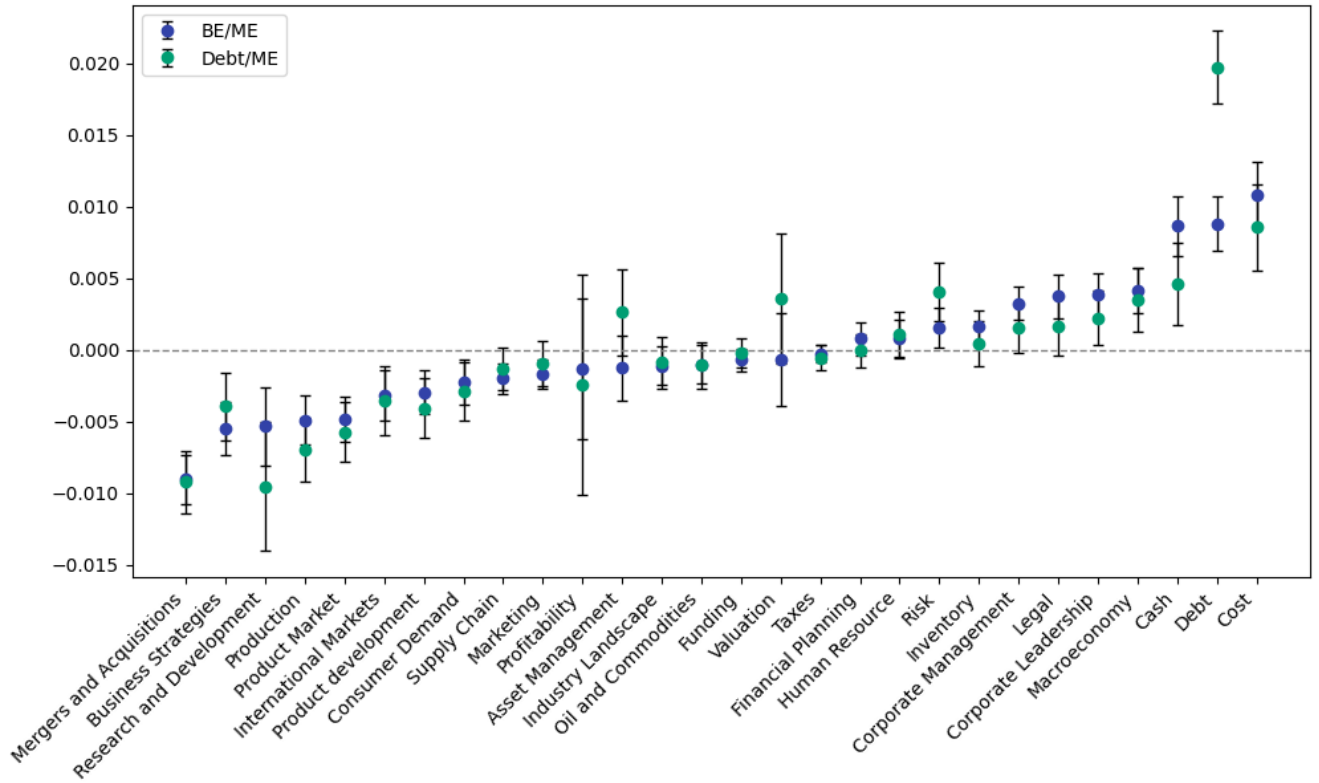


Figure OA.7: Regression coefficient of topic weight on  $BE/ME$  and  $Debt/ME$

*Notes:* This plot shows the panel regression coefficients of  $\bar{m}_{j,t}^k = \beta_k \cdot Z_{j,t} + \eta_t + \eta_j + \epsilon_{j,t}$ , where  $\bar{m}_{j,t}^k = \frac{1}{N_i} \sum_i m_{i,j,t}^k$  is the average attention across analysts within a quarter  $t$  for firm  $j$ , and  $Z_{j,t}$  is either a firm's book-to-market ratio (BE/ME) or debt-to-market ratio (Debt/ME). I rank-standardize  $Z_{j,t}$  within each time (quarter) to be within the  $[-0.5, 0.5]$  interval. Standard errors are two-way clustered by time and firm. The error bars indicate the 95% confidence intervals.

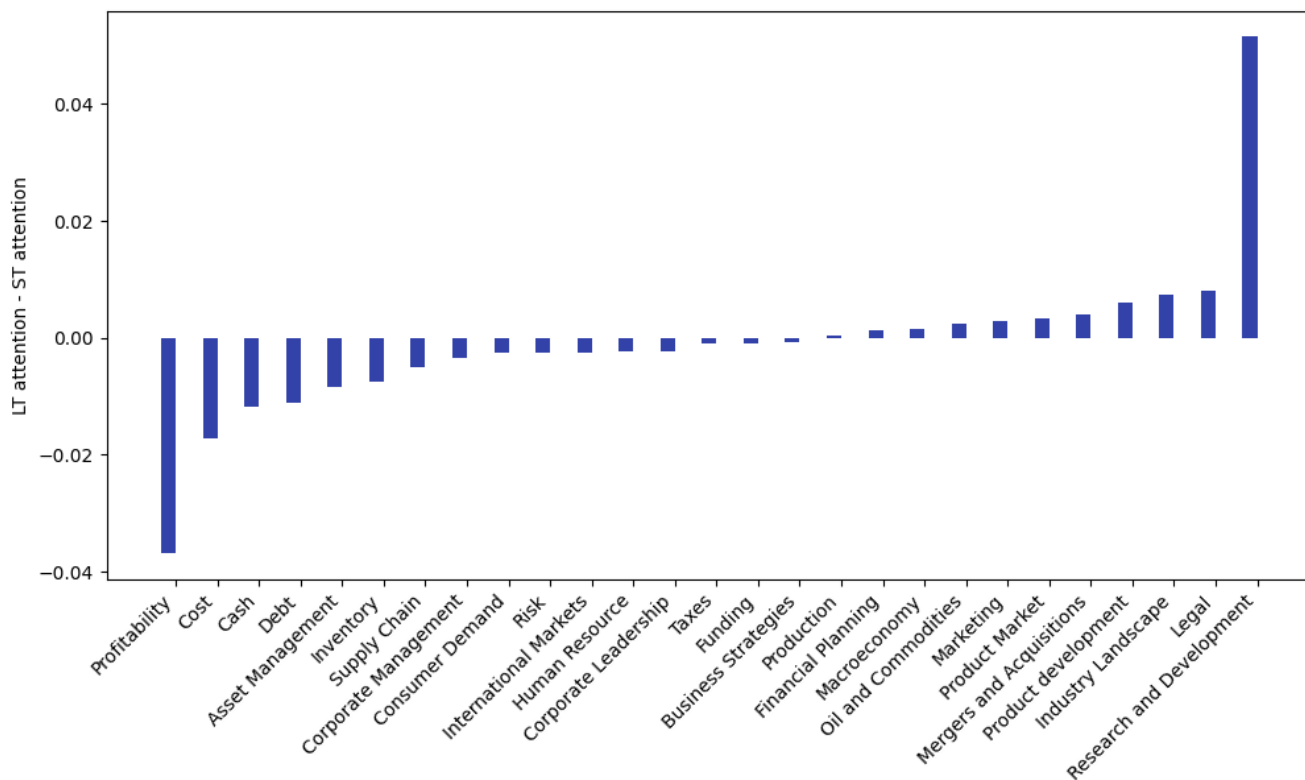


Figure OA.8: Differences in topic weights between short- and long-term EPS forecasts

*Notes:* This plot shows the differences in topic weights between announcements when analysts make short-term versus long-term EPS forecasts. Short-term forecasts are defined as forecasts of future 1-2 years EPS, and long-term forecasts are defined as forecasts of future 3 year or more EPS.

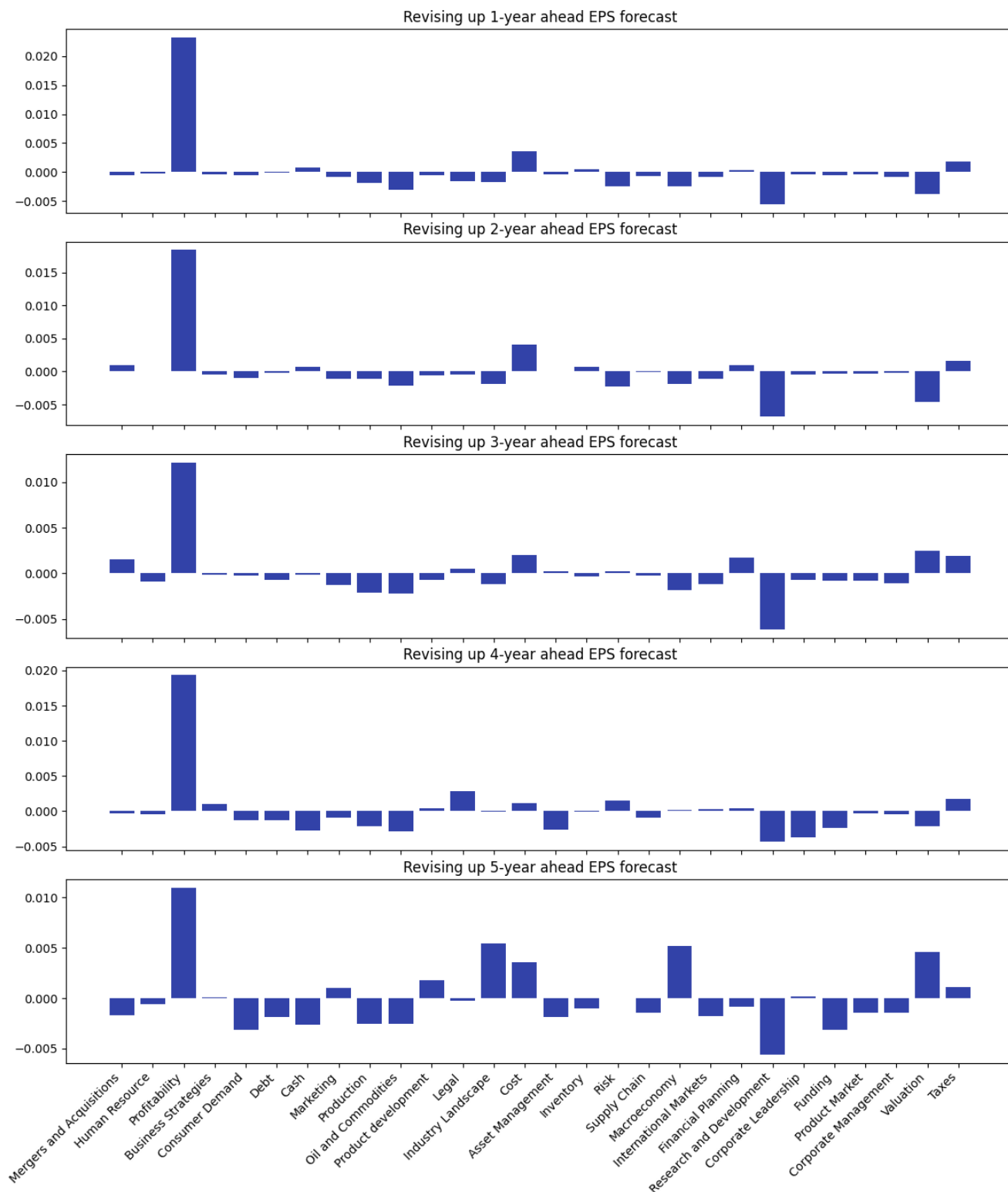


Figure OA.9: Differences in topic weights between upward revision and downward revision

*Notes:* This plot shows the differences in topic weights between announcements when analysts revise up their EPS forecasts versus when they revise down their forecasts.

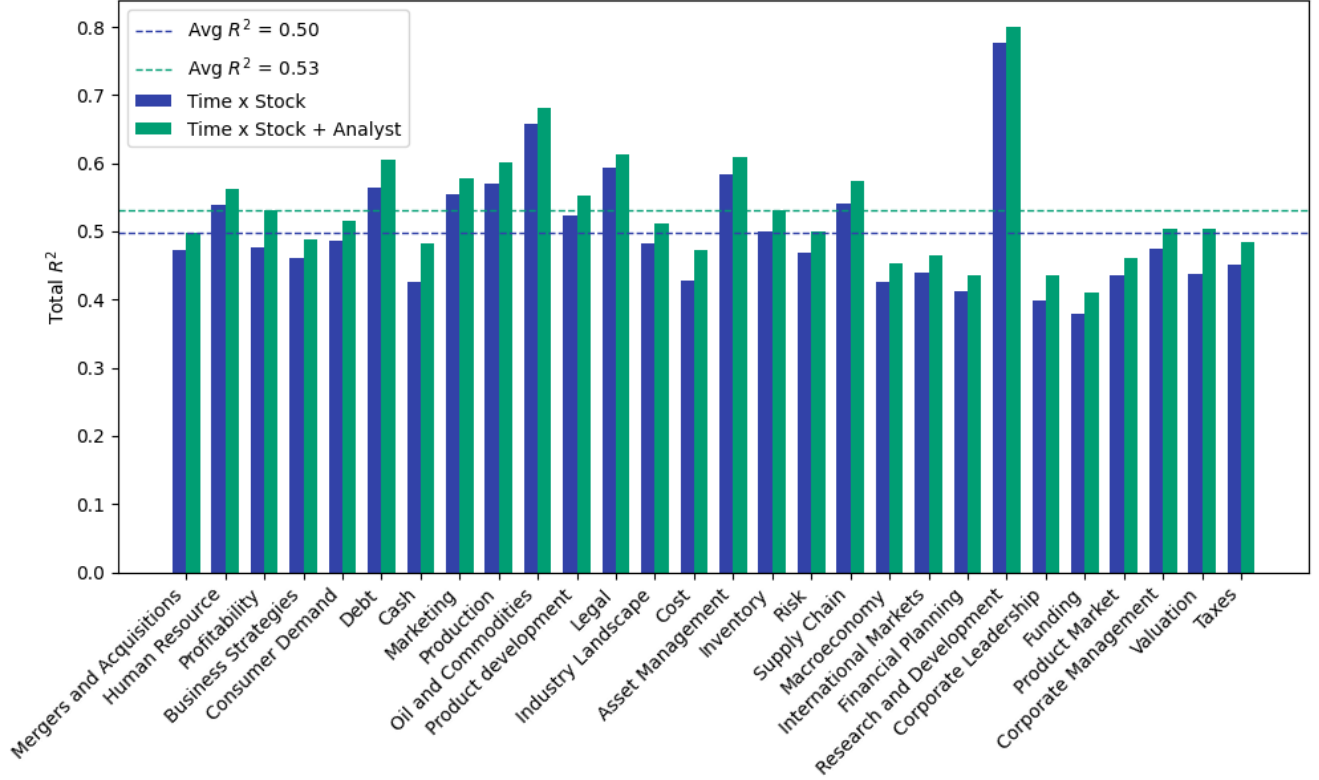


Figure OA.10:  $R^2$  in fixed-effect regressions: Time-, stock-, and analyst-fixed effects

*Notes:* This plot reports the total  $R^2$  in the regression  $m_{i,j,t}^k = \eta_{t,j} + \eta_i + \epsilon_{i,j,t}$  where  $m_{i,j,t}^k$  is the weight of topic  $k$  in the analyst report published by analyst  $i$  at time  $t$  for firm  $j$ .  $\eta_{t,j}$  captures the time (quarter)-  $\times$  firm-fixed effects, and  $\eta_i$  captures analyst-fixed effects. The blue (green) bar shows the total  $R^2$  without (with) analyst-fixed effects.



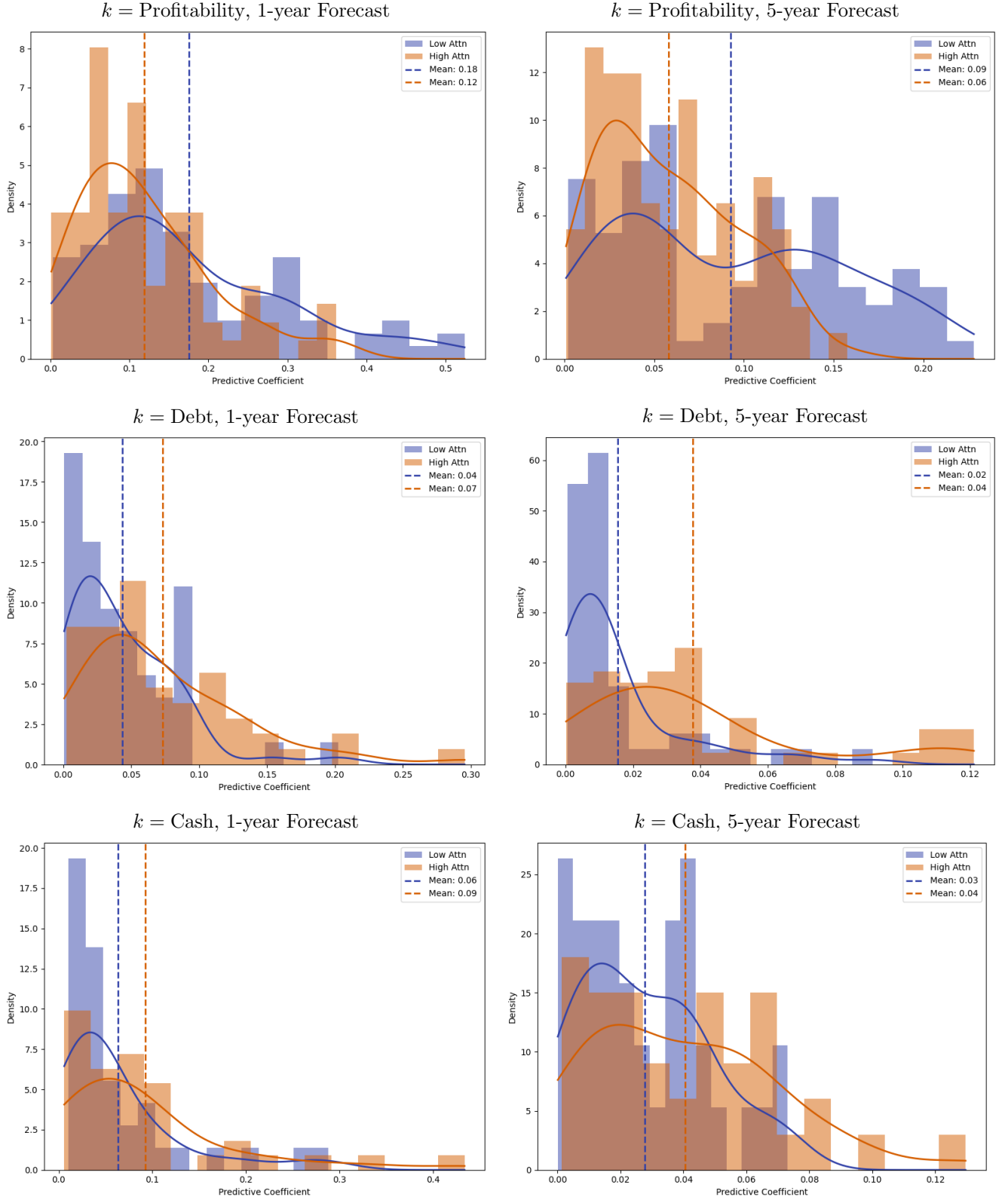


Figure OA.11: Histogram of predictive coefficients for stock-level predictors

*Notes:* This table reports the distribution of predictive coefficients of stock characteristics when predicting future earnings growth, i.e.  $\Delta y_{j,t} = \beta z_{c,j,t}^k + \epsilon_{j,t}$  where  $\Delta y_{j,t}$  is changes in earnings in future 1 year or 5 years, and  $z_{c,j,t}^k$  denotes the realization of characteristics  $c$  that belongs to topic  $k$ . Each stock characteristics  $z_{c,j,t}^k$  is standardized within each stock to have zero mean and unit standard deviation, and are signed to have positive predictive coefficient. It is further scaled within the high- and low-attention sample to have the same standard deviation.

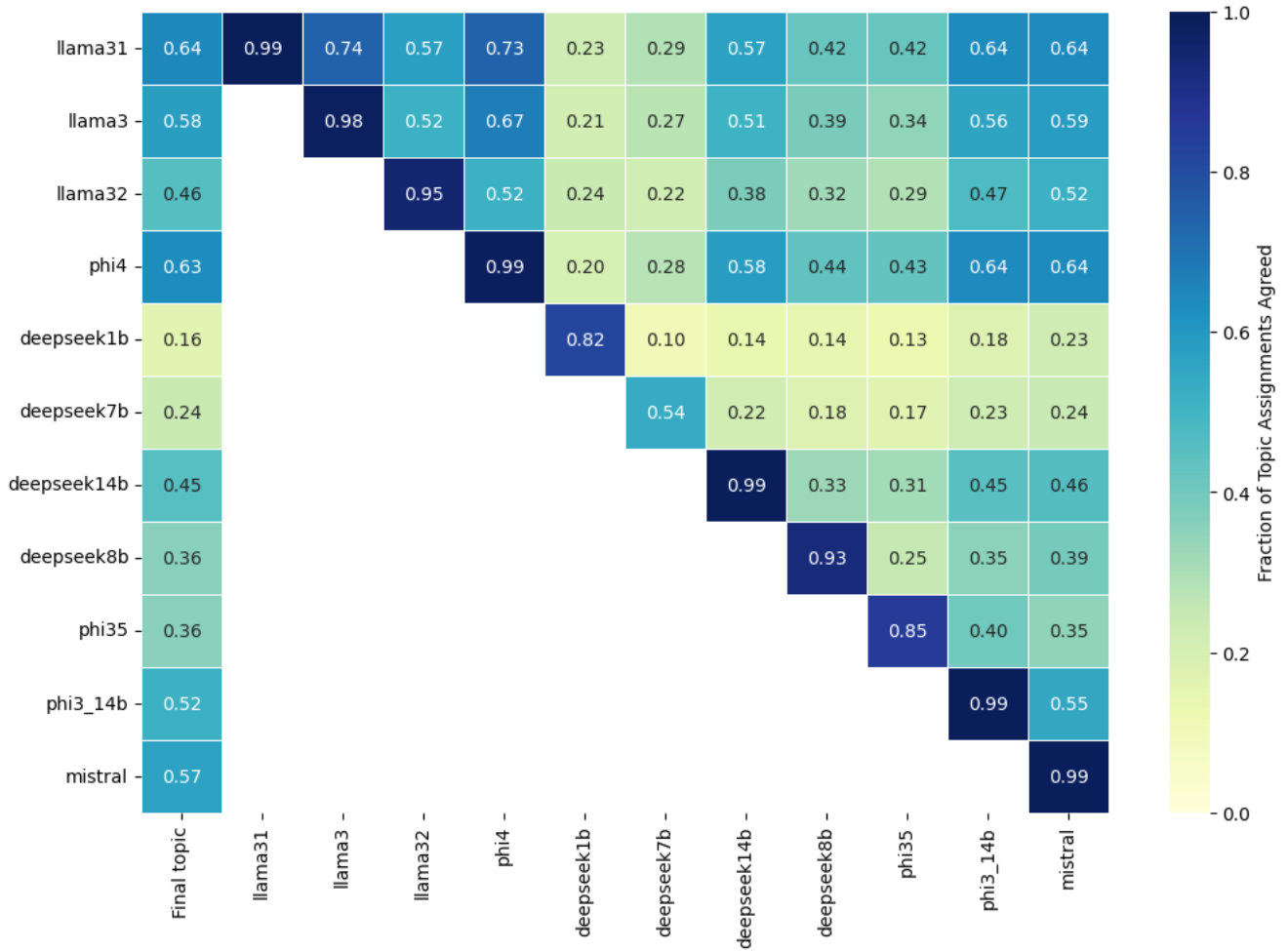


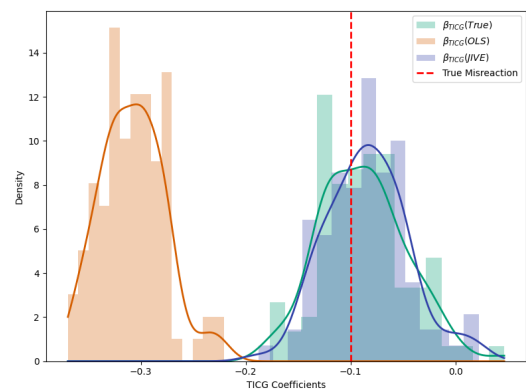
Figure OA.12: Topic Assignment Agreement

## OA.4 Simulation Results

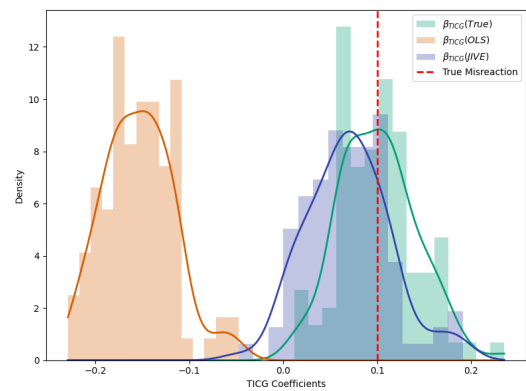
	True $Rev_i^k$	OLS	JIVE
$\rho_\epsilon = 0.0001$	0.870%	0.946%	0.803%
$\rho_\epsilon = 0.005$	0.870%	1.962%	1.821%

Table OA.11: Fraction of Total Variation in  $Rev_i$  Explained by Topic 1

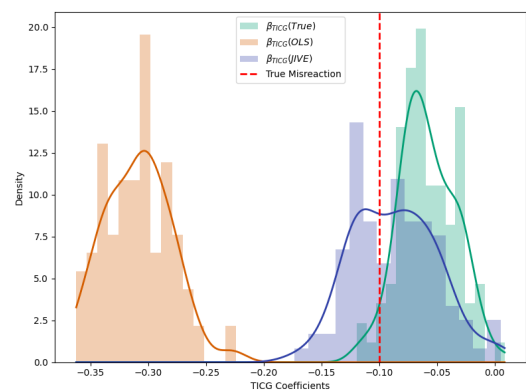
*Notes:* This table shows the average fraction of revision variation explained by the first topic in 100 simulations with settings  $b_1 = 0.1$ ,  $\sigma_e^2 = 0.2$ ,  $\rho_e = 0.4$ ,  $\rho_x = 0$ , and  $b_k \sim Uniform([-0.9, 0.9])$ .

Panel (A):  $\sigma_e^2 = 0, \rho_x = 0.4$ Panel (A):  $\sigma_e^2 = 0, \rho_x = 0.4$ 

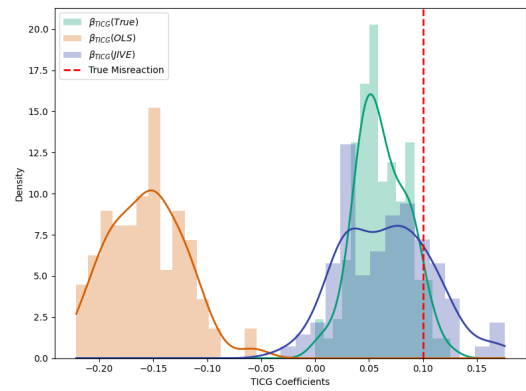
$$b_1 = -0.1, b_k = -b_1$$



$$b_1 = 0.1, b_k = -b_1$$

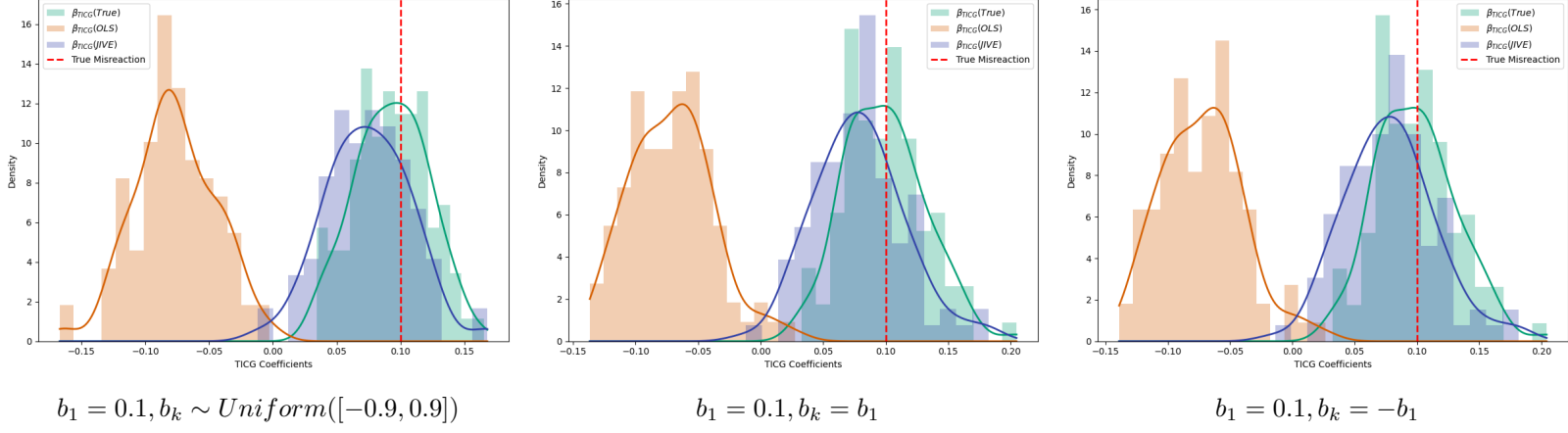
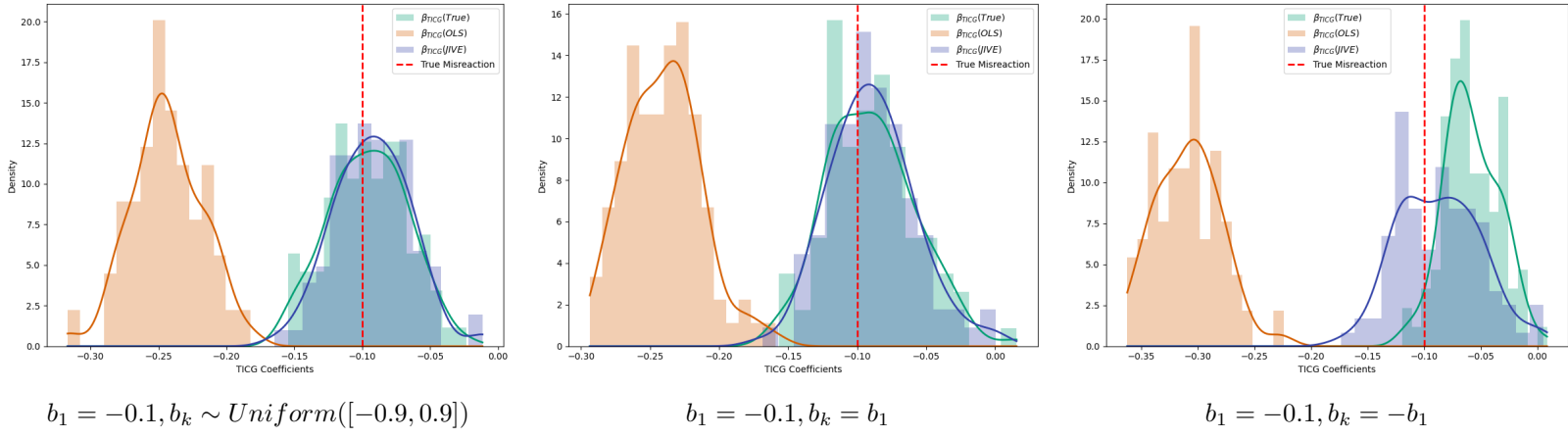
Panel (B):  $\sigma_e^2 = 0.2, \rho_x = \rho_e = 0.4, \rho_\epsilon = 0$ 

$$b_1 = -0.1, b_k = -b_1$$

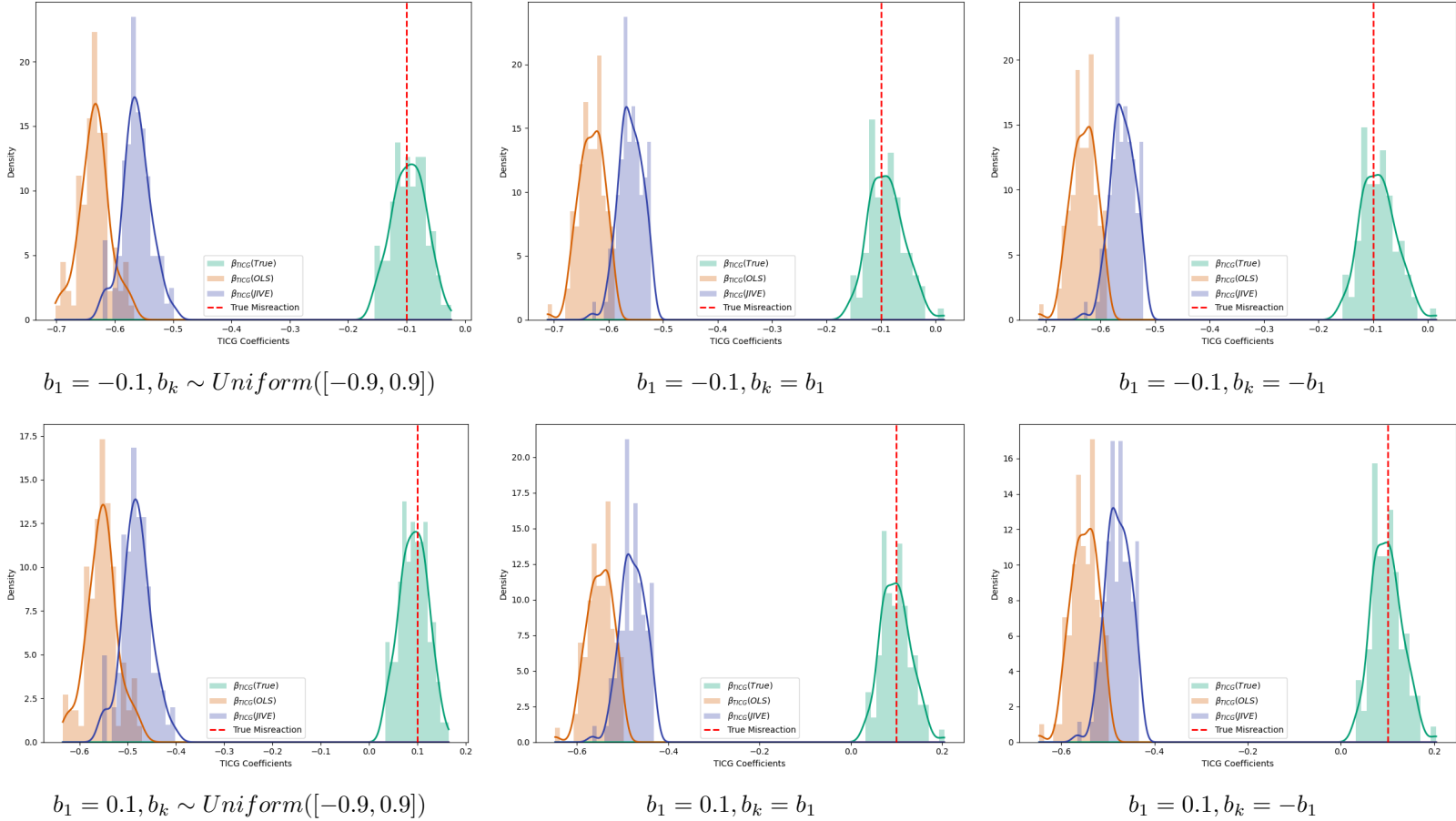


$$b_1 = 0.1, b_k = -b_1$$

Panel (C):  $\sigma_e^2 = 0.2, \rho_x = 0, \rho_e = 0.4, \rho_\epsilon = 0.0001$



Panel (D):  $\sigma_e^2 = 0.2, \rho_x = 0, \rho_e = 0.4, \rho_\epsilon = 0.005$



## OA.5 Prompts

### OA.5.1 Prompts for information retrieval

#### Prompt #0: Remove boilerplate content

##### SYSTEM PROMPT:

Forget previous prompts.

You are an assistant helping a research project and you will be given a paragraph from an equity analyst report from {broker}. Your task is to determine if the paragraph is disclosure or boilerplate content of {broker}, or if it contains meaningful information about the analyst's analysis of a stock. Boilerplate or disclosure content is defined as content that only used by {broker} to satisfy regulatory requirements, with no analysis of a stock. Output your answer in JSON.

##### USER PROMPT:

Here's the context written by the analyst:  
{context}

The above is the context. Below are the questions. Return your answers in JSON format.

Questions:

Q1: Is this a boilerplate or disclosure content? Answer Yes or No. When You have doubt or if it contains at least some analyst's analysis of a stock, answer No.

Output format:

{'Q1': ANSWER to Q1}

Prompt #0: Remove boilerplate content from analyst reports

#### Prompt #1: Identify company information

##### SYSTEM PROMPT:

Forget previous prompts.

You are a sell-side equity analyst specializing in summarizing other sell-side analyst research reports and an expert in causal reasoning, your task is to analyze an equity research report written by analyst at {broker name} on {report date}.

Your goal is to answer user's questions below by summarizing information in the research report. Answer the question based on the context below only. Do not make up the answers.

##### USER PROMPT:

Here's the context written by the analyst:  
{context}

The above is the knowledge base.

Answer the following questions and respond in JSON format.

Questions:

Q1: Is {ticker} mentioned in the context? Answer Yes or No.

Q2: Is {company name} mentioned in the context? Pay attention to its abbreviations. Answer Yes or No.

Q3: If {ticker} or {company name} is mentioned in the context, what is the analyst's EPS forecast or projection for it? Answer one numerical number. Answer 'N/A' if you cannot find answer in the context. Do not make up answer!

Output format:

{'Q1': ANSWER to Q1, 'Q2': ANSWER to Q2, 'Q3': ANSWER to Q3}

Prompt #1: Identify company information

## Prompt #2: Separating factual and subjective statements

### SYSTEM PROMPT:

Forget previous prompts.

You are an assistant helping a research project where you need to extract and summarize the factual information about a stock based on an analyst report about {ticker} ({company\_name}). Your task is to separate objective facts from subjective statements in an analyst report.

Objective facts are verifiable pieces of information, such as business developments, historical financial performance, company announcements, industry benchmarks, or anything that are objective about {ticker}.

Subjective statements reflect the analyst's opinions, predictions, interpretations, or valuation about {ticker}, often indicated by qualitative language, speculations, or value judgments.

Answer the question based on the context below only. Do not make up the answers. Output your answer in JSON.

### USER PROMPT:

Here's the context chunk written by the analyst:

{context}

The above is the context. Answer the questions and return in JSON format.

Questions:

Q1: Based on the context, what are the objective facts about {ticker} ({company\_name})? Include all details about the firm's operations, business, financials, and other information that are factual.

Write your answer in a passage in natural language and keep the original wording in the analyst reports as much as possible.

Start your answer with "{ticker} ...". Answer 'N/A' if you cannot find factual information about {ticker} in the context, or if the context is a disclosure. Do not make up your answer!

Q2: Based on the context, what are the analyst's subjective statement about {ticker} ({company\_name})? Your answer should not overlap with the objective facts in your answer to Q1.

Include all details about analyst's opinion about {ticker} in the future. Write your answer in a passage in natural language and keep the original wording in the analyst reports as much as possible.

Start your answer with "The analyst believes {ticker}...". Answer 'N/A' if you cannot find analyst's subjective statement about {ticker} in the context, or if the context is a disclosure. Do not make up your answer!

Output format:

{'Q1': ANSWER to Q1, 'Q2': ANSWER to Q2}

## Prompt #2: Separating factual and subjective statements

## Prompt #3: Checking factual and subjective statements are complete

### SYSTEM PROMPT:

Forget previous prompts.

You are an assistant helping a research project where you need to extract and summarize the factual information about a stock based on an analyst report. You will be provided with one paragraph that summarizes the factual information, one paragraph summarizing the subjective statements, and the original content from analyst report. Your task is to determine if there is missing factual information or missing subjective statements.

Objective facts are verifiable pieces of information, such as business developments, historical financial performance, company announcements, industry benchmarks, or anything that are objective about the stock.

Subjective statements reflect the analyst's opinions, predictions, interpretations, or valuation about ticker, often indicated by qualitative language, speculations, or value judgments.

Answer the question based on the context below only. Do not make up the answers. Output your answer in JSON.

#### USER PROMPT:

Here's the context chunk written by the analyst (CONTEXT):

{context}

Here's the factual information summarization (PARAGRAPH 1):

{factual content paragraph }

Here's the subjective statement summarization (PARAGRAPH 2):

{subjective content paragraph}

The above is the context. Answer the questions and return in JSON format.

Questions:

Q1: Does PARAGRAPH 1 include all factual details about the firm in CONTEXT? If so, return "Yes". If not, please rewrite the paragraph to include all details about the firm's operations, business, financials, and other information that are factual.

Write your answer in a passage in natural language and keep the original wording in the analyst reports as much as possible.

Start your answer with "{ticker} ...". Answer 'N/A' if you cannot find factual information about {ticker} in the context, or if the context is a disclosure. Do not make up your answer!

Q2: Does PARAGRAPH 2 include all subjective statements about the firm in CONTEXT? If so, return "Yes". If not, please rewrite the paragraph to include all details about analyst's opinion about ticker in the future. Your answer should not overlap with the objective facts in your answer to Q1.

Write your answer in a passage in natural language and keep the original wording in the analyst reports as much as possible.

Start your answer with "The analyst believes {ticker}...". Answer 'N/A' if you cannot find analyst's subjective statement about {ticker} in the context, or if the context is a disclosure. Do not make up your answer!

Output format:

{ 'Q1': ANSWER to Q1, 'Q2': ANSWER to Q2 }

Prompt #3: Checking factual and subjective statements are complete

#### Prompt #4: Extracting financial economic information

##### SYSTEM PROMPT:

Forget previous prompts.

You are an assistant helping a research project where you need to extract topics from factual information in a passage from a sell-side analyst report. Your task is to list what information does the analyst pay attention to, and what is the pieces of information. Identify the nouns and the associated description. Include all the details in your answer. If no description is associated with the noun, note "N/A". Below are some examples:

Example 1: "MKC has a flavor category growing at a healthy pace. The company has a strong M&A track record and sometimes dominant market shares."

Output: {'flavor category': 'growing at a healthy pace', 'M&A track record': 'strong', 'market share': 'dominant'}

Example 2: "QFAB announced the offering of 3.2 million shares of common stock on July 29, with 3 million being sold by the company and 200,000 being sold by Nortek. The offering was priced at \$13.00 per share, well below the approximate \$18 filing price."

Output: {'common stock shares offering': 'announced', 'offering price': 'below the filing price'}

Example 3: "QFIN reported second-quarter EPS of \$0.19 versus \$0.21 a year ago. This was two

pennies ahead of our \$0.17 estimate. Net sales for the quarter increased 22.1% to \$64.1 million and essentially were in line with our expectations. Revenues across the divisions were as follows: domestic fabric sales increased 31.8% to \$47.3 million and accounted for 72.5% of the sales mix, up from 66.9% a year ago; export fabric sales rose 2% year over year to \$10 million and represented 15.4% of the sales mix, down from 18.2% a year ago, and export sales were strongest in Canada and Mexico; lastly, yarn sales were flat year over year and accounted for 12.1% of the sales mix, down from 14.8% a year ago."

Output: {'second-quarter EPS': 'above forecast', 'net sales': 'in line with expectation', 'domestic fabric sales': 'increase', 'export fabric sales': 'increase', 'export sales': 'strongest in Canada and Mexico', 'yarn sales': 'flat'}

Example 4: "GCO reported solid FQ3 results with a \$0.13 beat mainly due to outperformance at the Journeys chain (+8% comp). November comps declined 4%, and the company's Q4 outlook is for 'flat'. The company's sales rate trends are moderating quickly."

Output: {'Journeys chain': 'outperformance', 'comps': 'declined', 'sales rate trends': 'moderating quickly'}

Example 5: "IT (Gartner, Inc.) reported third-quarter results on November 2 with adjusted EPS of \$0.35, which was \$0.01 lower than the estimate and consensus. Revenue was about \$4 million below the estimate and consensus. The shortfall is attributed to a shift of events into the fourth quarter, weaker-than-expected consulting revenue, and currency headwinds."

Output: {'adjusted EPS': 'lower than the estimate and consensus', 'revenue': 'below the estimate and consensus', 'shift of events': 'N/A', 'consulting revenue': 'weaker-than-expected', 'currency headwinds': 'N/A'}

Answer the question based on the context below only. Do not make up the answers. Output your answer in JSON.

---

#### USER PROMPT:

Here's the context chunk written by the analyst:  
{context}

The above is the context. Below are the questions.

Based on the context, extract information about the stock and summarize in {NOUN: DESCRIPTION} pair, where NOUN is an object, concept or event mentioned in the context and DESCRIPTION is a short verb or adjective phrase that describes NOUN.

Extract all relevant information in the paragraph with as much detail as possible. Answer 'N/A' if you cannot find any factual information. Do not make up your answer!

Output format:

{NOUN 1: DESCRIPTION 1, NOUN 2: DESCRIPTION 2, ...}

#### Prompt #4: Extracting Information

#### Prompt #5: Checking information retrieval is complete

#### SYSTEM PROMPT:

Forget previous prompts.

You are an assistant helping a research project where you need to extract topics from factual information in a passage from a sell-side analyst report. You will be given a list of ("NOUN": "DESCRIPTION") tuple, and your goal is to find any missing nouns with their descriptions in the context provided.

Answer the question based on the context below only. Do not make up the answers. Output your answer in JSON.

---

#### USER PROMPT:

Here's the context chunk written by the analyst:  
{context}



Here's the list of (NOUN: DESCRIPTION) tuple:  
{(NOUN: DESCRIPTION) tuples}

Based on the context, does the list contain all information in the context? If so, return "Yes". If not, regenerate the list by extract information about the stock and summarize in {NOUN: DESCRIPTION} pair, where NOUN is an object, concept or event mentioned in the context and DESCRIPTION is a short verb or adjective phrase that describes NOUN.

Extract all relevant information in the paragraph with as much detail as possible. Answer 'N/A' if you cannot find any factual information. Do not make up your answer!

Output format:

{NOUN 1: DESCRIPTION 1, NOUN 2: DESCRIPTION 2, ...}

Prompt #5: Checking information retrieval is complete

## OA.5.2 Prompts for topic generation and classification

Prompt #6: Assign a label to each noun

SYSTEM PROMPT:

Forget previous prompts.

You are an economic & finance topic research assistant. Your task is to classify nouns into labels. Output your answer in JSON.

---

USER PROMPT:

Here are the nouns:

{context}

- 1 Group them meaningful economics or finance labels. Labels must be related to economics, finance, or business operations. If you cannot find a proper label for a noun, put it in the "other" label.
2. For each label, provide
  - A concise name (2–5 words).
  - A 1-sentence economic/financial description of the label.
  - A list of member nouns.

You should classify all the nouns given to you! Make sure you are not missing nouns in your classification.

Output format:

{label: ..., label description: ..., members: [noun1, noun2, ...] }, .....

Prompt 6: Assign a label to each noun

Prompt #7: Generate topic list based on labels of nouns

SYSTEM PROMPT:

Forget previous prompts.

You are an economic & finance topic research assistant. You will be given a list of labels and their descriptions, and your task is to group these labels into several topics based on the similarities in their descriptions. Output your answer in JSON.

---

USER PROMPT:

Here is the list of labels with their descriptions:

{context}

Your task: group similar labels into a topic, and describe the meaning of the topic. I am looking for a list of topics that are unique and well-differentiated, with no duplicated or highly similar topics in the topic list you come up with. If a label is hard to be grouped in any topic, put it in the "Other" topic.

For each topic, provide:

- A concise topic label (2-5 words)
- A 1-sentence economic/financial description of the topic.
- A list of member labels that belongs to this topic.

Output format:

{topic: ..., topic description: ..., members: [label1, label2, ...] }, .....

Prompt #7: Generate topic list based on labels of nouns

Prompt #8: Checking topic list generation is complete

SYSTEM PROMPT:

Forget previous prompts.

You are an economic & finance topic research assistant. You will be given a list of labels and their descriptions, as well as a list of topic assignments. Your task is to determine whether the topic list is complete, and if not, identify missing topics. Output your answer in JSON.

USER PROMPT:

Here is the list of labels with their descriptions:

{context}

Here's the list of topics for these labels:

{topic list}

Is the topic list complete? If so, return "Yes". If not, modify the topic list by grouping similar labels into a topic, and describe the meaning of the topic. I am looking for a list of topics that are unique and well-differentiated, with no duplicated or highly similar topics in the topic list you come up with. If a label is hard to be grouped in any topic, put it in the "Other" topic.

For each topic, provide:

- A concise topic label (2-5 words)
- A 1-sentence economic/financial description of the topic.
- A list of member labels that belongs to this topic.

Output format:

{topic: ..., topic description: ..., members: [label1, label2, ...] }, .....

Prompt #8: Checking if topic generation is complete

### Prompt #9: Classify nouns into topics

#### SYSTEM PROMPT:

Forget previous prompts.

You are an economic & finance topic research assistant. You will be given some nouns or noun phrases, and your task is to classify each noun into one or several of the following topics:

#### LIST OF TOPICS

Only answer the topic names that best fit the noun. If the noun doesn't belong to any topic, answer "N/A". Output your answer in JSON.

---

#### USER PROMPT:

Here are the noun or noun phrase:

{context}

Which topic does each noun belong to? Answer only the topic name.

Output format:

{"Noun1": Topic 1, Noun2: Topic 2, .....}

### Prompt 9: Classify nouns into topics

### Prompt 10: Classify (Noun: Description) tuple into "Qualitative" or "Quantitative"

#### SYSTEM PROMPT:

Forget previous prompts.

You are an assistant helping a research project where you need to classify a piece of information extracted from a sell-side analyst report. The information provided to you will be presented in the form of {"NOUN": "DESCRIPTION"} tuple.

Qualitative information includes non-numerical factors that provide insight into its operations, culture, and narratives. Quantitative information includes verifiable numerical business or accounting data, or other kinds of statistics.

Return letter A if it is Quantitative, letter B if it is Qualitative, and letter C if unclear. Please write only a letter A, B, or C. Add no other formatting or bolding.

---

#### USER PROMPT:

Here is the information tuple:

{context}

Is this information tuple quantitative or qualitative? Return letter A if the information is the Quantitative, letter B if the information is Qualitative, and letter C if unclear.

Please write only a letter A, B, or C. Add no other formatting or bolding.

### Prompt 10: Classify (Noun: Description) tuple into "Qualitative" or "Quantitative"

### Prompt 11: ChatGPT-4o's evaluation of Top 1000 topic classifications

#### SYSTEM PROMPT:

You are a financial expert helping a research project. You will be given a noun or noun phrase, a topic name, and a list of candidate topic names.

Your task is to determine if the noun is best characterized by the chosen topic name among all the topics.

Return A if it is the best topic name, and B if it is not the best topic name.

Please write only a letter A or B. Add no other formatting or bolding.

---

#### USER PROMPT:

The noun or noun phrase is: {NOUN}  
The chosen topic is: {CHOSEN TOPIC}.  
Is {CHOSEN TOPIC} the best topic for {NOUN} among all the topics in the list? Answer only A (for yes) or B (for no).

Prompt 11: ChatGPT-4o's evaluation of Top 1000 topic classifications

#### Prompt 12: Find Assertive Language

##### SYSTEM PROMPT:

You are an economic and finance research assistant. You will be given a passage of analyst reports, and your task is to identify if the analyst uses assertive language. Please analyze the tone and language, especially for signs of overconfidence, such as assertive phrasing, unwarranted certainty, or definitive predictions without qualifiers.

##### USER PROMPT:

Here's a passage from the analyst report:

{Context}

Does the analyst use any assertive phrasing, unwarranted certainty, or definitive predictions without qualifiers. Answer 0 if no, and 1 if analyst uses assertive, definitive, or absolute language like "undoubtedly," "we are very confident," "definitely," "no doubt," etc. Answer only 0 or 1.

Prompt 12: Find Assertive Language

#### Prompt 13: Find Reference of Past Events

##### SYSTEM PROMPT:

You are an economic and finance research assistant. You will be given a passage of analyst reports, and your task is to identify if the analyst talks about any historical events or episodes about the firm. Note that I am not looking for current events or events that happened in the recent past.

##### USER PROMPT:

Here's a passage from the analyst report:

{Context}

Does the analyst talk about any historical events or episodes? Answer 0 if no, and 1 if analyst talks about historical events or episodes about the firm. Please don't answer any current events or events that happened in the recent past. Answer 0 or 1 only.

Prompt 13: Find Reference of Past Events

#### Prompt 14: Find Reference of Consensus Forecast

##### SYSTEM PROMPT:

You are an economic and finance research assistant. You will be given a passage of analyst reports, and your task is to identify if the analyst talks about street or consensus estimates.

##### USER PROMPT:

Here's a passage from the analyst report:

{Context}

Does the analyst talk about consensus or street estimates? Answer 0 if no, and 1 if analyst describes "street estimates" or "consensus estimates". Answer only 0 or 1.

Prompt 14: Find Reference of Consensus Forecast

Prompt 15: Find Reference of Past EPS Forecast

SYSTEM PROMPT:

You are an economic and finance research assistant. You will be given a passage of analyst reports, and your task is to identify if the analyst talks about his previous forecasts or research reports. Note that I am not looking for current forecast.

---

USER PROMPT:

Here's a passage from the analyst report:

{Context}

Does the analyst talk about his previous EPS forecast? Answer 0 if no, and 1 if analyst talks about his previous forecast or past research report. Answer only 0 or 1.

Prompt 15: Find Reference of Past EPS Forecast

## OA.6 Information Extraction Example

**From Credit Suisse’s report of Bath & Body Works (BBWI) on 02/25/2022:**

*“Importantly, the rollout of the Loyalty program will be a -\$50m drag to revenues and gross profit dollars in 2H. BBWI’s 2022 guidance embeds SG&A dollars -1% to +4% YOY. But the company noted that its model includes \$75m of one-time costs to separate IT systems from Victoria’s Secret. Ex those costs, SG&A dollars are implied flat to -5% YOY. BBWI noted investments in 2022 in improving IT, but also to bring marketing spend back up to pre-COVID levels. On its face, we think it will be difficult to hold SG&A dollars flat or even down YOY given those investment priorities. But we’d note that BBWI is carrying the vast majority of the full corporate costs of the combined legacy been working to bring the total corp cost structure down to the right sized for the standalone BBWI organization (explaining the guidance for SG&A to be lower YOY in dollars).”*

**Factual content summarized by LLM:**

“BBWI has provided guidance for its 2022 performance. The company’s 2022 guidance embeds SG&A dollars -1% to +4% YOY. Additionally, BBWI is carrying the vast majority of the full corporate costs of the combined legacy companies and is working to bring the total corp cost structure down to the right size for the standalone BBWI organization. In 1Q, SG&A is guided down -3% YOY, partly due to lapping a \$25m charitable contribution in 1Q21 (which will likely repeat but at a lower rate this year). The rollout of the Loyalty program will be a -\$50m drag to revenues and gross profit dollars in 2H. BBWI is also investing in improving IT and bringing marketing spend back up to pre-COVID levels.”

**Subjective content summarized by LLM:**

“The analyst believes BBWI will have difficulty holding SG&A dollars flat or even down YOY due to investments in improving IT and bringing marketing spend back up to pre-COVID levels.”

**Information tuple extracted by LLM: factual content:**

- “SG&A”: “-1% to +4%”
- “Corporate costs”: “right size”
- “IT”: “improving”
- “Marketing spend”: “pre-COVID”

**Information tuple extracted by LLM: subjective content:**

- “SG&A”: “difficulty”

Table OA.12: Example of factual and subjective content summarization

## OA.7 Additional Examples of LLM Responses

### OA.7.1 Validity of identifying stock information

As a proof of validation, Figure OA.15 reports the fraction of IBES announcements for which the LLM can find the corresponding ticker, for which LLM can find the ticker’s EPS forecast, and for which the EPS forecast extracted from the textual content by the LLM exactly matches the record in IBES.<sup>56</sup> The LLM can correctly find the stock in the relevant context 91.4% of the time, and 77.3% of the time the LLM can find the numerical number for the analyst’s forecast. Furthermore, 56% of the time the numerical number extracted by the LLM exactly matches the IBES record. This high accuracy in extraction performance assures the LLM’s ability to read through the context as well as the relevance of the retrieved contexts.

I also conduct two placebo tests for the efficacy of using LLM to identify stock information. In the first placebo test, I replace the actual stock ticker with MSFT, and prompt the LLM to identify MSFT in contexts associated with announcements not related to MSFT. As shown in the second column of Figure OA.15, the LLM correctly reports that it cannot find MSFT nor its EPS forecasts. In the second placebo test, I replace the actual stock ticker with “S&P500”. While the LLM reports that it can find reference to S&P500 in many contexts (as many equity reports will mention what the overall market is doing), it cannot find EPS forecasts for S&P500, which are in fact non-existent in the context. These placebo tests validate that RAG with CoT will make LLM correctly report “I don’t know” when the information is in fact not present in the context.

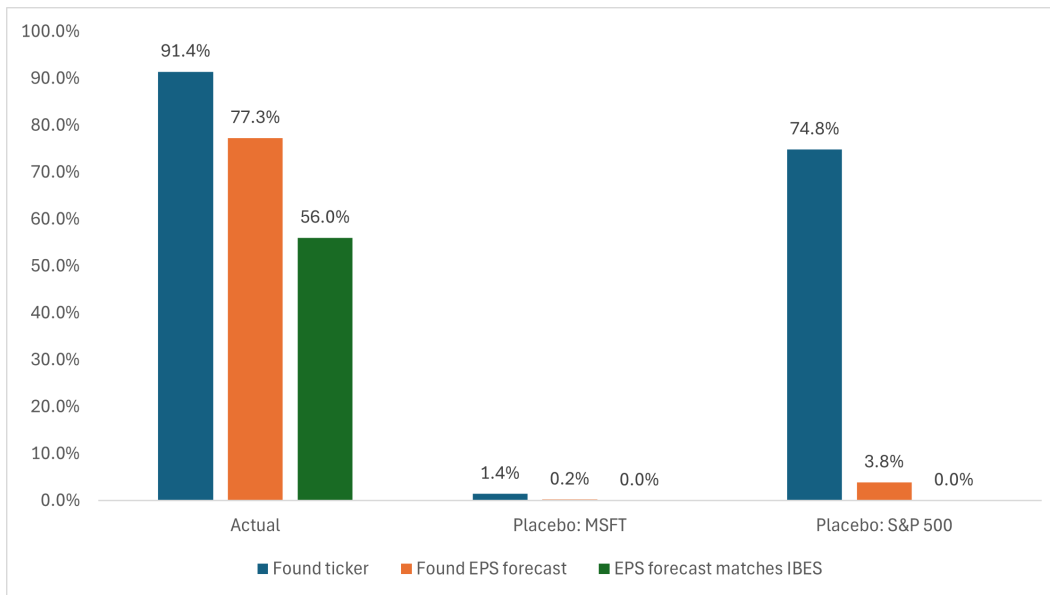


Figure OA.15: RAG retrieval validity of finding stock EPS information.

<sup>56</sup>Since the prompt does not specify the horizon for the analyst’s EPS forecast, I count an exact match if the extracted EPS forecast matches any EPS forecast across horizons.

## OA.7.2 Separating factual and subjective content

### **From JPMorgan's report of Incyte Corp (INCY) on 08/21/2021:**

*"We have a Neutral rating on INCY. We see long-term value in the company primarily driven by the Jak inhibitor, Jakafi, as well as an increasingly broad new product pipeline with late-stage shots on goal (notably tafasitamab and topical ruxolitinib). In the near term we look for greater visibility into these ongoing and/or pending new product launches to have greater conviction in the upside case for this story. INCY is well capitalized with \$2.1B in cash."*

### **Subjective content summarized by LLM:**

"The analyst believes INCY will have long-term value driven by the Jak inhibitor, Jakafi, as well as an increasingly broad new product pipeline with late-stage shots on goal (notably tafasitamab and topical ruxolitinib). In the near term, they look for greater visibility into these ongoing and/or pending new product launches to have greater conviction in the upside case for this story."

### **Factual content summarized by LLM:**

"Incyte Corporation is well capitalized with \$2.1B in cash."

### **From Wells Fargo's report of SunCoke Energy (SXC) on 05/01/2012:**

*"SXCP closed its acquisition of Lakeshore Coal Handling for \$28.6M in cash. The company provides coal handling and blending services to Indiana Harbor coke making operations, which are currently owned by SXCP's parent (SXC). The deal is expected to boost DCF by \$4M (\$0.12 per unit), which represents 6% accretion. More M&A to follow; another 19% of potential accretion. The Kanawha transaction, which also provides coal blending and handling services, closes in Q4 2013 and is expected to add \$0.18 per unit to DCF (9% accretive). We also estimate another \$0.20 per unit (10% accretive) related to assets drop downs from parent SxC as early as Q1 2014 (remaining stake in Haverhill and Middletown assets), SXCP a top pick. We continue to like SXCP due to M&A catalysts combined with attractive valuation. Post transaction, we estimate SXCP trades at 9X 2014 DCF vs. the MLP universe near 12X. We also like parent SXC, which should benefit from M&A at SXCP, and believe shares will trade better once the Indiana Harbor contract with ArcelorMittal is renewed."*

### **Subjective content summarized by LLM:**

"The analyst believes SXC will benefit from M&A at SXCP, which should drive its shares to trade better once the Indiana Harbor contract with ArcelorMittal is renewed."

### **Factual content summarized by LLM:**

"SXC owns Indiana Harbor coke making operations, which are serviced by Lakeshore Coal Handling. The company also has a remaining stake in Haverhill and Middletown assets."



**From KeyBank Corp's report of Watts Water Technologies (WTS) on 05/01/2012:**

*"Watts Water Technologies, Inc. WTS: Solidly In-Line 3Q15; High Impact Items Announced/Underway/Completed. Following 3Q15 earnings, we are reiterating our Overweight rating given the Company's encouraging operational execution and multitude of high-impact initiatives either announced, completed or underway. We continue to view WTS's accelerated DIY exit positively, and see meaningful earnings power ahead amid management's progressing Transformation efforts. Solid FCF and active capital allocation support WTS's internal improvement story, with favorable exposure to eventual N.A. and EMEA recoveries."*

**Subjective content summarized by LLM:**

"The analyst believes WTS will have meaningful earnings power ahead amid management's progressing Transformation efforts. The company's accelerated DIY exit is viewed positively, and solid FCF and active capital allocation support the internal improvement story, with favorable exposure to eventual N.A. and EMEA recoveries."

**Factual content summarized by LLM:**

"Watts Water Technologies, Inc.'s 3Q15 earnings were solidly in-line. The company has announced, completed, or is underway with various high-impact initiatives."

**From KeyBank Corp's report of Watts Water Technologies (WTS) on 05/01/2012:**

*"KeyBanc Capital Markets Inc. ("KBCML") does and seeks to do business with companies covered in its research reports. As a result, investors should be aware that the firm may have a conflict of interest that could affect the objectivity of this report. Investors should consider this report as only a single factor in making their investment decision. This report has been prepared by KBCML. "*

**Subjective content summarized by LLM:**

"N/A"

**Factual content summarized by LLM:**

"There is no factual information about WTS or WTS mentioned in the text that relates to the company's operations, business, and financials."

### OA.7.3 Extracting financial concepts

**Factual content summarized by LLM:**

Vishay expects its contract wins to generate revenues at an annualized \$50-\$60 million in the back half of 2014. The company's Capella business has a break-even point of about \$25-\$30 million in annual revenues, which suggests that it will be contributive to margins even at current revenue levels. Despite weak distribution trends and currency headwinds in 3Q14, Vishay kept its fixed costs under control and prevented significant margin contraction.

**Financial economic concepts extracted by LLM:**

- "Revenue": "\$50-\$60 million"
- "Contract wins": "annualized"
- "Break-even point": "\$25-\$30 million"
- "Fixed costs": "under control"
- "Margin": "prevented contraction"

**Factual content summarized by LLM:**

The factual information about HII (or H2) related to the company's operations, business, and financials is that there were operational issues in P&S-US. The Electronic Systems and Cyber segment performed better than expected, which offset some of these issues. There was also a change in the comment on debt from 'broadly unchanged' to 'little changed'. The company's underlying operating income for H1 came in at GBP874m, which is a 10% decrease compared to H1 17.

**Financial economic concepts extracted by LLM:**

- "Debt": "little changed"
- "Orders": "N/A"
- "Operating income": "10% decrease"

**Factual content summarized by LLM:**

MLTX has completed enrollment for its Phase 2b ARGO trial with a sample size of 200 patients ahead of schedule. The company plans to announce the topline 12-week data in the first half of November 2023. Additionally, MLTX is expected to host a Capital Markets Day in New York on September 11, 2023, to provide updates and frame expectations ahead of both the 12-week topline ARGO and 24-week MIRA readouts. MLTX ended 2Q23 with \$501.8M in cash and equivalents, which is sufficient to fund operations until 2026.

**Financial economic concepts extracted by LLM:**

- "Enrollment": "ahead-of-schedule"
- "Data readout": "N/A"
- "Cash": "sufficient"

## OA.8 Topic Classification

Topic	Sub-topic	Topic	Sub-topic	Topic	Sub-topic
Asset Management	Asset Growth	Asset Management	Asset Management	Asset Management	Financial Instruments
Business Strategies	Business Growth	Business Strategies	Business Operations	Business Strategies	Growth Strategies
Business Strategies	Legacy Business Elements	Business Strategies	Contract Negotiations and Agreements	Business Strategies	Digital and Business Transformation
Business Strategies	Pricing Strategies	Business Strategies	Websites and Online Presence	Business Strategies	Retail and Commerce
Business Strategies	Royalty Payments	Cash	Cash Flow	Cash	Cash Management
Cash	Free Cash Flow	Cash	Liquidity and Funding	Consumer Demand	Consumer Demand
Consumer Demand	Customer Base	Consumer Demand	Customer Engagement	Consumer Demand	Customer Membership
Consumer Demand	Demand	Consumer Demand	Subscriptions and Services	Corporate Leadership	Corporate Leadership
Corporate Leadership	Executive Personnel	Corporate Leadership	Management Guidance and Outlook	Corporate Leadership	Leadership and Governance
Corporate Management	Company Restructuring	Corporate Management	Business Investment	Corporate Management	Real Estate Operations
Cost	Cost Management	Cost	Costs of Goods Sold (COGS)	Cost	Expenses
Cost	Depreciation and Amortization	Cost	Operational Costs	Debt	Debt Management
Debt	Liabilities and Debt	Debt	Leverage	Debt	Credit and Lending
Debt	Interest Expense	Financial Planning	Financial Planning and Budgeting	Financial Planning	Working Capital Management
Financial Planning	Financial Management	Financial Planning	Financial Transactions	Funding	Funding and Capitalization
Funding	Equity and Financing	Funding	IPO	Human Resource	Education and Research
Human Resource	Employee Compensation and Benefits	Human Resource	Employee Management and HR	Human Resource	Human Capital
Human Resource	Pension and Benefits	Human Resource	Talent Management	Industry Landscape	Competitive Strategy
Industry Landscape	Industry Consolidation	Industry Landscape	Industry Peers	Industry Landscape	Industry Trends
Industry Landscape	Peer Comparison and Analysis	International Markets	Foreign Market	International Markets	International Trade
International Markets	Global Trade	International Markets	Currency Exchange	International Markets	Geopolitics
Inventory	Backlog and Orders Management	Inventory	Inventory Management	Legal	Compliance and Regulations
Legal	Litigation and Disputes	Legal	Legal Disputes	Legal	Regulatory Compliance
Legal	Financial Compliance	Legal	Insurance	Macroeconomy	Economic Conditions
Macroeconomy	Economic Impacts of Fiscal Policy	Macroeconomy	Economic Impacts of Monetary Policy	Macroeconomy	Economic Policies and Regulations
Macroeconomy	Inflation	Macroeconomy	Macroeconomic Trends	Macroeconomy	Interest Rates
Macroeconomy	Government	Macroeconomy	Demographics	Macroeconomy	Financial Crises
Macroeconomy	Central Bank	Macroeconomy	Mortgage and Real Estate	Marketing	Branding and Franchise
Marketing	Branding and Marketing	Marketing	Marketing	Marketing	Advertising and marketing
Mergers and Acquisitions	Acquisitions	Mergers and Acquisitions	Mergers and Acquisitions	Mergers and Acquisitions	Collaboration and Partnership
Mergers and Acquisitions	Synergies	Oil and Commodities	Oil and Energy Price	Oil and Commodities	Oil Price
Oil and Commodities	Commodities	Product Market	Market Expansion	Product Market	Market Opportunities
Product Market	Market Presence	Product Market	Market Share	Product Market	Market Trend
Product development	Product and Service Development	Product development	Product Development and Launch	Product development	Product Development and Pipelines
Product development	Product Lifecycle	Product development	Product Management	Production	Capacity Expansion
Production	Manufacturing and Production	Production	Production and Operations	Production	Productivity
Production	Raw Materials Management	Production	Mining and Resources	Production	Infrastructure Management
Profitability	Dividend	Profitability	Earnings	Profitability	Earnings Growth
Profitability	Profitability	Profitability	Profitability Margins	Profitability	Revenue
Profitability	Sales	Profitability	Sales and Revenue	Profitability	Sales Performance
Profitability	Net Income	Research and Development	Clinical Trials	Research and Development	Intellectual Property
Research and Development	Innovation and Development	Research and Development	Research and Development	Research and Development	Patent and Licensing
Research and Development	Technology Development and Management	Research and Development	Therapeutic Development	Research and Development	Medical and Healthcare
Risk	Cybersecurity	Risk	Environmental Impact	Risk	ESG
Risk	Social Responsibility	Risk	Risk Management	Risk	Bankruptcy
Risk	Delinquencies and Credit Risk	Risk	Disaster Impact	Risk	Pandemic
Risk	Weather Impact	Risk	Financial Risk	Risk	Volatility
Supply Chain	Logistics and Supply Chain	Supply Chain	Supply Chain Management	Supply Chain	Distribution Channels
Supply Chain	Shipments and Deliveries	Taxes	Taxation and Regulations	Taxes	Taxes
Taxes	Taxes and Benefits	Valuation	Stock Market Valuation and Performance	Valuation	Valuation and Multiple
Valuation	Valuation Methods				

Table OA.13: Topic Definitions

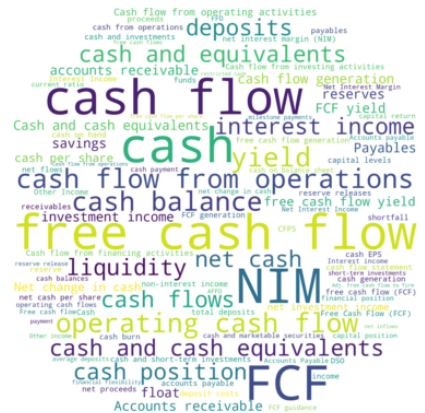
## OA.9 Word Clouds of 28 Meta-topics



## Asset Management



## Business Strategies



Cash



Customer Demand



## Corporate Leadership



Corporate Management



Cost



Debt



## Financial Planning



secondary offering  
equity offering  
shareholders' equity  
capitalization  
equity issuance  
fully diluted shares  
book value per share  
common stock  
equity value  
share repurchase  
authorization  
preferred stock  
equity ratio  
shareholders' equity  
common shares  
shares repurchased  
shareholders' equity  
initial public offering  
share authorization  
buyback  
equity  
dilution  
equity  
funding  
diluted share  
public offering

## Funding

Latin American  
China  
California  
Brazil  
Americas  
geographies  
Germany  
Japan  
Canada  
FX  
headwinds  
United States  
Latin America  
North America  
U.S. countries  
Asia  
Europe  
region  
regions  
international markets

## International Markets

purchase price  
merger  
integration  
deal  
partnerships  
joint venture  
acquisition  
synergies  
acquisitions  
partnership  
deals  
M&A  
activity  
relationship  
target  
purchase  
collaborations  
recent acquisitions

## M&A

commodity price  
oil price  
oil price  
gold price  
commodities  
natural gas prices  
proved reserves  
oil production  
crude oil  
commodity prices  
Eagle Ford  
drilling  
natural gas  
oil production  
steel prices

## Oil and Commodities

compensation costs  
incentive expense  
pension expense  
workforce reduction  
stock option expense  
headcount reduction  
stock-based compensation  
headcount  
stock-based compensation  
attendance  
incentives  
headcount reductions  
personnel costs  
workforce  
students  
staff layoffs  
stock-based compensation expense

## Human Resource

inventory management  
vehicles  
inventory days  
inventory levels  
backlog  
inventory turns  
inventory growth  
inventory turnover  
items  
total backlog  
new orders  
inventory build  
merchandise

## Inventory

oil prices  
natural gas prices  
inflation  
economic conditions  
FX headwinds  
funding costs  
market conditions  
economy  
FX headwind  
expense growth  
economic downturn

## Macroeconomy

product line  
products and services  
product introductions  
new products  
product offerings  
product mix  
product portfolio  
product launches  
product lines  
product offerings  
accessories

## Product Development

peer group  
competitors  
industry trends  
semiconductor  
peers  
competition  
strategy  
competitive pressures

## Industry Landscape

financial statements  
litigation  
FDA approval  
regulations  
combined ratio  
regulatory environment  
report date  
disclosures  
regulatory risk  
regulatory changes

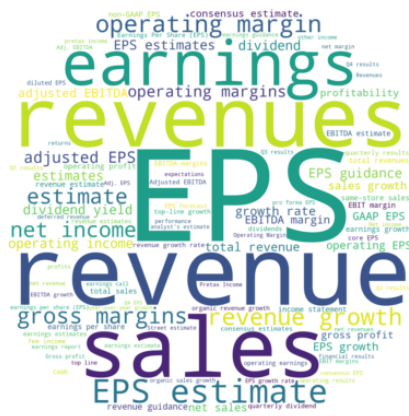
## Legal

advertising  
franchise  
brands  
brand  
promotions  
marketing  
sales and marketing  
advertising expenses

## Marketing

sales trends  
market share  
markets  
end markets  
prices  
market share  
average daily volume  
international sales  
mix shift

## Product Market



*Notes:* The size of the word represents the log frequency of the word in the topic.

## OA.10 Estimating Traditional Topic Models

To showcase the necessity of using LLMs in information extraction from analyst reports, I compare the performance of two traditional topic modeling approaches.

**Latent Dirichlet Allocation** First, I estimate Latent Dirichlet Allocation (LDA) models. LDA is a widely used model that decomposes the term-document frequency matrix into topics (distribution of words) and the document composition of topics. It is a simple model because it only uses the counts of vocabulary in documents, and does not utilize information on the semantics of the words.

To estimate LDA model, I select the longest analyst report in each month, and use the most frequent 5000 words as my vocabulary to construct the term-document frequency matrix. In the estimation, the topic number and Dirichlet prior parameters are hyperparameters. I vary the topic number to be from 2 to 200 and set the Dirichlet priors to be the  $1/\text{topic number}$ , the default choice in many LDA packages. To evaluate the model performance, I estimate the model using half of the documents, and I compute the perplexity score and (negated) log-likelihood of the other half of the documents. Topics are better when both scores are low. Figure OA.16, however, shows that both scores increase in the number of topics, and the global minimum occurs with only eight topics. This suggests the term-document frequency matrix is not informative enough for the LDA to discover more granular topics.

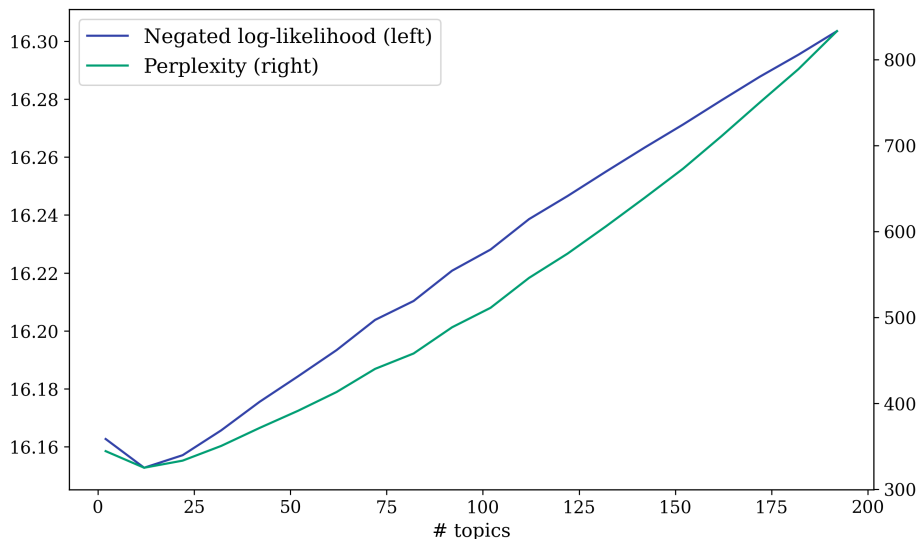


Figure OA.16: Perplexity and Log-likelihood of LDA models with different number of topics

**Embedding clustering** Another popular method to generate topics is to cluster words based on their embedding vectors. As embedding vectors capture semantic meanings of words, unsupervised clustering algorithms can be performed to search and group similar words together, which forms



topics in the semantic space. However, this approach requires clustering structure of words in the embedding space. In Figure OA.17, I report the Hopkins statistics of the embedding vectors of the most frequent 5000 words.<sup>57</sup> Because the Hopkins statistics only works for low dimensional vectors, I consider the top 5 and 10 principal components of the embeddings. I find that for all embedding models considered, from simple Word2Vec to complex Llama3.1, the Hopkins statistics are lower than 0.9, which is the common threshold to reject that the embedding vector distributions are different from a random uniform distribution.

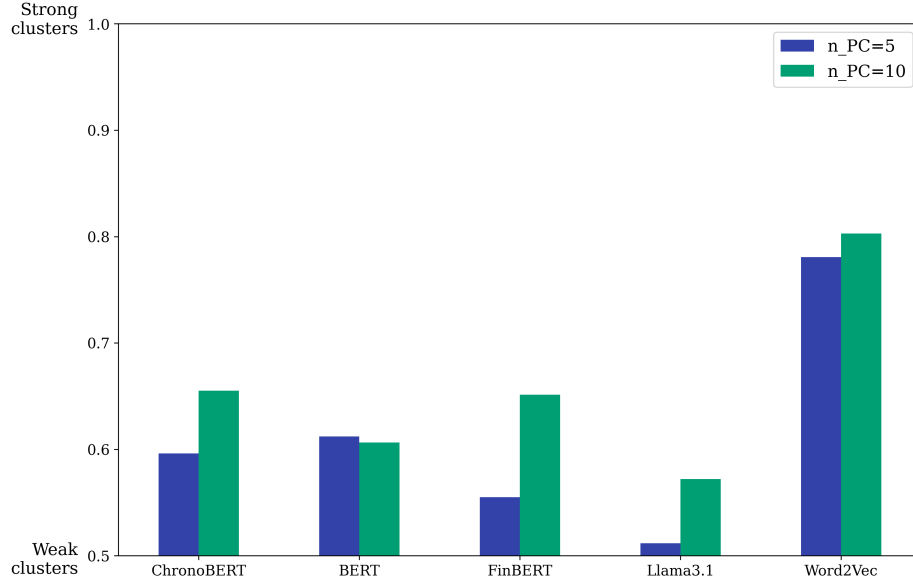


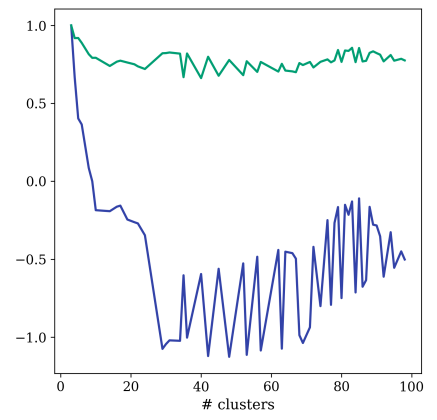
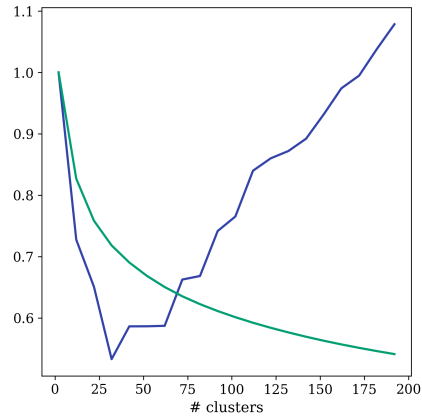
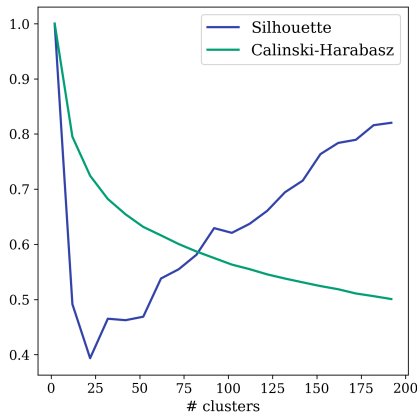
Figure OA.17: Hopkins Statistics of Different Embedding Vectors

In addition, I perform clustering algorithms including K nearest neighbors (KNN), agglomerative hierarchical clustering (AggHier) and DBSCAN on the embedding vectors, and I compute the Silhouette and Calinski-Harabasz scores with different cluster numbers. These two scores measure the internal similarity and external dis-similarity of the embeddings within and across clusters, and higher scores are better.<sup>58</sup> Figure OA.17 shows, that for most the embeddings and clustering algorithms considered, the highest score occurs when the algorithm only estimates two clusters. This indicates that these algorithms cannot cluster words into meaningful groups based on their embeddings alone when we look for granular clusters.

<sup>57</sup>Hopkins statistics is defined as  $H = \frac{\sum_{i=1}^d u_i}{\sum_{i=1}^d u_i + \sum_{i=1}^d w_i}$ , where  $w_i$  is the distance from  $x_i$  to its nearest neighbor in dataset  $X$ , and  $u_i$  is the distance from  $y_i$  generated from uniform distribution to  $X$ , and  $d$  is the data dimensionality.

<sup>58</sup>Silhouette score is defined as  $s(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}}$  where  $a(i)$  is average distance between  $i$  and all other points in the same cluster, and  $b(i)$  is the smallest average distance between  $i$  and all points in any other cluster, and the aggregate score is computed by averaging  $s(i)$  across data points. Calinski-Harabasz index is defined as  $CH = \frac{BCSS/(k-1)}{WCSS/(n-k)}$ , where  $BCSS$  is the between-cluster sum of squares and  $WCSS$  is the within-cluster sum of squares. For each score, I normalize it by the score with  $k = 2$  clusters for better comparison.

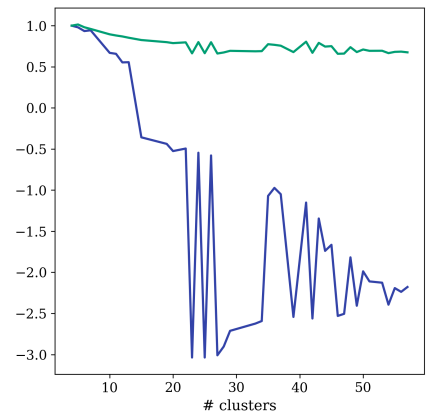
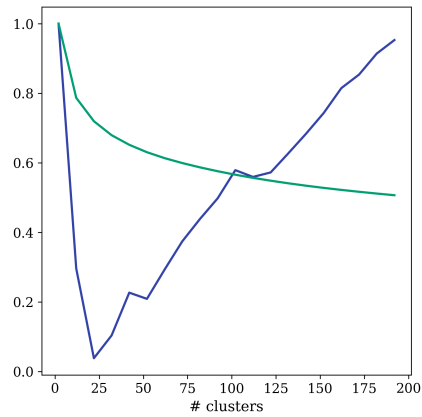
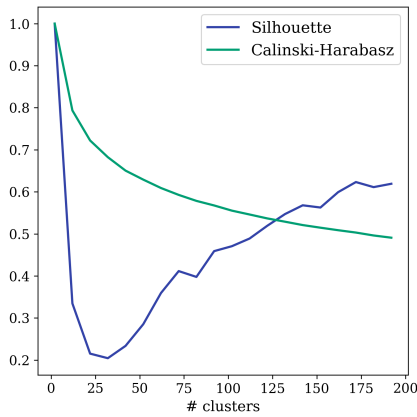




KNN: ChronoBERT

AggHier: ChronoBERT

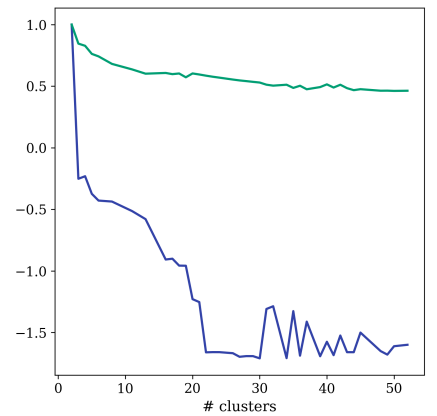
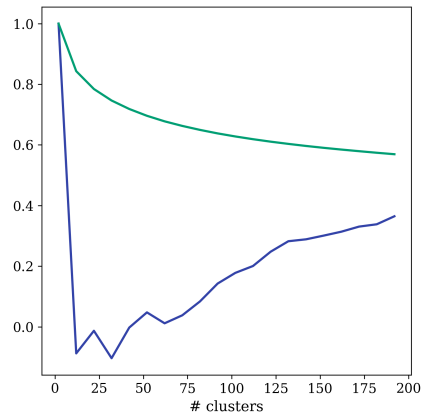
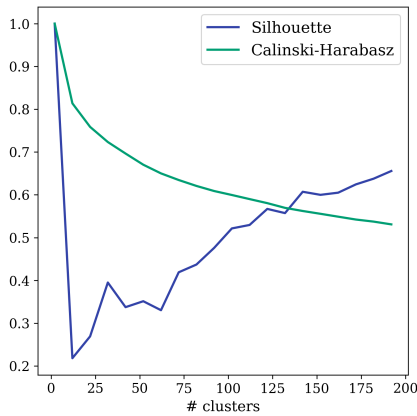
DBSCAN: ChronoBERT



KNN: BERT

AggHier: BERT

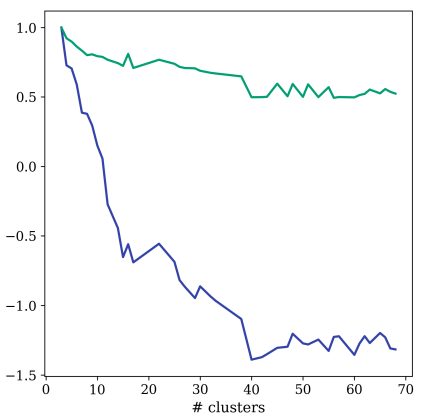
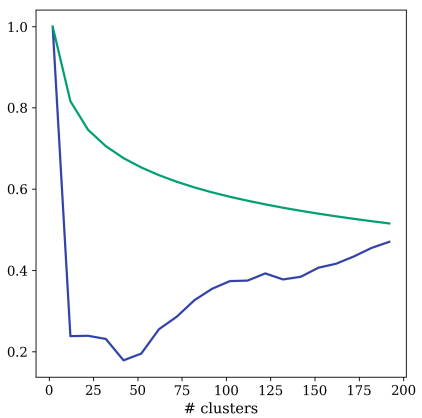
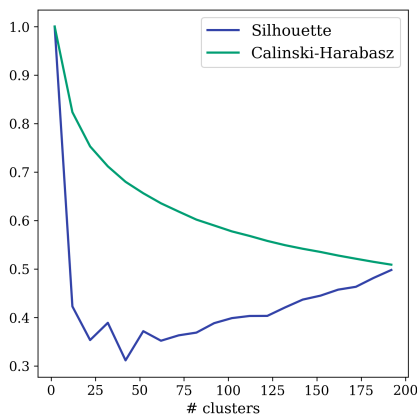
DBSCAN: BERT



KNN: Word2Vec

AggHier: Word2Vec

DBSCAN: Word2Vec



KNN: Llama3.1

AggHier: Llama3.1

DBSCAN: Llama3.1

Figure OA.17: Model performance of embedding clustering algorithms

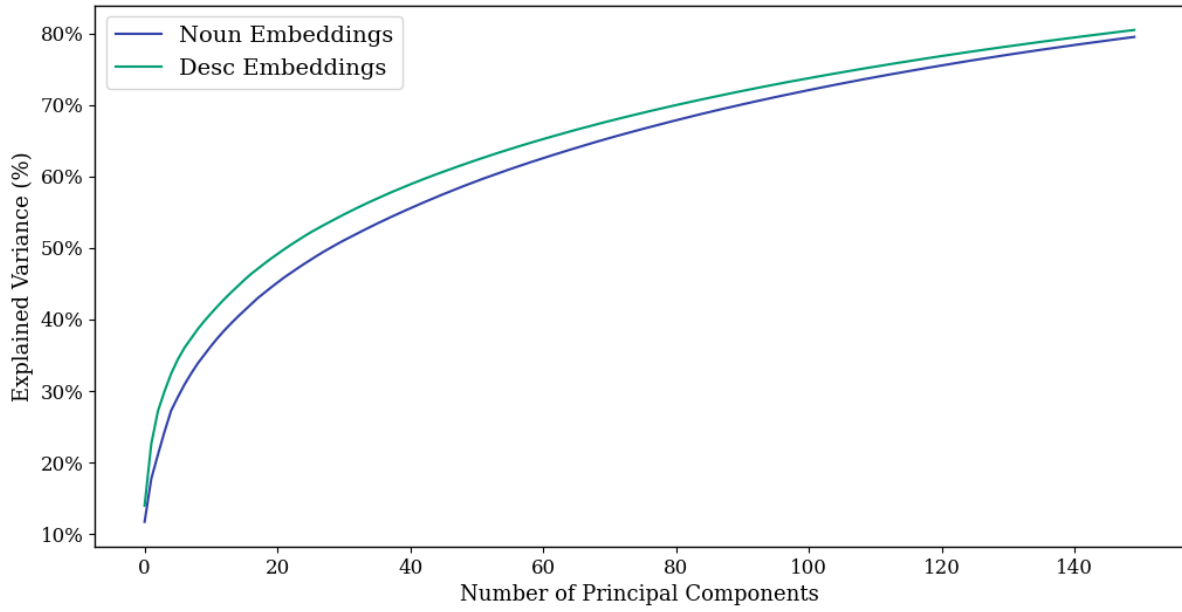


Figure OA.18: Explained variance ratios of topic principal components of embeddings

*Notes:* This plot shows the ratios of explained variance of topic principal components of the embeddings vectors for all *Nouns* and *Descriptions*, respectively.

Profitability characteristics			
sales	Revenue	fml_gr1a	1-year financial liability growth
sale_gr1	1-year Revenue growth	dbnetis_at	Debt issuance to asset
sale_be	Revenue to book equity	debt_gr1	1-year debt growth
net_income	Net income	debtlt_gr1a	1-year long-term debt growth
ni_gr1a	1-year net income growth	dlnetis_at	Long-term debt issuance to asset
ni_be	Net income to book equity	debtst_gr1a	1-year short-term debt growth
gp_sale	Profit margin	dstnetis_at	Short-term debt issuance to asset
dgp_dsale	Profit margin change	dlnetis_mev	Long-term debt issuance to market equity
ope_gr1a	1-year operating profit growth	dstnetis_mev	Short-term debt issuance to market equity
ope_be	Operating profit to book equity	dbnetis_mev	Debt issuance to market equity
ope_bell	Operating profit to lagged book equity		
cop_bev	Operating cash flow to book equity		
cop_mev	Operating cash flow to market equity		
ebit_gr1a	1-year EBIT growth		
ebit_sale	EBIT to revenue ratio		
ebitda_gr1a	1-year EBITDA growth		
ebitda_sale	EBITDA to revenue ratio		

Table OA.14: List of stock characteristics as proxies for Profitability and Financial Conditions

## OA.11 Examples of Overreaction to Qualitative Information

### JCPenney (JCP) and its new CEO

**Background and news** After experiencing lack of sales growth, JCP hired Rob Johnson from Apple as its new CEO, who introduced a multi-year turnaround roadmap, including a new pricing strategy (“price clarity”) in 2011.

**Analyst’s reaction** Matthew Boss at JPMorgan raised 2012 EPS forecast to \$1.62 from \$0.89 and 2013 forecast to \$2.76 from \$1.98.

**Reason for revision** Boss noted that *“Rob Johnson spoke to strong initial traction and increased vendor base buy-in to the company’s new pricing strategy”* and that (JPMorgan) *“believe the setup remains favorable with positive near-term catalysts in place (additional brand announcements, Town Square center core details, timing of top-line traction)”*. He wrote that *“every journey begins with a first step”*.

**Result** JCP had a 22% sales decline by mid-year 2012 and cut 600 more staff in April. Rob Johnson was fired in April 2013.

---

### General Motors (GM) and Cruise

**Background and news** Cruise was a self-driving car company that was acquired by General Motors (GM) in 2016, focusing on producing a fleet of driverless taxis. On December 1, 2017, GM held an investor event focusing on the recent developments on AVs.

**Analyst’s reaction** Guggenheim Securities upgrade GM to BUY and raise price target from \$48 to \$52.

**Reason for revision** Guggenheim noted that *“We expect GM to showcase the advanced status of its technology, which could give it a competitive advantage in putting autonomous vehicles on the road and monetizing the ride. GM has expressed notable optimism about the status of its AVs, and seems confident it can launch a ride-hailing fleet of robo-taxis well before 2020 for geo-fenced urban areas.”*.

**Result** GM’s stock traded in the \$30-40 range for the next 3 years. Its robo-taxi fleet never come to life, and in December 2024 GM stopped funding Cruise.

---

### Twitter

**Background and news** Twitter announced a better-than-expected Q3 result on Oct 26 2015, but provided a lowered guidance for Q4.

**Analyst’s reaction** on Nov 16, 2015, Roth Capital Partners lowers its 2016 EPS forecast to \$0.51 from \$0.65.

**Reason for revision** Roth Capital Partners noted that “A difficult user experience, high attrition and declining engagement continued to be challenges.” and that “We have not noticed improvement in matching content to our personal interests in our anecdotal experience.”

**Result** Twitter had a \$0.57 EPS for 2016. Roth Capital Partner’s revision from \$0.65 to \$0.51 therefore is an overreaction.

## OA.12 Examples of Behavioral Mechanism Detection

### OA.12.1 Confidence

Jeffrey Swarz from Credit Suisse for Gilead Science (GILD) on 6/12/1998:

*That process typically involves an evaluation of the potential efficacy and safety of that new drug, the potential economic impacts of adoption of the drug, where that drug should be adopted, and how that drug compares with existing therapies whether they are drugs or other forms of intervention. **This process is the reason that we are very confident patient, a really refractory unstable angina patient, go ahead and use aggressive therapy.***

---

Ellen Zickmann from William Blair for Abercrombie & Fitch (ANF) on 7/28/1999:

*Product is secondary to maintaining the brand. The clothes are just one of many vehicles used to convey the branding message. Other components critical to creating the image are: 1) its energetic stores staffed with great-looking “brand reps”, 2) its edgy, skin-bearing magalog, the A&F Quarterly, 3) its newly created catalog, 4) its Web site, and 5) print advertising. Customers aspiring to be part of A&F’s lifestyle want to “hang out” in the stores, carry the shopping bag, read the magalog, and wear the logo. **We are very confident in management’s abilities to continue to successfully expand its current businesses and leverage its core competencies of branding and merchandising into new growth vehicles.***

---

Jonathan Braatz from Oppenheimer for Neogen Corp (NEOG) on 6/13/2002:

*Moreover, we have little doubt that the private sector will independently adopt tighter controls and testing procedures. **As a result, we are very confident that most food producers/processors/millers/importers/distributors/retailers will devote greater resources in the future to the detection and elimination of food pathogens before the food reaches the consumer***

---

Safa Rashtchy from Piper Sandler for Netflix (NFLX) on 8/6/2002:

*We believe some of the recent pullback in Netflix shares may have been caused by confusion regarding Blockbuster’s new subscription plan. However, due to the reasons above, we do not expect short- or long-term pressure on Netflix business from the Blockbuster roll out..... **We are very confident in our estimates, and we believe they may prove conservative, as the Company has significant leverage in its customer acquisition costs.***

---

Johan Bergtheil from JPMorgan for Freeport-McMoRan Inc (FCX) on 9/29/2004:

*Our commodity research team believes that the global copper supply/demand balance will go from its current deficit to equilibrium in 2005. **We are very confident in FCX’s increased copper production given management’s***

*plan for a group analyst mine tour in Indonesia next week—FCX's first in the past four years.*

---

Brian McKenna from JMP Securities for Carlyle Group (CG) on 5/4/2023:

*It was the first time we heard from Carlyle's new CEO, Harvey Schwartz, and he made it very clear that he was "not pleased" with first-quarter earnings, although he does believe there is a "massive" long-term potential for the firm (we are looking forward to hearing more from him in the future around how he sees the business evolving longer term). It will definitely take some time for Mr. Schwartz to start to make a noticeable impact broadly at the company (and he clearly cannot control what happens in the macro), but given his impressive tenure in the industry, we are very confident in his ability to lead the firm through this tough time and ultimately re-accelerate growth at Carlyle to a level more in line with the broader industry.*

---

## OA.12.2 Memory

Scott Reamer from TD Cowen for Amazon (AMZN) on 10/13/1998:

*Though we expect Amazon to reach that customer level sometime in mid-1999, we believe AMZN may be given more berth thanks to the structurally lower profit potential of the book business. We remember vividly when AOL was spending upwards of \$300 per subscriber, a condition that the market promptly rewarded by whacking 66% off the stock in five months in 1995.*

---

Beth A. Cariello from Deutsche Bank for Pfizer (PEE) on 11/16/1999:

*Depending on the quality and validity of the data, we believe the sales potential could be between \$200 million and \$500 million. We remember PFE's optimistic story on Tenidap several years ago, and do not want to fall into that same trap again, as the real FDA issues were far greater than PFE acknowledged.*

---

Gary J. Holdsworth from Wedbush for PetSmart Inc. (PETM) on 4/20/2004:

*Once upon a time in, the upstart dot.com pet stores sought to bury "brick and mortar" retailing in the "old economy" boneyard. For all their flash and style (we remember the obnoxious sock-puppet from Pets.com which became a symbol of the entire dot.com craze) at the end of the day, the dot.coms were surprised to discover that business plans that sought to make money shipping \$20, 20-pound bags of dog food around at negative margins just did not work once the VC money was burned through. Many of the dot.coms went bankrupt or were acquired and absorbed into their brick and mortar cousins. Today, PETM and PETC each sell products online, but each company's online sales represent less than 5% of total company sales.*

---

Kevin W. Sterling from BB&T Capital Markets for Fedex (FDX) on 11/29/2012:

*Could an extended strike cause chaos and panic? We remember in 2002 when shippers were frustrated and saw their goods sitting on vessels for days on end, thus many have diversified their supply chain, expanding ports of call to Oakland, Seattle, Houston, etc. However, shippers who still have exposure to the ports of LA and Long Beach may be in a bind if an extended strike occurs.*

---

James K. Wicklund from Credit Suisse on Weatherford International PLC (WFT) on 5/13/2013:

*There are a number of investors that are surprised the CEO is still there. We are not. Having done this for a*

*while, we remember when Bernard ..... were hailed for generating more shareholder value than any other OFS CEO for a decade. The board of directors approved the promotion of the CFO in 2006, which was the beginning of the dramatic change .....It would be a mistake to under-estimate Bernard. He is very smart, he has proven his ability to generate shareholder value, he has not missed the issues that have put the company where it is today. Credibility takes time to be rebuilt but a couple of quarters of clean, on-target results and no more black swans could do it.*

---

Sterling Auty from JP Morgan on Zscaler Inc. (ZS) on 5/27/2021:

*Monday at the conference we had the CEO of Palo Alto Networks who talked about the challenges of complex applications like treating that have difficulty utilizing proxy-based architectures over the cloud. ZS refuted this claim suggesting that properly built applications of all types absolutely can utilize and benefit the security focus architecture of a proxy-based system. **We have covered the cybersecurity industry for greater than 20 years and we remember back to the initial architecture arguments of the firewall vendors that utilize proxy-based architecture.** The common belief was that proxy-based firewalls like the Secure Computing Sidewinder offer greater security, but stateful inspection (later to become deep packet inspection or DPI) based firewalls offered greater performance.*

### OA.12.3 Reference to past beliefs

Nancy Benacci from KeyBanc for AVX Corp (AVX) on 6/10/2005:

*Although distributor inventory liquidation appears complete, our checks indicate that there is no desire on the part of distributors to restock inventory above current levels. AVX's questionable fiscal 4Q05 results also **raise concerns about our previous belief that AVX should see near-term benefits from any potential improving trends in the component industry and substantial operating leverage.***

---

Jimmy Bhullar from JPMorgan for Principal Financial Group (PFG) on 8/5/2008:

*The shortfall in deposits, in turn, pressured net flows, which declined from \$3.0 billion in 1Q08 to \$1.1 billion, below our \$1.9 billion forecast. **Despite our previous beliefs, we have yet to see a noticeable impact on deposits or lapses due to the weak economy and the decline in the equity market.** However, we remain concerned about the long-term impact on net flows and margins from increasing competition in the pension market and a steady shift in PFG's AUM mix toward more ESOP business and larger accounts*

---

Hideyuki Maekawa from Credit Suisse for Samsung (SSNLF) on 11/4/2008:

***We believe our previous belief that some yield loss and a supply disruption would turn the cycle is too optimistic now,** as demand is worsening and supply has been and is expected to be rather resilient in the face of the financial crisis.*

---

Mark Mahaney from RBC for Trade Desk Inc. (TTD) on 5/8/2020:

***We are stepping away on rich valuation and our previous belief that Ad Revenue would hold up better than it did because of programmatic and CTV.** Revenue turned out to be less resilient (see Exhibit 3) given the "switch off / switch on" nature of Programmatic and exposure to harder hit channels like Digital Display.*

---

David Lewis from Truist Securities on Unum Group (UNM) on 5/11/2001:

*We do not believe that was the case in 4Q00, but viewed part of 4Q00's robust growth a product of easy comparisons, a stabilized sales force, and easing competition. UNM's "more normal" 12% sales growth in 1Q01 confirms our previous beliefs.*

---

David MacDonald from Leerink for CVS (CVS) on 1/14/2002:

*At this point, we have not changed our estimates. However, given the sizable contribution from the hepatitis C product line (10-15% of revenues and 20-25% of profits), and a supply picture that may be choppy, we have less conviction in our previous belief that our 2002 estimates could have meaningful upside.*

---

David Bofit from CIBC for Interpublic (IPG) on 3/30/2003:

*Interpublic reiterated guidance for an organic revenue decline of 1%-4% in 2003. The filing confirmed our previous beliefs that the March convertible offering and tender offer for the zero coupon notes should alleviate some of the restrictions on Interpublic's cash usage.*

---

## OA.12.4 Reference to consensus forecasts

Ryan Brinkman from JPMorgan for Tesla (TSLA) on 7/20/2023:

*We are lowering our earnings estimates for Tesla shares today—and suspect the Street will as well—after the company's fourth straight operating profit miss vs. Bloomberg consensus, brought about again by the continued trading of profits for sales.*

---

Daniel Binder from The Buckingham Research Group for Costco (COST) on 7/20/2023:

*Original EPS guidance of \$0.46 to \$0.48 for the quarter had been based on a 3-5% comp store sales gain in 4Q00, but with comps falling below plan we believe the consensus will probably gravitate toward our more conservative EPS estimate of \$0.44 for the quarter. Since we had already lowered our EPS estimates early last week, we are maintaining our current forecasts for the quarter and for the full-year 2000.*

---

Kit Case from Southwest Securities for TTEC Holdings (TTEC) on 8/3/2021:

*While reported results were just \$0.01 below the consensus, management gave additional guidance on the expected future impact of enhansiv (\$0.02 per quarter). Our original model included a significant cash requirement through FY02, which we do not believe many other analysts had incorporated these costs into their models. As such, we believe the consensus estimates will be reduced closer to our numbers.*

---

Rupak Ghose from Credit Suisse for Hexcel (HXL) on 12/7/2007:

*To reflect a slightly more conservative outlook for 2008 (given the difficulty of timing the 787 and A380 ramp-up) we are reducing our 2008 EPS by \$0.03 to \$0.91—still within the guidance. That said, since we are more comfortable with the long-term outlook (as the company's expectations confirm many of our internal estimates), we are publishing our 2009 estimate, which calls for 38% growth to \$1.26—which is well above the consensus and we believe the street will push higher over the next 6-12 months.*

---

John McNulty from Credit Suisse for Ashland (ASH) on 9/21/2011:

*For F2012, in speaking with the company, we believe the expense will be \$40-50 mil (down from \$90-100 mil under the old policy), which will result in a similar boost of \$0.41 (using the midpoints). While our estimates for F2011/12*

*by new policy could potentially increase 12%/7.5% to \$3.90/5.86, we are not adjusting our estimates until they report their F4Q earnings, when we believe the consensus will also move up.*