Analysts' Belief Formation in Their Own Words

Shikun (Barry) Ke*

First draft: November 16, 2024
This draft: March 15, 2025

Abstract

I study the formation of analysts' subjective beliefs about firms' earnings using analysts' own written text from over 1.1 million equity research reports. Text in analyst reports strongly predicts analysts' forecast revisions and forecast errors. Using a Large Language Model, I distinguish between factual and subjective content and distill it into interpretable topics on firm fundamentals. I document three sets of novel findings regarding analysts' subjective beliefs. (1) I show that analysts' attention allocation varies significantly over business cycles, firms, and forecast horizons. Analysts pay more attention to profitability information during booms and pay more attention to financial conditions and macroeconomics during recessions. These patterns align with a model of rational inattention. (2) I introduce a novel text-instrumented Coibion-Gorodnichenko regression to study analysts' misreaction to specific information. I find a pervasive underreaction across topics in analysts' short-term earnings forecasts, while their overreaction in long-term forecasts is mainly significant for business operations, corporate management, and macroeconomic information. This pattern is consistent with a "story-statistics gap" in associative memory being an important driver of overreaction to qualitative, story-like information. (3) I find that both asymmetric information and differences of opinion contribute to disagreement in earnings forecasts. Together, these results offer new insights into the formation of subjective beliefs about firms' earnings.

Keywords: Earnings Forecast, Belief Formation, Analyst Reports, Large Language Model, Attention

*Yale School of Management (barry.ke@yale.edu). I thank Nick Barberis, Bryan Kelly, Stefano Giglio, Kelly Shue, Alp Simsek, Kaushik Vasudevan, Jeff Pontiff, Theis Jensen, Mohammad Pourmohammadi, and Tobias Moskowitz, as well as seminar participants at Oxford SMLFin Seminar and Yale for their comments. I thank Jared Flake and Kelvin Law for sharing the IBES brokerage matching data.

1 Introduction

A large body of research studies subjective beliefs captured in survey data. By examining these survey-based measures of beliefs, researchers have documented deviations from rationality in these beliefs and have shown that these deviations can explain important economic phenomena, such as variation in asset prices.[1] However, several questions remain open regarding the *belief formation* process: What information do agents pay attention to? What information do agents misreact to? Why do they misreact? And what explains disagreement in agents' beliefs? Answering these questions is essential for understanding agents' beliefs in reality, yet traditional numerical or categorical survey responses offer limited guidance. To address these questions, one needs to open the black box of agents' own thinking.

In this paper, I propose to study agents' belief formation by examining how they explain their own beliefs. Specifically, I study sell-side equity research analysts' subjective beliefs about stock earnings, using the written text from over 1.1 million equity research reports. These analyst reports provide detailed analyses, forecasts, and recommendations on stocks to help investors make informed decisions. The quantitative measures in these reports—such as earnings forecasts, price targets, and stock recommendations—are collected by third-party aggregators (e.g., IBES) and are often used by researchers as measures of subjective earnings or return expectations.[2] Consequently, written text in these reports—which provides the rationale behind analysts' quantitative forecasts—offers a great setting to study the formation of analysts' subjective beliefs. I match detailed earnings forecast records in IBES with their corresponding analyst reports and use these reports to examine the drivers of analysts' earnings forecasts.

To begin, I demonstrate that the text in analyst reports is informative about analysts' subjective earnings beliefs. I represent the text using high-dimensional embedding vectors, and use them to predict analysts' forecast revisions and forecast errors. I find strong out-of-sample predictability, indicating that analyst reports contain qualitative information related to systematic variation in analyst beliefs, as well as their systematic mistakes. I also find that the text in analyst reports is predictive of future realized earnings and returns, beyond analysts' numerical forecasts and observable stock characteristics.

Having established that analyst reports are informative about analysts' beliefs, I then use a Large Language Model (LLM) within a Retrieval Augmented Generation (RAG) framework to

[1] A non-exhaustive list of papers that study survey-based measures of subjective expectations includes Vissing-Jorgensen (2003), 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), Nagel and Xu (2022), Nagel and Xu (2023), Bordalo et al. (2024a), Bordalo et al. (2024b). See Adam and Nagel (2023) for a review of the use of subjective expectation data in asset pricing.

[2] For references, see Chen et al. (2013), Bouchaud et al. (2019), De La O and Myers (2021), Bordalo et al. (2024a), and Bordalo et al. (2024b) for using IBES EPS forecasts as subjective cash-flow expectations, and Bastianello (2022), Jensen (2024), and Andrei et al. (2023) for using IBES price targets as subjective return expectations.

extract and summarize detailed content from analyst reports. Compared to traditional textual analysis methods like dictionary-based searching and topic modeling, LLMs better capture semantics and comprehend natural language in a human-like manner. By leveraging LLMs in a RAG framework, where the model generates answers based on retrieved context, I am able to extract information with great flexibility.

I extract information using a multi-step procedure. In the first step, I retrieve relevant context about analysts' earnings forecasts from reports based on embedding similarities, and use the LLM to verify the relevance by extracting analysts' EPS forecasts. Validation results show that the LLM can accurately extract detailed information and correctly report "no results found" if the targeted information does not exist. In the second step, I use the LLM to summarize the text into factual content—–where analysts describe objective facts and actual developments in the firm, and subjective content—–where analysts make subjective statements about what they believe will happen to the firm in the future. In the third step, I use the LLM to represent the factual content as a collection of narratives, defined as {*Noun*: *Description*} tuples that captures detailed factual information. Finally, I use zero-shot classification to categorize the most common *Noun*s into 7 topics related to different aspects of a firm's fundamentals.[3]

With the separated factual and subjective narratives, as well as the topics associated with different aspects of a firm's fundamentals, I make three contributions to the study of analysts' subjective belief formation.

First, I study analysts' attention to different information. I measure analysts' attention to a particular fundamental component by computing the frequency of related *Noun*s in the factual content, and examine how their attention varies across time, stocks, analysts, and forecast horizons. I find that aggregate attention varies significantly over time and across business cycles: Analysts' attention to Profitability and Growth information is pro-cyclical, while their attention to Financial Conditions and Macroeconomic information is counter-cyclical.

There is also significant cross-sectional variation in attention. For example, analysts pay more attention to Profitability for value firms, and they pay more attention to financial conditions when they make earnings forecasts about firms with high debt levels. Finally, analysts pay more attention to Profitability when they are making short-term forecasts, while paying more attention to non-profitability information when making long-term forecasts.

Is the variation in attention rational? While it is difficult to provide a definite answer—because the true rational attention is unobservable—I provide evidence that the variation in analysts' attention is consistent with a rational inattention model. The model posits that, when an agent forecasts a target by learning about multiple fundamental components and faces a cost of attention,

[3]These components are Profitability, Financial Conditions, Business Operation, Corporate Management, Growth, Valuation, and Macroeconomics. See Section 4.2 for details.

the rational allocation will concentrate attention on fundamental components that have higher predictive power for the target. I provide empirical evidence consistent with this prediction: in the subsample of high analyst attention to Profitability or Financial Conditions, the respective realized fundamental component (using relevant observable stock characteristics as proxies) does have stronger predictive power for future earnings growth, compared with the subsample of low analyst attention.

Second, I revisit the previous literature's findings on analysts' systematic misreaction in their earnings forecasts, by investigating what information analysts misreact to. It has been established in the literature that analysts' earnings forecasts exhibit underreaction to information in their short-term forecasts (e.g. Bouchaud et al. (2019)), while their long-term forecasts exhibit overreaction (e.g. Bordalo et al. (2024b)). In these studies, researchers conduct Coibion-Gorodnichenko regressions (Coibion and Gorodnichenko (2015), henceforth CG-regressions) that regress forecast errors on forecast revisions, where a positive (negative) coefficient indicates under(over)-reaction to information. However, an important open question remains: What specific information do analysts misreact to?

As the analyst reports provide detailed factual content on the information the analysts observe when making their forecasts, they provide a great setting for studying misreaction to information at a more granular level. I develop a novel "text-instrumented CG-regression" that aims to capture misreaction to textual information on different topics. The procedure first estimates a mapping from textual information (captured by the embedding vector of the text) to forecast revision, and then regresses forecast error on the fitted forecast revision. Under the assumption that the textual embeddings provide sufficient statistics on the informational drivers of analysts' forecast revisions, the second stage regression coefficients can be used to identify the misreaction to different information.

I conduct text-instrumented CG-regressions in the data and document two significant patterns of misreaction in earnings forecasts. First, the underreaction in short-term earnings forecasts is pervasive across information types; it is manifested in significantly positive text-instrumented CG coefficients for five out of the six topics (except Corporate Management). Second, in their long-term forecasts, analysts significantly overreact to information related to Business Operation, Corporate Management, and Macroeconomics in their long-term forecasts, while the overreaction to Profitability is small and insignificant. Overall, these results highlight the heterogeneity in misreaction to different types of information. In particular, the results on long-term overreaction is consistent with the findings of Daniel and Titman (2006), who argue that investors mainly overreact to intangible information, rather than past tangible, i.e. accounting information.

Why do analysts have heterogeneous overreaction to different information? While it is difficult to provide a definitive answer, I propose one possible explanation: Analysts are more prone to overreaction when the information is presented in qualitative descriptions, rather than quantitative statistics. Graeber et al. (2024) finds a significant "story-statistics gap" in associative memory, such that agent's beliefs are more persistently influenced by information presented in qualitative forms (story) than quantitative forms (statistics), and easier associative recall can be associated with higher overreaction (e.g. Bordalo et al. (2023), Enke et al. (2024)). I show that the topics analysts overreact to the most—Business Operations, Corporate Management, and Macroeconomics—have fewer *Description*s that contain numerical values, indicating that these topics are mostly presented in qualitative forms. Furthermore, I run the original CG regression conditional on the fraction of qualitative information in analyst reports, and the result supports more overreaction to qualitative information.

In my third contribution, I investigate the drivers of analyst disagreement. Conceptually, disagreement can arise from either asymmetric information or differences of opinion based on the same information, but there has been no attempt in the literature to study which channel drives the observed disagreement empirically.

The separation of factual versus subjective content in the analyst reports provides a way of gauging which disagreement channel is more relevant empirically, as the differences in factual content capture asymmetric information while differences in subjective content capture different interpretations. By regressing pairwise absolute differences in earnings forecasts on the cosine distances between two analyst reports, I find that pairwise differences in short-term EPS forecasts are greater when there is more dissimilarity in both factual and subjective content. This result suggests that both the asymmetric information and differences of opinions are important drivers of disagreement.

Related Literature This paper connects to the large literature on subjective beliefs about earnings. Prominent belief formation models of subjective cash-flow expectations include sticky beliefs (Bouchaud et al. (2019)), belief in the reversal of earnings growth (De La O and Myers (2021)), diagnostic expectations (Bordalo et al. (2024a)), learning with fading memory (Nagel and Xu (2022)), noisy cognitive defaults (De Silva and Thesmar (2024)), among others. These papers build belief formation models based solely on empirical patterns in analysts' numerical earnings forecasts and are intentionally stylized. In contrast, this paper takes a different approach by directly examining how analysts explain their belief formation using their written text. By studying the text in analyst reports, I provide direct evidence on how analysts form beliefs in reality, shedding light on important aspects of belief formation that researchers cannot study by using only numerical measures of beliefs, such as analysts' attention and heterogeneous misreaction to information. This approach is similar to an emerging literature in behavioral economics that uses open-ended surveys (e.g Haaland et al. (2024)).

The findings in this paper connect to three aspects of subjective beliefs. First, they add to the literature on asymmetric attention (e.g. Sims (2003), Kohlhas and Walther (2021)) by directly measuring sell-side analysts' attention to different aspects of fundamentals and studying its variation. My finding on how analysts' attention varies across business cycles complements research on firm managers' time-varying attention to macroeconomic conditions (e.g., Song and Stern (2021), Flynn and Sastry (2024)). Second, it contributes to the assessment of misreaction in analysts' forecasts. The text-instrumented CG-regression provides a new avenue to examine what information analysts misreact to, an aspect that traditional tests like CG-regression cannot address. The finding suggests analysts have heterogeneous misreaction to different types of information, complementing the previous research on misreaction in earnings forecasts that uses only numerical forecast data (Bouchaud et al. (2019), Bordalo et al. (2024b)). The paper also provides evidence that the "story-statistics gap" in associative memory could be driving the overreaction to certain types of information (Graeber et al. (2024)). Finally, it contributes to the literature on disagreement by showing that both asymmetric information and differences in beliefs are important drivers of disagreement.[4]

This paper also relates to the study of textual information in analysts' written reports.[5] Despite their apparent relevance, analyst reports have not been systematically used to study the drivers of analysts' subjective expectations, likely due to challenges in text parsing and information extraction. Some recent papers use analyst reports to study how analysts form 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 Chinco (2024) focus on price multiples. In contrast, this paper studies the textual content in analyst reports and how it relates to analysts' beliefs about stock earnings.

Finally, this paper contributes to the emerging agenda on using generative AI and LLMs in economic research (see Korinek (2023) for a review). I show that using LLMs in a RAG framework enables researchers to flexibly extract information from unstructured documents that can be important for studying agents' beliefs. More broadly, this paper connects to the large literature on using machine learning in financial research (see Kelly et al. (2023) for a review). In particular, I contribute to the literature on using machine learning to study analysts' earnings forecasts (e.g. Van Binsbergen et al. (2023), De Silva and Thesmar (2024)) by showing that analysts' own written text is an important and informative source for studying the systematic mistakes in their forecasts.

[4]See Diether et al. (2002), Anderson et al. (2005), Jiang and Sun (2014), and Bali et al. (2023) for more references on empirical research on disagreement in stock markets.

[5]Asquith et al. (2005), Huang et al. (2014), and Huang et al. (2018) study investor reactions to analyst reports by looking at short-term stock price movements. Recent work by Bellstam et al. (2021) employs topic modeling to measure corporate innovation in analyst reports. Li et al. (2023) uses a RAG framework similar to mine to assess corporate culture. Chi et al. (2024) identifies the usage of alternative datasets in analyst reports and study how it affects forecasts.

Outline The paper is structured as follows. Section 2 introduces the text data in analyst reports and the RAG framework for information extraction. Section 3 presents a conceptual framework of attention and misreaction to information, and describes how text in analyst reports can be mapped to model quantities, including the measurement of attention and text-instrumented CG-regression. Section 4 details the implementation of information extraction using RAG. Section 5 studies the variation in analyst attention. Section 6 studies how analysts misreact to different types of information. Section 7 studies the drivers of disagreement in analysts' earnings forecasts. Section 8 concludes.

2 Data and Methodology

2.1 Data

2.1.1 Analyst Reports

An analyst report is the primary research output of a sell-side equity research analyst. It typically includes detailed textual content covering a wide range of topics, such as a stock's recent financial performance, business strategies, competition with industry peers, and risk exposures. The report also provides several quantitative measures, including earnings forecasts, price targets, and stock recommendations, which are collected by third-party data aggregators, such as IBES.

Analyst reports are reliable sources for studying analysts' belief formation. First, analyst reports must adhere to regulatory requirements, such as FINRA Rule 2241, which mandates that "purported facts in research reports are 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 favor the interests of the member or a current or prospective customer or class of customers."[6] Second, 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 justifications for their forecasts.[7]

[6]See https://www.finra.org/rules-guidance/rulebooks/finra-rules/2241. Similar requirements can be found in SEC Regulation Analyst Certification: https://www.sec.gov/rules-regulations/2003/02/regulation-analyst-certification. Regulators do enforce these rules and can impose fines or suspensions of analyst certification, including in cases involving major investment banks like Deutsche Bank and Barclays. Andrew J. Ceresney, a former SEC director of enforcement, said to 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."

[7]For example, a survey conducted by *Institutional Investor Magazine* shows that the written report is the fifth most important consideration when investors vote for "All-Star" analysts, while earnings forecast ranks only 12th. "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 more 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

I obtain sell-side equity research analyst reports from Investext, an aggregator provided by Mergent Online. Investext stores these reports in PDF format, with each publication accompanied by metadata that includes the report title, publisher (investment bank or brokerage firm), authors (sell-side analysts), company names and tickers mentioned in the report, 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 authors of the research reports with IBES Analyst Codes (`amaskcd`) using their first and last names, as well as stock coverage. Additionally, I match contributors of the reports with IBES Estimator IDs (`estimator`) based on name resemblance and analysts' affiliations (see Appendix A.4 for details). I focus on 115 brokerage firms where the mapping between Investext contributor and IBES Estimator IDs are validated by the matching tables provided by Flake (2023) and Law (2023). These brokerage firms include major investment banks like 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, for each announcement (identified by firm–estimator–announcement date triplet) made to IBES, I attempt to match it with research reports written by the same institution about the same stock within ± 1 business day of the announcement date. Appendix A.6 gives more details about 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, which covers about 50% of all announcements made by the 115 institutions in IBES.[8] 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. To process these reports, I use the `deep-document-understanding` pipeline developed by InfiniFlow. This pipeline uses various deep learning models in computer vision and text analysis, including Optical Character Recognition (OCR), Layout Recognition, Text Recognition, and Table Recognition. These steps ensure that textual information is preserved during parsing, thereby maintaining data quality. See Appendix A.2 for details on the parsing procedure.

an important part of their research process for generating trade ideas.

[8]The number of reports exceeds the number of announcements because multiple reports can be associated with a single announcement. For example, a broker may issue a report on an EPS announcement, and the following day publish a research digest summarizing the content of the original report. This is not an issue for my analysis, as the texts are split into chunks, and the retrieval step in RAG will retrieve the relevant text chunks from multiple reports.

2.1.2 Subjective Expectation Data

I obtain analyst forecast data from the Institutional Broker's Estimate System (IBES), a widely used data source for studying equity analysts' subjective expectations. My primary focus is on 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 cleaning and processing of IBES data largely follows the literature and is detailed in Appendix A.5. The sample period spans from 01/01/1998 to 09/30/2023 to align with the sample period for the Investext analyst report sample.

Other data used in this paper include stock price and return data from the Center for Research in Security Prices (CRSP), as well as stock characteristics from Jensen et al. (2023). These characteristics cover a broad spectrum of stock-level variables that have been studied in the literature.

2.2 Information Extraction in the Age of Generative AI

The primary goal of this paper is to study analysts' belief formation process by way of their written reports. Achieving this requires advanced natural language processing (NLP) techniques. Traditional textual analysis in economics and finance typically relies on a bag-of-words approach, where a corpus is represented as a collection of word counts. Researchers either count the frequency of specific words from a predefined dictionary or apply unsupervised topic modeling techniques, such as Latent Dirichlet Allocation (LDA), to identify a set of topics, each characterized by a distribution of word frequencies.[9] However, to fully understand analysts' justifications for their forecasts, one ideally wants to answer highly semantic questions, such as "What factual information about stock X does the analyst use to justify her EPS forecast?" or "What is the analyst's opinion about the future profitability of stock X?" Extracting such information requires NLP tools to *comprehend* the text and generate answers tailored to specific questions.

In this paper, I implement a Retrieval-Augmented Generation (RAG) pipeline that leverages recent developments in Large Language Models (LLMs) for information extraction. LLMs are statistical models designed to generate human-like text. They achieve this by predicting the next token in a sequence, given all previous tokens, using a neural network architecture known as transformers. With billions of parameters and training on vast corpora with diverse content and semantics, LLMs have shown significant improvements in context-based NLP tasks, including text summarization, zero-shot classification, translation, and question answering, reaching performance levels comparable to human benchmarks.[10]

[9] Examples of textual analyses that uses the dictionary-based approach include Tetlock (2007), Loughran and McDonald (2011), Hassan et al. (2019), and examples of topic modelling in economics and finance include Hansen et al. (2018), Bybee et al. (2024), and Ke et al. (2024), among others. See Gentzkow et al. (2019) for a recent review for these traditional approaches in textual analysis.

[10] See Brown et al. (2020) and Wei et al. (2022) for surveys of recent developments in large language models.

A Retrieval-Augmented Generation (RAG) framework is a novel AI approach that combines generative LLMs with traditional information retrieval methods to provide advanced retrieval results based on the context of the input corpus. The framework consists of two components: a "retrieval" component that retrieves the most relevant textual chunks based on their contextual similarity to a user-defined query, and a "generation" component that uses an LLM to answer the question based on these relevant chunks. Compared to bag-of-words approaches, RAG retrieves information based on the contextual meaning of the text, simulating a retrieval process that mimics human comprehension and judgment. Figure 7 in Appendix A.7 illustrates the RAG workflow.

In this paper, I use `Llama-3.1-8b-Instruct` as the generator in RAG. `Llama-3.1-8b-Instruct` is a recent LLM developed by Meta AI that has demonstrated state-of-the-art performance in various textual tasks that is comparable to other LLMs such as GPT, Claude, and Mistral, while the open-source nature of `Llama-3.1-8b-Instruct` allows me to flexibly deploy the model locally, run inferences without limits, and gain full control of the generation process.[11] In Section 4 and Appendix E, I provide further details on the RAG implementation, including the retrieval technique and the various prompts used.[12]

2.3 The Informativeness of Analyst Reports

The purpose of this paper is to study analysts' belief formation by examining their written reports. This study is based on the premise that analyst reports provide valuable insights into analysts' subjective earnings beliefs. In this section, I demonstrate the usefulness of textual content in analyst reports by showing that the text embeddings predict analysts' individual forecast errors. I also show that the embeddings predict future realized earnings and returns, beyond analysts' consensus forecasts and common stock characteristics. As these analyses are intended to showcase the general usefulness of analyst reports without specific information extraction, in this section I only summarize the main takeaways briefly. Full details are presented in Appendix B.

Embedding representation To represent the textual information in analyst reports, I obtain the embeddings of the context using `Llama-3.1-8b-Instruct`. An embedding is a (potentially high-dimensional) vector of numerical values designed to capture the contextual meaning of the text.[13] In `Llama-3.1-8b-Instruct`, each context provided to the model is embedded in a 4,096-dimensional space. Therefore, I obtain a 4,096-dimensional vector $X_{i,j,t}$ for each report written by analyst i about stock j at time t. To mitigate the concern that an analyst's numerical forecasts are also included in the report, I remove all numerical symbols from the text before getting the

[11] See https://ai.meta.com/blog/meta-llama-3/ for description of `Llama-3.1-8b-Instruct`.

[12] Some recent literature raised concerns about using LLM in empirical research pipelines, including issues on reproducibility and look-ahead bias. I address some issues in Appendix A.8.

[13] See Liu et al. (2020) for a survey of contextual embedding techniques and Touvron et al. (2023) and MetaAI (2024) for specific training details of `Llama-3.1-8b-Instruct`.

embedding vectors.[14]

Predicting forecast error In the first exercise, I predict an analyst's individual forecast error $Error^h_{i,j,t}$ with the associated embedding of his report $x_{i,j,t}$

$$Error^h_{i,j,t} = \gamma_h' X_{i,j,t} + \epsilon_{i,j,t} \tag{1}$$

where $Error^h_{i,j,t}$ is defined as the difference between realized h-year ahead EPS and analyst i's forecast. Details on variable definition, sample selection, training and evaluation are presented in Appendix B.1. I find that the text-based prediction of forecast error achieves an out-of-sample R^2 ranging from 0.08 to 0.30 for different forecast horizons, and the predictability is not spanned by forecast revision, a usual predictor for forecast error predictability.

This result suggests that text in written reports is informative about analysts' behavioral biases. For example, if an analyst is usually too optimistic when a firm launches a new product, the predictive regression (1) will pick up a relationship between a negative forecast error and an embedding that captures text saying "new product launch". The association between textual content and systematic errors leads to the predictability.

Remark on institutional incentive distortion While this paper focuses on belief formation and proposes misreaction to qualitative information as the explanation for the forecast error predictability, an alternative explanation for systematic forecast errors is that analysts' numerical forecasts are distorted by institutional incentives in the equity research industry.[15]

While it is not possible to entirely rule out this explanation, one would expect that if these institutional incentives align with observable analyst or brokerage characteristics, we would see differences in the forecast error predictability for different types of brokerages. Table 21 in the Appendix shows, however, that there is no difference in the overall out-of-sample R^2 between experienced and inexperienced analysts, nor between large and small brokerage firms. A comprehensive study of whether institutional incentives distort the written text in analyst reports is beyond the scope of this paper, and I leave it for future research.

Predicting future earnings In addition to predicting individual forecast errors, in Appendix B.3 I show that the textual information in analyst reports is useful for predicting future realized

[14] Most of the analyst reports can be fit into the 128k token context window in `Llama-3.1`. I truncate excess tokens if the report contains more than 128k tokens.

[15] Different institutional incentives might push analysts to issue more or less accurate forecasts. For example, analysts might issue too optimistic forecasts in order to win underwriting business or to please the firm management in order to gain access to private information (e.g. Michaely and Womack (1999), Mayew (2008), Kolasinski and Kothari (2008)). On the other hand, reputational concerns and job outcomes related to forecast accuracy push analysts to issue unbiased forecasts (e.g. Hong and Kubik (2003), Groysberg et al. (2011)).

earnings, and that the predictability gain is beyond what can be obtained from the consensus IBES earnings forecast or from common stock characteristics.

Predicting future returns Finally, in Appendix B.4 I show that the textual information in analyst reports is predictive of future individual stock returns. An equal-weighted decile spread based on text-based 12-month future return prediction achieves an annualized Sharpe ratio of 0.82, representing a 36.5% increase compared to a similar decile spread formed based on IBES's consensus price-target implied expected return. The text-based strategy achieves significant alphas relative to the price-target-based strategy, as well as relative to asset pricing models. These results suggest that there is additional return predictability in the textual information in the analyst reports that adds significant economic value beyond the numerical price target predictions.

3 A Framework for Belief Formation and Text

Before diving into information extraction and the textual analysis of analysts' belief formation, in this section I present a theoretical framework for thinking about belief formation, which features disaggregated information sets and attention choice. In Section 3.2 I discuss how the empirical models of text map to the variables of interest in the model, and in Section 3.3 I introduce text-instrumented Coibion-Gorodnichenko regression, a novel procedure that leverages text to identify an analyst's misreaction to different types of information. Finally, in Section D.1 I solve the inattention model with rational expectation, and derive implications for earnings and return predictability. Together, this section guides the analysis of analyst attention in Section 5 and misreaction to information in Section 6.

3.1 Model Setup

The framework is motivated by Kacperczyk et al. (2016) and Kohlhas and Walther (2021). For simplicity, I describe the model for the case where one analyst is trying to make a forecast for one firm. Let y_{t+1} denote the firm's future earnings that the analyst is trying to forecast. The future earnings is driven by the sum of K structural components $y_{k,t+1}$,

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

These structural components represent different 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, he does not treat the firm's earnings as a single variable. Instead, he first builds a financial model to make projections about different components in the income statement, and then aggregates. In the analyst's financial model, the firm's earnings (net income) can thus be (roughly) represented as

$$\text{Earnings} = \text{Revenue} - \text{COGS} - \text{Operating Expense} - \text{Interest Expense} - \cdots$$

which resembles the formulation in (2). I assume that each structural component is driven by a latent factor $\theta_{k,t}$ such that

$$y_{k,t+1} = a_k \theta_{k,t}. \tag{3}$$

In this formulation, a_k denotes the importance of component k. A larger a_k (in absolute value) means that component k contributes more to future earnings. One can think of this as capturing different determinants of future earnings: for a firm that is revamping its business operation, the operating expense might matter more for future earnings, while for a firm that relies more on external financing, the interest expense might matter more. As we see later, different importance parameters a_k determines the agent's relative attention allocation.

I assume the latent factor $\theta_{k,t}$ evolves according to the autoregressive process

$$\theta_{k,t} = \rho_k \theta_{k,t-1} + \sigma_\theta \epsilon_{k,t} \tag{4}$$

where $\epsilon_{k,t}$ are standard normal shocks. For simplicity, I assume all latent factors are uncorrelated ($\epsilon_{k,t} \perp \epsilon_{k',t}$) and their shocks have the same variance σ_θ^2. I study correlated fundamental components in Appendix D.5.

Belief formation The analyst observes a signal $s_{k,t}$ of the latent factor about component k. The signal takes the form of

$$s_{k,t} = \theta_{k,t} + u_{k,t}; \quad u_{k,t} \sim N(0, \tau_k^{-1}) \tag{5}$$

where τ_k is the precision (or attention) in the analyst's signal about the k'th component. I define the *relative attention* the analyst pays to component k as the ratio between τ_k and total precision across all K signals

$$m_k := \frac{\tau_k}{\sum_{k'=1}^{K} \tau_k'} \tag{6}$$

Given that the earnings are the sum of K structural components and the independence assumption of $\theta_{k,t}$ across k, I model the analyst's subjective forecast about earnings as the sum of his belief revisions about the K structural components. The reduced-form belief formation model follows

$$\begin{aligned} F_t[y_{t+1}] &= \sum_{k=1}^{K} F_t[y_{k,t+1}] + \epsilon_t \\ F_t[y_{k,t+1}] &= \mathrm{E}_t[y_{k,t+1}] + A_k(\mathrm{E}_t[y_{k,t+1}] - \mathrm{E}_{t-1}[y_{k,t+1}]) \end{aligned} \tag{7}$$

where $F_t[\cdot]$ denotes the analyst's subjective belief. The second line specifies a general belief updating rule, where $\mathrm{E}_t[\cdot]$ denotes the rational expectation conditional on the analyst's time t information set $\Omega_t = \{s^0 \cup (s_{1,\tau}, \cdots s_{K,\tau})_{\tau=1}^t\}$, which is the history of her past signals.[16] A_k denotes the analyst's potential deviation from rational reaction to information: if $A_k > 0$, his belief overreacts to information about k relative to the rational expectation, and vice versa for underreaction. ϵ_t denotes beliefs due to reasons other than the information about these K components, such as optimism bias, reaction to information not included in the K components, institutional incentive distortion, or simply noise in expectations. This belief formation specification also implies that the forecast revision can be decomposed into the revisions due to different structural components

$$Rev_t[y_{t+1}] = \sum_{k=1}^{K} Rev_t[y_{k,t+1}] + \tilde{\epsilon}_t \tag{8}$$

where $Rev_t[\cdot] = F_t[\cdot] - F_{t-1}[\cdot]$ denotes the revision operator.

This belief formation model can be directly mapped to diagnostic expectation that stems from representative heuristic (Bordalo et al. (2020)). Importantly, the heterogeneous A_k coefficient relaxes the usual diagnostic expectation model by allowing for different severity of judging by representativeness for different fundamental components. In Appendix D.2, I provide alternative belief formation models for over/under-estimation of signal precision τ_k (where overestimating signal precision can come from overconfidence), belief stickiness, extrapolative beliefs, and misperception about component importance a_k. In each case, I show that the text-instrumented CG-regression coefficient introduced in Section 3.3 can be interpreted as over/underreaction due to the specific behavioral distortions in each model.

3.1.1 A Rational Inattention Benchmark

In Appendix D.1, I solve for a rational attention allocation, where the analyst chooses signal precision τ_k in order to maximize his expected forecast accuracy, but incurs a cost for acquiring a more precise signal. Proposition 1 shows that the key prediction for a rational attention allocation is that, the analyst should allocate more attention to a fundamental component with higher predictive power of future earnings. Formally, denote the rational attention for component k to be m_k^*, we have

$$\frac{\partial m_k^*}{\partial a_k} > 0. \tag{9}$$

In Section 5.4, I will directly investigate how earnings predictability changes with analyst's attention. In Appendix D.4 I also study the implication of rational inattention for return predictability,

[16] Following the convention in the literature, I assume that, in the initial period $t = 0$, the analyst has received an (infinitely) long sequence of signals. This ensures that the analyst's signal extraction problem is initialized in steady state.

and show that the relationship between rational attention and return predictability from $\theta_{k,t}$ can be non-monotonic.

3.2 Uncovering Belief Formation from Text

The analyst reports provide detailed context on how analysts form beliefs, but text imposes an additional layer of difficulty: the researcher needs to first extract information and then represent it in numerical forms in order to study belief formation. In this section, I discuss how information extraction and text representation in Section 4 map into the theoretical framework introduced above.

First, the K structural components that drive the firm's earnings can be interpreted as different aspects of a firm's fundamentals—or "topics"—that the analyst describes in his report. These topics can further be viewed as a collection of nouns. For example, when the analyst writes about "revenue," "profit margin," or "sales," he is describing information about the firm's profitability, whereas when he writes about "inventory," "products," or "customers," he is describing information about the firm's business operations. I thus represent an analyst's report as a collection of $\hat{s} = \{Noun : Description\}$ tuples, where *Noun* is a noun word that is classified into K topics, and *Description* is a phrase that describes the information about *Noun*.

Next, I describe how attention m_k cam be measured from text. Suppose the report contains M such tuples, $\hat{s}_1, \hat{s}_2, \cdots, \hat{s}_M$, I measure the attention in the report to topic k as the fraction of *Noun* that belongs to topic k,

$$\widehat{m}_k = \frac{\sum_{i=1}^{M} \mathbf{1}\{\hat{s}_i(Noun) \in \text{topic k}\}}{M} = \frac{M_k}{M} \tag{10}$$

Measuring attention to a topic as the fraction of text content devoted to the topic has been the standard practice in NLP (e.g. Bybee et al. (2024)), but here the measurement also can be mapped to the definition of attention in (6). If one interprets each tuple $\hat{s}_i$ as an independent draw of a normally distributed signal with precision $\hat{\tau}$, the total precision the analyst can obtain by observing M tuples is $M\hat{\tau}$. If the analyst devotes M_k of the tuples to topic k, it is equivalent to him observing a signal for topic k with precision $M_k\hat{\tau}$. Therefore, the ratio $\hat{m} = M_k/M$ directly maps to the definition in (6).

Finally, I map the text in $\hat{s}$ in the information shock to component k. To do so, I first gather the embedding vector for $(\hat{s}_i(Noun), \hat{s}_i(Description))$, which I denote using X_i. The information about topic k in the report is simply captured by the average of embeddings of tuples that belong to topic k, $X^k = Avg(X_i|\hat{s}_i(Noun) \in \text{topic k})$. The analyst's reaction to information about topic k, therefore, can be described as a response function $f_k(\cdot)$ that maps the embedding vector X^k to his revision about structural component k. For simplicity, I assume that the response function is linear and takes the form $f_k(X^k) = \lambda_k' X^k$, where λ_k is a coefficient vector that has the same dimension as X^k. The analyst's revision about structural component k, therefore, can be represented as

$$Rev_t[y_{k,t+1}] = \lambda_k X_t^k. \tag{11}$$

Given the revision formation in (8), one can estimate the analyst's revision due to component k by running the regression regression

$$Rev_t[y_{t+1}] = \sum_{k=1}^{K} \lambda_k X_t^k + \epsilon_t \tag{12}$$

and defining the estimated $Rev_t[y_{k,t+1}]$ as

$$\widehat{Rev}_t[y_{k,t+1}] = \widehat{\lambda}_k X_t^k. \tag{13}$$

One of the main benefits of using text in analyst reports to study analysts' belief formation is that the researcher can directly condition on the information set the analyst responds to. In the context of estimating $\widehat{Rev}_t[y_{k,t+1}]$, this benefit translates to an identification condition that guarantees the estimated $\widehat{\lambda}_k$ identifies the structural parameter λ_k. Formally, the identification result is stated in the following proposition:

Proposition 1 (Identification of λ_k)
Suppose ϵ_t in (7) are uncorrelated with all the embedding vectors X_t^k, i.e. $X_t^k \perp \epsilon_t$ for all k, then a linear regression estimate $\widehat{\lambda} = [\widehat{\lambda}_1', \cdots \widehat{\lambda}_K']'$ from regressing $Rev_t[y_{t+1}]$ onto stacked $X_t = [X_t^1, \cdots X_t^K]$ identifies $\{\lambda_k\}$.

This identification result follows directly from the fact that linear regression estimator achieves identification under a conditional independence assumption. Intuitively, the assumption $X_t^k \perp \epsilon_t$ states that the text about topic k in analyst report contains all information about structural component k that the analyst responds to. In other words, text in analyst reports provide sufficient statistics for analyst's belief formation.

This result highlights the benefit of working with text data to directly study belief formation. Without directly conditioning on the analyst's information set, a researcher might be tempted to regress $Rev_t[y_{t+1}]$ on a signal the researcher observes—for example past return—and then call the fitted value the analyst's "revision due to past return." However, this estimate might not truly capture the analyst's revision due to past return: the analyst may be responding only to past earnings growth, which is correlated with past return.[17] By working with text data in analyst

[17] Some recent papers attempt to instrument revisions using plausible exogenous shocks. For example, Angeletos et al. (2021) and Kučinskas and Peters (2024) instruments macroeconomic expectation revisions using macroeconomic or inflation shocks, and Chaudhry (2023) instrument earnings expectation revisions using flow-induced price

reports, the researcher can avoid this kind of omitted variable bias.

3.3 Text-Instrumented Coibion-Gorodnichenko Regression

The uncovering of belief revision due to different information provides an avenue for studying misreaction to information at a micro-level, therefore providing a more detailed diagnosis of an analyst's subjective belief mistakes.

To implement this, I generalize the seminal Coibion-Gorodnichenko regression (Coibion and Gorodnichenko (2015), henceforth CG-regression) to uncover the relationship between forecast error and forecast revision due to specific information. The traditional (individual level) CG-regression takes the form

$$y_{t+1} - F_t[y_{t+1}] = \alpha + \beta \cdot Rev_t[y_{t+1}] + \epsilon_t. \tag{14}$$

Without knowing the agent's information set, CG-regression proposes to use a positive (negative) forecast revision as a proxy for good (bad) information about y_{t+1}. Thus a positive (negative) β estimate can be interpreted as underreaction (over-reaction) to information. Despite its popularity, the CG-regression cannot shed light on a fundamental question about belief formation: What specific information does the agent over- or underreact to?

Given the identification results in Proposition 1, the researcher can identify an analyst's forecast revision due to different information. This allows one to study analyst's misreaction to detailed information. I thus generalize the usual CG-regression setup to multiple revision components uncovered using text, which I call "Text-instrumented CG-regression":

$$y_{t+1} - F_t[y_{t+1}] = \alpha + \sum_{k=1}^{K} \beta_k \cdot Rev_t[y_{k,t+1}] + \epsilon_t. \tag{15}$$

The next proposition states that the text-instrumented CG regression coefficient can be interpreted as over/under-reaction due to information about structural component k, in the context of subjective belief updating (7) and the rational belief updating in Proposition 3:

Proposition 2 (Identifying misreaction to specific information using text-instrumented CG-regression)
Suppose ϵ_t in (7) are uncorrelated with all the embedding vectors X_t^k, i.e. $X_t^k \perp \epsilon_t$ for all k. Consider a two-stage process where the researcher first runs the regression

$$Rev_t[y_{t+1}] = \sum_{k=1}^{K} \lambda_k X_t^k + \epsilon_t \tag{16}$$

changes. For many types of information, however, such exogenous shocks might not be easily found.

and then computes the fitted revision $\widehat{Rev}_t[y_{k,t+1}] = \widehat{\lambda}_k X_t^k$. *Then the text-instrumented CG-regression*

$$y_{t+1} - F_t[y_{t+1}] = \alpha + \sum_{k=1}^{K} \beta_k \cdot \widehat{Rev}_t[y_{k,t+1}] + \epsilon_t \tag{17}$$

finds $\beta > 0$ *if* $A_k < 0$ *(under-reaction to information about k) and* $\beta < 0$ *if* $A_k > 0$ *(over-reaction to information about k), provided that* $Cov(F_t[y_{t+1}], E_t[y_{t+1}]) > 0$.[18]

In summary, because text in analyst reports reveals rich information that the analyst reacts to, the researcher can identify the analyst's reactions to different types of information. By studying their relationship with forecast error, one can thus identify over/under-reaction to different information at the micro level. I implement the text-instrumented CG regression in Section 6.

4 Information Extraction from Analyst Reports

In this section, I describe the information extraction procedure using an LLM. To fully digest and summarize the textual information in analyst reports, one needs a way to comprehend the content in the reports and to extract relevant information. The traditional approach usually involves some degree of manual reading by humans, but given my sample of more than 1.2 million analyst reports, manual reading comprehension and information extraction is prohibitively inefficient. Instead, I leverage the recent development of LLMs and extract information in a RAG framework.

4.1 RAG Implementation

4.1.1 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 `FinBERT`.[19] I then generate another embedding y for the retrieval query "{TICKER} EPS forecast", 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

[18] The condition $Cov(F_t[y_{t+1}], E_t[y_{t+1}]) > 0$ is to ensure the deviation from rationality is not too large, so that the direction of updating is the same for subjective and rational expectation, where the over/under-reaction is in the sense of magnitudes of updating.

[19] 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.

the query embedding:

$$\text{Relevant Context Set} = \{n \text{ for } n \in \text{ Top 20 of } \{\frac{x_n \cdot y}{||x_n||||y||}\}\} \tag{18}$$

By retrieving using similarity in the embeddings, the retrieved context will be more relevant for the EPS 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.

4.1.2 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.[20]

Step 1: Identify Stock Information

In the first step, the LLM is asked to identify the stock in the context. 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. Figure 11 in Appendix E shows the exact prompt used for identifying stock information. I conduct a validation check in Appendix G.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 vs. Subjective Content

In the next step, I use the LLM to separately summarize the factual information and the analyst's subjective statements presented in the context. The reports contain a rich set of information underpinning the analyst's subjective view of a firm. In the textual content of a report, the analyst typically describes the recent developments in a public firm's operations and financials, and then states his own subjective opinion about the firm's future outlook. As an example, consider two statements:

1. "TSLA reported lower EV units delivery due to supply chain disruption in Q1."
2. "We believe TSLA will trade lower as TSLA will face more challenges in EV delivery."

While both sentences talk about TSLA and its EV delivery, the first sentence states an objective fact about TSLA that has already happened and that the analyst has paid attention to, while the

[20] I also use LLM to identify and remove boilerplate content, using prompt in Figure 10 in Appendix E.

second sentence is a subjective statement reflecting the analyst's own opinion about the future. To study the drivers of subjective expectations, it is important to separate the factual content from the subjective content, because the former is connected to the information set that the analyst pays attention to, while the latter maps to the analyst's subjective narrative about the future.

I leverage the LLM's textual comprehension and summarization ability to separate the factual versus subjective content in the analyst reports. Specifically, for each context chunk retrieved by the retriever and passed Step 1 that identifies the correct stock, I feed it into the LLM and ask the LLM to summarize the factual and subjective content separately. Figure 12 in Appendix E describes the prompt used in this task.

Step 3: Extracting Narratives from Factual and Subjective Content

In the next step, I extract narratives from both factual and subjective content. As introduced in Section 3.2, I define narrative as a tuple of $\{Noun\text{: } Description\}$, 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*. This is essentially a dimensionality reduction step: the text of factual information is an ultra high-dimensional objective, and to construct interpretable variables of interest, one needs to further condense the information in the text.[21] Figure 13 shows the actual prompt used.

Table 23 in Appendix F shows an example separating factual versus subjective content and narrative extraction. In the original context, the analyst from Credit Suisse writes detailed factual information about BBWI's loyalty program and its SG&A expense. 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 a subjective view expressed by the analyst. The LLM can successfully identify and summarize this subjective content. The last part shows the financial economic concepts extracted by the LLM, along with their descriptive adjectives. Appendix G reports more examples of LLM responses, which further validate that the LLM gives satisfactory results that separate factual information from subjective statements and extracts meaningful narratives.

4.2 Topic Classification

The LLM inference step extracts 46.11 million occurrences of 7.48 million unique nouns. I focus on the 3.48 million most frequent nouns, which covers over 90% of the occurrences. To further distill the information for better interpretability and map clearly to the idea of structural components

[21] This can be viewed as an NLP task known as Semantic Role Labeling (SRL). See Ash et al. (2024) for an application of SRL analysis of US congress corpus.

in Section 3.2, I classify the nouns using zero-shot classification.[22] Zero-shot classification is an NLP task where a model is trained on a set of labeled examples but is then able to classify new examples from previously unseen classes by using the embeddings of the text and the class label. The main advantage of using zero-shot classification is that the user has the flexibility to input candidate labels which the model is not trained on, without retraining the model. I use the zero-shot classification pipeline of `bart-large-mnli`, the most downloaded zero-shot classifier on `huggingface`.

Specifically, based on manual reading of multiple analyst reports, I design 25 labels that are supposed to capture the majority of the report contents. For each of the 3.48 million nouns, I use the zero-shot classifier to infer the probability of the term belonging to one of the labels. The noun is classified into the label when its predicted probability is greater than 0.2 (vs. an uninformed guess of 0.04). I allow for multiple label classification, so the noun phrases such as "profit margin growth" can be classified into both "margin" and "growth" label.

These 25 labels are then aggregated into 7 topics: Profitability, Financial Conditions, Business Operations, Corporate Management, Growth, Valuation, and Macroeconomics. Table 1 reports the topic assignments of classification labels. I then manually check for false positive and false negatives. Figure 14 shows the word cloud for the seven topics, and Figure 16 in Appendix shows the word cloud for the 25 labels.

Topic	Classification Label
Profitability	"earnings", "revenue", "expense", "margin", "net income"
Financial Conditions	"cash", "debt", "assets", "balance sheet"
Business Operations	"customer", "product", "inventory", "industry competition"
Corporate Management	"management", "merger and acquisition", "research and development", "tax", "regulation"
Growth	"growth"
Valuation	"valuation", "stock price"
Macroeconomics	"macroeconomic condition", "international", "demand", "supply"

Table 1: Topic and zero-shot classification label assignment

[22] I explored two alternative unsupervised topic modeling approaches: LDA and embedding-based clustering. Both methods yield uninterpretable topics. The LDA model estimates topics based on word-count distribution over documents, but unlike other text corpus like news articles where there is a natural thematic structure, the semantic structure in analyst reports is less clear. For unsupervised clustering based on embedding vectors, I find the embedding vectors based on general-purpose LLMs are usually very similar to the financial terms in analyst reports, and the clustering algorithms cannot discover any meaningful clustering structure when unsupervised.

5 An Anatomy of Analysts' Attention

In this section, I study how analyst's attention to different information varies in the time series and in the cross-section. As introduced in Section 3.2, I measure analyst i's attention to topic k when forecasting firm j's EPS at time t as the fraction of topic k-related nouns in the factual content in the analyst report:

$$Attn_{i,j,t}^{k} = \frac{\sum_{noun \in A_{i,j,t}^{factual}} \mathbf{I}(noun \in C^{k})}{\sum_{noun \in A_{i,j,t}^{factual}} \mathbf{I}}, \tag{19}$$

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 noun $noun$ belongs to the topic of k.

Figure 1 shows the full sample average attention to the seven topics in factual content. Analysts' attention gravitates mostly toward firm profitability, which takes more than 40% of their overall attention allocation. Analysts' attention to the other six topics are roughly equal, each taking about 10% of the attention allocation.

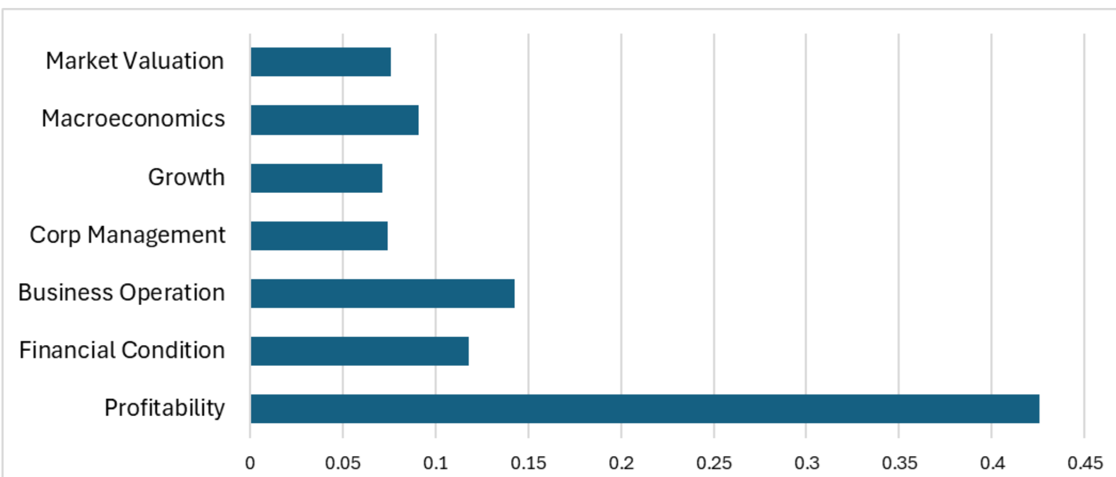


Figure 1: Full sample average of analyst attention.

Because an analyst's valuation is typically determined using earnings forecasts as an input, I exclude the "Market Valuation" topic from subsequent analysis. This is to ensure that I focus on topics that are potential informational drivers for earnings belief formation.

5.1 Stylized Facts about Attention: Time Series Variation

Analysts' attention exhibits significant variation over time and over business cycles. I construct an aggregate measure of attention to topic k information as simply the average $Attn_{i,j,t}^{k}$ measure across all analysts and firms at time t.

Figure 2 plots analysts' attention to different topics of factual information over time. One prominent pattern in analysts' attention is how it changes during recessions. In a recession, analysts pay more attention to a firm's financial condition and macroeconomic topics, while reducing their attention to profitability, corporate management, and growth.

As a formal statistical test, I run a time-series regression of aggregate attention to each topic on log industrial production, the VIX index, and a trend indicator:

$$\overline{Attn}_t^k = \alpha Trend_t + \beta_1 \log IP_t + \beta_2 VIX_t + \epsilon_t. \tag{20}$$

The coefficient on industrial production measures how analysts' attention changes with the aggregate economy, and the coefficient on VIX measures how analysts' attention moves with financial uncertainty. Together, the regression serves to diagnose how attention varies with the business cycle.

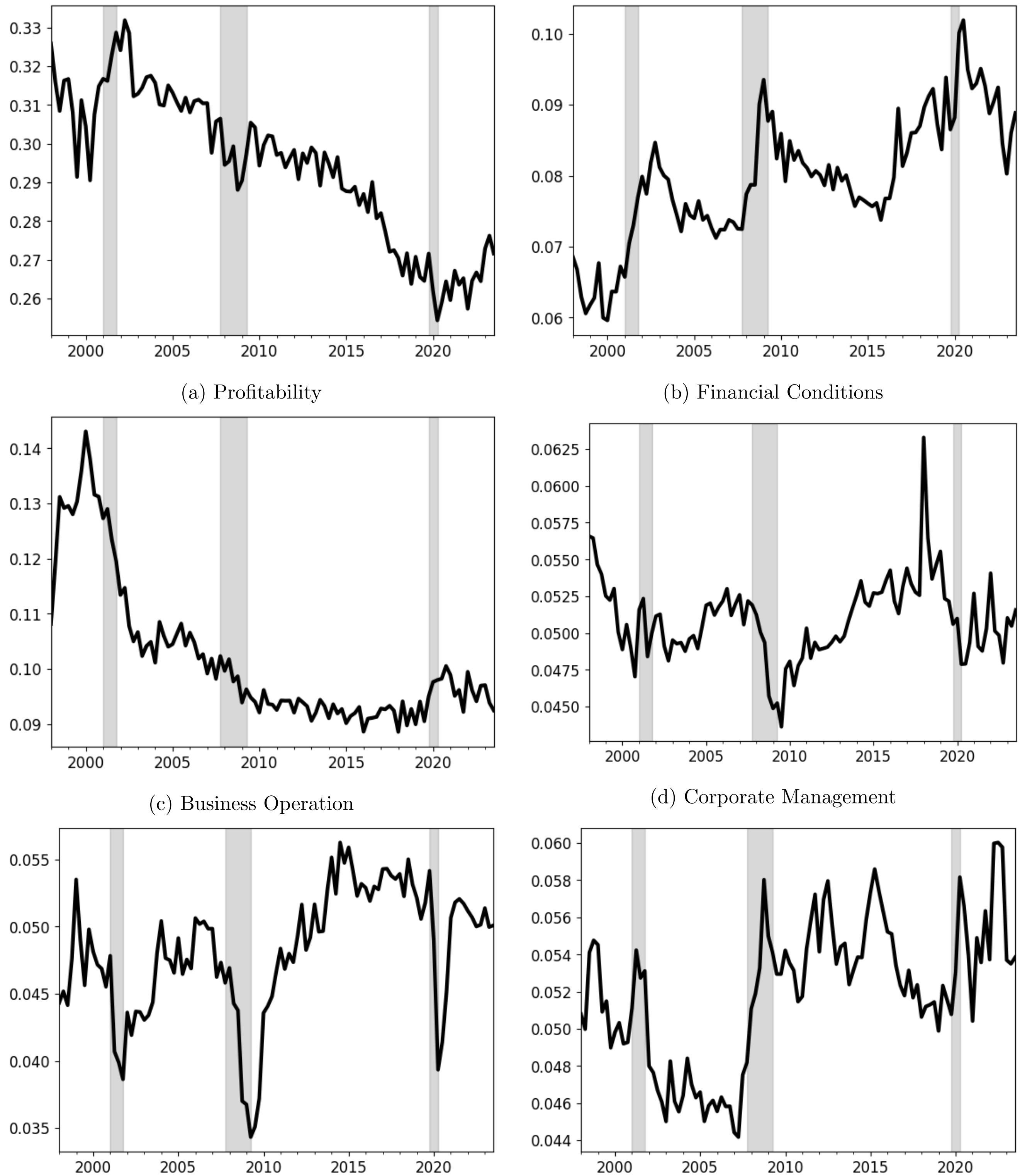
The regression results are presented in Table 2. I find analysts' attention to Profitability, Business Operations, Corporate Management, and Growth are pro-cyclical, as they move positively with industrial production and negatively with VIX. On the other hand, their attention to Financial Conditions and Macroeconomics is counter-cyclical.[23] These patterns are consistent with the idea that in economic downturns, firm's future fundamentals becomes more exposed to financial distress and aggregate risk, and analysts are correctly paying more attention to information about these variables in a recession. Notably, analysts' counter-cyclical attention to macroeconomic information is also consistent with recent findings that firm managers also have counter-cyclical attention to macroeconomic states (e.g. Song and Stern (2021); Flynn and Sastry (2024)). Finally, attention to Financial Conditions and Macroeconomics also increases over time, suggesting an increasing importance of financial distress and aggregate economic conditions in determining firms' fundamentals.

5.2 Stylized Facts about Attention: Cross-Sectional Variation

In this section, I explore how analysts' attention varies across stocks in the cross-section.[24] I focus on three prominent stock attributes: a stock's value (measured by book-to-market ratio), profitability (measured by operating profit-to-book ratio), and leverage (measured by total debt to market cap). I run a regression of attention on the three stock characteristics, together with

[23] One might be concerned that the analyst or firm composition in the IBES forecast sample might change over time, and that the time-series variation in attention might be driven by these changes. In Table 24 in Appendix H.1, I run a panel regression controlling for analyst and firm fixed effects, and I find similar results as in Table 2.

[24] In Appendix H.2, I explore how analysts' attention varies across industries.



(a) Profitability

(b) Financial Conditions

(c) Business Operation

(d) Corporate Management

(e) Growth

(f) Macroeconomics

Figure 2: Analysts' quarterly average attention to different topics. Shaded region denotes recession periods identified by NBER.

	Profitability	Financial conditions	Business operation	Corporate management	Growth	Macroeconomics
Log IP	0.432***	-0.505***	0.212**	0.260***	0.223***	-0.493***
	(10.20)	(-8.46)	(2.18)	(3.62)	(3.51)	(-5.29)
VIX	-0.021***	0.048***	0.026	-0.064***	-0.081***	0.059***
	(-3.02)	(4.40)	(1.00)	(-5.27)	(-7.07)	(3.89)
Trend	-3.242***	2.943***	-2.579***	-0.027	1.144***	2.202***
	(-10.57)	(7.56)	(-4.78)	(-0.04)	(3.20)	(5.13)
R^2	0.83	0.69	0.59	0.17	0.53	0.44
Nobs	103	103	103	103	103	103

Table 2: Regression results of topic attention on macroeconomic variables. The attention measure is normalized to have zero mean and unit standard deviation. *Trend* is a variable that equals 0 at the beginning of the sample and 1 at the end. For time-series regression, standard errors are HAC-adjusted with 4 lags. *, **, *** denote significance level at 10%, 5% and 1%.

time and stock fixed effects:

$$\overline{Attn}^k_{j,t} = \beta_1 BE/ME_{j,t} + \beta_2 OP/BE_{j,t} + \beta_3 Debt/ME_{j,t} + \eta_t + \xi_j + \epsilon_{j,t}. \tag{21}$$

Table 3 reports the results. It shows that analysts' attention to different topics differs across different types of stocks. Specifically, for value firms, analysts pay more attention to profitability, financial conditions, and corporate management, while paying less attention to business operations and growth. This is consistent with the idea that value firms operate with stable business models with less growth potentials and therefore current profitability and financial conditions are more reflective of the firm's fundamentals. For firms with high profitability, analysts focus more on growth, corporate management, and macroeconomics. This indicates that analysts are gauging whether firm profitability has more growth potential and can weather broader market conditions. Finally, for firms with high leverage, analysts focus more on profitability and financial conditions, consistent with financial distress being the major concern for highly leveraged firms.

Overall, these cross-sectional attention results provide suggestive evidence that analysts pay attention to the defining feature of the stock that are consistent, and these focused features do appear to be more relevant for a firm's fundamental.

5.3 Stylized Facts about Attention: Short vs. Long-term forecasts

The literature has documented that analyst forecasts have different behavioral patterns when they forecast short-term versus long-term targets (e.g. Van Binsbergen et al. (2023); De Silva and Thesmar (2024)). A natural question to ask is: Do analysts pay attention to different information

	Profitability	Financial conditions	Business operation	Corporate management	Growth	Macroeconomics
BE/ME	0.186***	0.088***	-0.094***	0.067***	-0.463***	-0.015
	(7.44)	(3.76)	(-3.97)	(2.61)	(-15.82)	(-0.71)
OP/BE	-0.038*	-0.031	-0.014	0.062**	0.079***	0.052**
	(-1.65)	(-1.39)	(-0.62)	(2.30)	(2.80)	(2.08)
Debt/ME	0.131***	0.266***	-0.067**	-0.189***	-0.286***	-0.033
	(3.78)	(8.04)	(-2.28)	(-6.00)	(-8.10)	(-1.01)
Time FE	✓	✓	✓	✓	✓	✓
Stock FE	✓	✓	✓	✓	✓	✓
Total R^2	0.003	0.003	0.001	0.001	0.014	0.000
Nobs	115159	115159	115159	115159	115159	115159

Table 3: Regression results of regressing quarterly topic attention on selected stock characteristics 21. Attention is measure is normalized to have zero mean and unit standard deviation. Each stock characteristics is cross-sectionally rank-standardized and scaled to be in [0, 1] interval. Standard errors are two-way clustered at firm and time level. *, **, *** denotes significance level at 10%, 5% and 1%.

when they make short-term and long-term forecasts?

In this section, I explore how analysts' attention choice depends on the forecast horizon. Specifically, I compare analyst's attention choice to different components of firms' fundamentals in three different scenarios: when the analyst only announces a short-term (FY1 or FY2) forecast (denoted as "Only ST"); when the analyst only announces a long-term (FY3 to FY5) forecast (denoted as "Only LT"); and when the analyst announces both a short-term and long-term forecast (denoted as "Both ST and LT").

Table 4 reports the attention in "Only ST", "Both ST and LT", and "Only LT" announcements, as well as the relative difference (in percentages) with respect to short-term forecast announcements. I find that when analysts make long-term forecasts, they pay less attention to a firm's Profitability, and pay more attention to all other topics. The increase in attention (in relative percentages) is largest for Financial Conditions, suggesting that analysts might perceive financial conditions as a particularly important driver of long-term earnings.

One might be concerned that the differences in attention in short-term and long-term forecasts are driven by different compositions of analysts, firms, or time periods. In Appendix H.3 I conduct a formal test while controlling for fixed effects, and I find that the significant differences in attention generally continue to hold, although the economic magnitudes are smaller due to the fixed effect controls.

Overall, these results suggest that analysts do pay attention to different information when they forecast earnings at different horizons. When they make long-term forecasts, they pay less attention to firms' profitability information, while paying more attention to all other information.

	Avg. Attention			% Diff from Only ST		
	Only ST	Both ST & LT	Only LT	Only ST	Both ST & LT	Only LT
Profitability	0.302	0.272	0.250	(-)	-9.8%	-17.1%
					(-68.2)	(-39.4)
Financial conditions	0.067	0.079	0.082	(-)	17.9%	22.3%
					(6.8)	(19.7)
Business operations	0.098	0.096	0.107	(-)	-1.8%	9.0%
					(-7.0)	(10.6)
Corp Management	0.050	0.052	0.055	(-)	3.1%	8.5%
					(10.8)	(9.1)
Growth	0.048	0.051	0.053	(-)	4.9%	8.5%
					(14.6)	(8.0)
Macroeconomics	0.051	0.054	0.053	(-)	5.8%	2.1%
					(17.6)	(2.1)
Num. Anns.	265785	112927	10469			

Table 4: Average topic attention in announcements with different forecast horizons. The numbers in parentheses report the t-stat of differences in means.

Cross-analyst variation in attention In Appendix H.4, I investigate the cross-analyst variation in attention to firm-level information. The optimal attention choice in Section D.1 predicts that all analysts should have the same attention choice when they make a forecast for the same firm at the same time, because they are facing the same information environment. However, I show that there is significant heterogeneity in the attention to different topics across analysts even when controlling for a stock-time fixed effect. This heterogeneity does not seem to be explained by the usual analyst characteristics, such as analyst experience and stock or industry coverage. Understanding the drivers of analyst-specific attention choice is an interesting direction for further exploration.

5.4 Is Analyst Attention Allocation Rational?

In this section, I provide evidence that the variation in analysts' attention documented in the previous sections is consistent with the prediction of the rational inattention in Section 3.1.1, which says that if analysts want to maximize forecast accuracy but face cost of signal precision, they should focus their attention to the components that have more predictive power of future earnings.

I directly test this relationship between earnings predictability and attention in the data. I focus on the attention to Profitability and Financial Conditions, the two topics whose current realizations are easier to capture using publicly available accounting data. Specifically, I run a predictive regression

$$\Delta y_{j,t+h} = \beta_1 Z^k_{j,t} + \beta_2 Z^k_{j,t} \times \mathbf{1}\{\text{High Attn}\}_{j,t} + \epsilon_{j,t+h} \tag{22}$$

where $\Delta y_{j,t+h}$ is the h-year change in a firm's earnings, $Z^k_{j,t}$ is a stock characteristics that captures the current realization of component $k \in \{\text{Profitability}, \text{Financial Condition}\}$, and $\mathbf{1}\{\text{High Attn}\}$ is an indicator for a firm-quarter observation where analysts have high (above median) attention.[25] The magnitude and significance of β_2 thus provide a gauge of the difference in the predictive power of component k for future earnings growth in the high versus low attention subsample. Table 28 in Appendix H.5 lists the stock characteristics I use for capturing realized profitability and financial conditions.[26]

	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: Profitability						
$\overline{Z}^{profit}_{j,t}$	0.042***	0.062***	0.043***	0.026***	0.025***	0.026***
	(3.96)	(7.97)	(4.12)	(8.53)	(13.39)	(8.68)
$\overline{Z}^{profit}_{j,t} \times \mathbf{1}\{\text{High Attn}\}$			0.018**			-0.001
			(1.99)			(-0.19)
Panel B: Financial Condition						
$\overline{Z}^{fin.\ cond.}_{j,t}$	0.004**	0.009***	0.004***	0.001	0.004***	0.001
	(2.31)	(2.35)	(2.38)	(1.43)	(3.27)	(1.49)
$\overline{Z}^{fin.\ cond.}_{j,t} \times \mathbf{1}\{\text{High Attn}\}$			0.005			0.004***
			(1.07)			(2.48)
Quarter FE	✓	✓	✓	✓	✓	✓
Firm FE	✓	✓	✓	✓	✓	✓
Nobs	53658	53658	53658	53658	53658	53658

Table 5: Regression results of (22). Standard errors are double clustered at firm and quarter level. *** denotes significance at 1% level.

Table 5 shows the earnings predictability results. In Panel (A), I examine the predictability of profitability characteristics in the sample of high analysts' attention to Profitability (firm-quarter observations with above median $\overline{Attn}$) versus in the sample of low attention. I use the average of profitability characteristics as the predictor. The first three columns show the results for 1-year ahead earnings prediction results, where I find that current profitability characteristics do predict future 1-year earnings growth more strongly when attention is high, which is consistent with rational attention allocation. The difference of predictability becomes insignificant for 5-year

[25] I scale earnings by current price, following Bouchaud et al. (2019).

[26] All stock characteristics are obtained from Jensen et al. (2023) and are cross-sectionally rank-standardized to be within [-0.5, 0.5] interval, following standard practice.

earnings growth, indicating that changes in attention to Profitability are mainly driven by changes in predictability in short-term earnings forecast, which is also consistent with the fact that analysts pay more attention to Profitability when they make short-term forecasts.

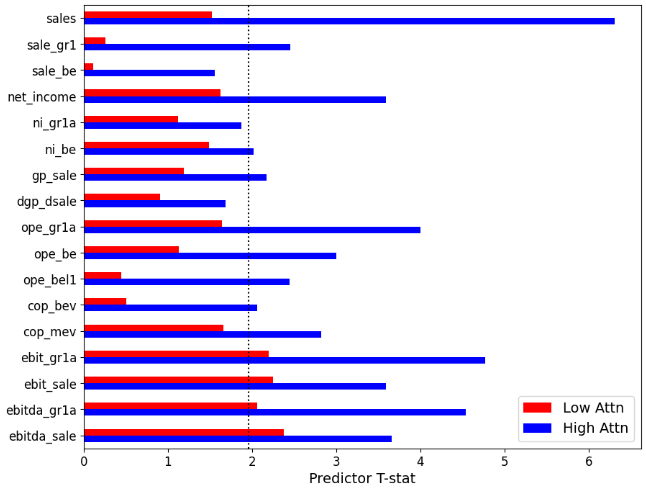
Pane (B) in Table 5 reports the results for financial condition characteristics. I again find that these characteristics predict future earnings more strongly in the subsample where analysts pay more attention to the Financial Condition topic. This difference in predictability is statistically significant for 5-year future earnings growth prediction, which—in conjunction with the evidence that analysts pay more attention to Financial Conditions when making long-term forecasts—is again consistent with rational attention allocation.

As a robustness check, I also conduct earnings prediction regressions for individual stock characteristics. Figure 3 reports the regression t-statistics for the high attention subsample (blue bars) and low attention subsample (red bars), respectively. I find that the differences in predictability are observed for many stock characteristics. In Panel (a), I find that for many profitability characteristics, their 1-year ahead earnings growth predictive power is statistically insignificant in the low-attention subsample, but almost all profitability characteristics have significant predictability for the high-attention subsample. In Panel (c), I find that for 5-year ahead earnings prediction, the predictability for many characteristics is still more significant in the high attention subsample.

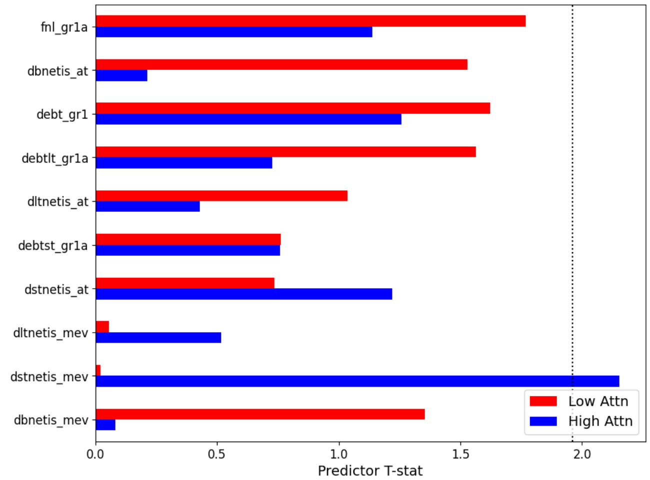
For financial condition characteristics, while most of them have no predictive power for future 1-year earnings growth, as shown in Panel (b), almost all of them have significant predictive power for 5-year ahead earnings growth in the high attention subsample, as shown in Panel (d). On the other hand, in the low attention subsample, their predictive power remains insignificant.

In summary, although determining the rationality of attention is difficult empirically—because it is hard to measure the "rational attention" benchmark in the data—by examining how the predictive power of different observable stock characteristics for earnings change with analysts attention, I provide evidence that analyst's attention allocation is consistent with rational inattention choice: analysts increase their attention to a fundamental component when that component is more important in driving future earnings.

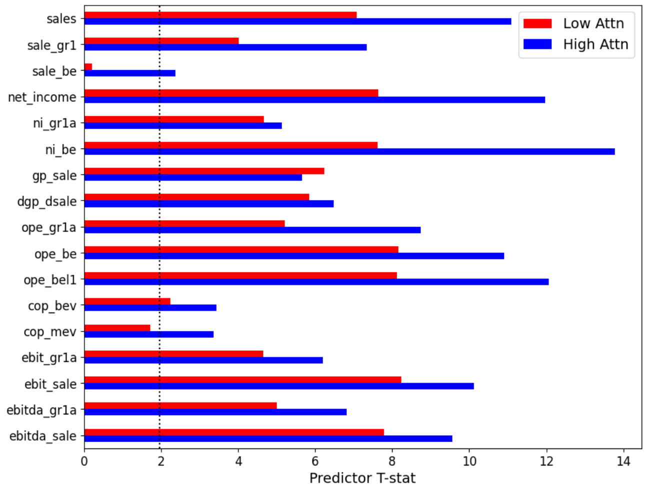
(a) Predicting 1-year future EPS growth: Profitability characteristics



(b) Predicting 1-year future EPS growth: Financial condition characteristics



(c) Predicting 5-year future EPS growth: Profitability characteristics



(d) Predicting 5-year future EPS growth: Financial condition characteristics

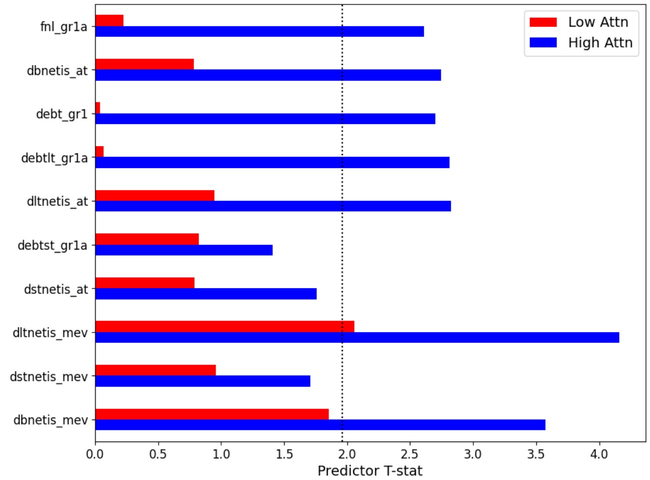


Figure 3: T-stat of earnings prediction regression of individual stock characteristics in high and low attention sample

6 Misreaction to Information

In this section, I move on to study how analysts misreact to different types of information using the text-instrumented CG regression introduced in Section 3.3.

6.1 Evidence from Original CG-regressions

As a starting point, I confirm that analysts underreact in their short-term forecasts but overreact in their long-term forecasts—consistent with prior literature (e.g. Bouchaud et al. (2019), Bordalo et al. (2024a))—by running the original CG-regression,

$$y_{j,t+h} - E_{i,t}[y_{j,t+h}] = \beta_{CG} \cdot Rev_{i,t}[y_{j,t+h}] + \epsilon_{i,j,t}, \tag{23}$$

where $y_{j,t+h}$ is firm j's h-year ahead EPS at time t and $E_{i,t}[y_{j,t+h}]$ is analyst i's forecast of $y_{j,t+h}$.[27] I control for analyst, firm, and quarter fixed effects in this regression. As in the original CG-regression framework, a positive β_{CG} indicates that, when analysts receive good news and revise their forecasts upwards, their new forecasts are still too low relative to the actual EPS, therefore indicating an underreaction. Similarly, a negative β_{CG} indicates an overreaction to information.

	Forecast Horizon				
	1 Year	2 Years	3 Years	4 Years	5 Years
β_{CG}	0.077***	0.089***	-0.160***	-0.347***	-0.413***
	(12.59)	(7.66)	(-11.10)	(-10.83)	(-8.57)
Quarter FE	✓	✓	✓	✓	✓
Analyst FE	✓	✓	✓	✓	✓
Firm FE	✓	✓	✓	✓	✓
Adj. R^2	0.003	0.001	0.003	0.016	0.022
Num Obs.	464164	420071	136176	26919	14240

Table 6: Original CG-regression results on individual earnings forecasts. Standard errors are double clustered at analyst-firm and firm-quarter level. *** denotes significance at 1% confidence level.

Table 6 reports the results. I confirm that the previously documented misreaction to information in earnings forecasts holds in my sample.[28] As the text in the analyst reports provides detailed information that analysts incorporate in their forecasts, I now study which information analysts misreact to, using text-instrumented CG regression.

[27] Following Bouchaud et al. (2019), I scale earnings and forecasts by the price at the time of forecast. Outliers outside 1% and 99% percentiles are dropped.

[28] Table 29 in Appendix I reports the results for the full IBES sample, where the underreaction in short-term forecasts and overreaction in long-term forecasts holds as well.

6.2 Evidence from the Text-Instrumented CG Regression

To implement the text-instrumented CG regression, one first needs to capture the information about different topics in text. I do so by using the embedding vector of both *Noun* and *Description* in the extracted narrative tuples. Specifically, consider a narrative tuple s=("net margin", "increased by 6%"). I represent the textual information for this tuple as

$$X_s = [\texttt{Embedding}(\text{"net margin"}),\ \texttt{Embedding}(\text{"increased by 6\%"})] \tag{24}$$

where $\texttt{Embedding}(\cdot)$ denotes the embedding output. I use `FinBERT` as my embedder, which gives a 768-dimensional embedding vector for each string input. To reduce dimensionality, I obtain the first 100 principal components of the embedding vector, which captures more than 77% of the variation in all the noun phrases and descriptions extracted. Therefore, each narrative tuple is represented by a 200-dimensional vector (100 dimensions for the embedding of *Nouns* and *Description*, respectively).[29]

After obtaining the embedding vectors for each narrative tuple, the information about topic k is captured as the average embedding vectors for narratives whose *Noun* belongs to topic k. I then implement the text-instrumented CG-regression introduced in Section 3.3 by first running

$$Rev_{i,t}[y_{j,t+h}] = \sum_k \lambda_k X^k_{i,j,t} + \epsilon_{i,j,t} \tag{25}$$

where $X^k_{i,j,t}$ is the embedding vector for topic k written by analyst i for firm j at time t. I then regress forecast error on fitted revision in each component

$$y_{j,t+h} - F_{i,t}[y_{j,t+h}] = \sum_k \beta_k \cdot \widehat{Rev}_{i,t}[y^k_{j,t+h}] + \epsilon_{i,j,t} \tag{26}$$

where $\widehat{Rev}_{i,t}[y^k_{j,t+h}]$ is the fitted value of revision using embeddings for topic k, i.e. $\widehat{Rev}_{i,t}[y^k_{j,t+h}] = \widehat{\lambda}_k X^k_{i,j,t}$. Since I control for analyst, quarter, and firm fixed effects in the original CG regression specification, to make the text-instrumented CG regression coefficients comparable, I estimate the first-stage regression for the demeaned revision at the analyst, quarter, and firm level. I also scale the demeaned fitted revision $\widehat{\lambda}_k X^k_{i,j,t}$ to have the same sample variance as the demeaned revision in original CG regression. To account for the estimation error in the first-stage, I bootstrap standard errors by repeatedly sampling (with replacement) at the analyst-firm and firm-quarter level.

Table 7 reports the text-instrumented CG regression results. Two distinct patterns stand out.

[29]The choice of embedding dimensionalities reflects the classic bias-variance tradeoff: if the dimensionality is too small, the embedding vector will not capture enough variation in the meaning of the text, resulting in biases in the fitted revision due to different topics. On the other hand, using a large embedding vector could lead to an uncertainty of the estimation of λ_k when the number of observations is small. Table **??** in Appendix I reports results for using the first 50 and 80 principal components, which shows similar pattern as Table 7.

First, the underreaction in short-term earnings forecasts is pervasive: there are significant positive coefficients on most of the six topics for 1-year earnings forecasts (except Corporate Management). The underreaction in analyst forecasts is consistent with the idea of belief stickiness (e.g. Bouchaud et al. (2019)), and my finding reveals that there is significant belief stickiness in all types of information in an analyst's short-term earnings forecast, especially for Profitability information.

Second, I find that analysts' overreaction in long-term forecasts is mainly concentrated in a subset of information, namely Business Operations, Corporate Management, and Macroeconomics. These results highlight the heterogeneous overreaction to different types of information in analysts' earnings forecasts. Surprisingly, even though analysts pay the most attention to Profitability information, they do not appear to overreact to it in their long-term forecasts; by contrast, the overreaction is mainly significant for the information they pay less attention to. The magnitude for overreaction is particularly large for macroeconomic information.

	Forecast Horizon				
	1 Year	2 Years	3 Years	4 Years	5 Years
Profitability	0.028***	0.038***	-0.010	-0.017	-0.053
	(10.80)	(5.88)	(-0.89)	(-0.57)	(-1.22)
Financial Condition	0.009***	0.014**	-0.021*	-0.026	-0.058
	(3.29)	(2.05)	(-1.68)	(-0.94)	(-1.42)
Business Operation	0.012***	0.007	0.002	-0.052*	-0.112***
	(4.09)	(1.04)	(0.17)	(-1.82)	(-2.62)
Corporate Management	0.003	-0.009	-0.013	-0.070***	-0.085**
	(1.30)	(-1.30)	(-1.07)	(-2.74)	(-1.98)
Growth	0.006**	0.012*	-0.019	-0.026	-0.054
	(2.21)	(1.87)	(-1.53)	(-0.99)	(-1.30)
Macroeconomics	0.007***	0.017***	-0.002	-0.032	-0.125***
	(2.37)	(2.44)	(-0.16)	(-1.08)	(-3.02)
Quarter FE	✓	✓	✓	✓	✓
Analyst FE	✓	✓	✓	✓	✓
Firm FE	✓	✓	✓	✓	✓
Num Obs.	464164	420071	136176	26919	14240

Table 7: Results of text-instrumented CG regression. Analyst, quarter and firm fixed effects are removed in the first-stage regression. Standard errors are estimated based on block bootstrap. *, **, *** denotes significance level at 10%, 5% and 1%.

Taken together, these results from text-instrumented CG regressions provide a more detailed view of how analysts misreact to different types of information. The evidence also provides guidance for models featuring earnings overreaction on what types of information to focus on. In particular, the significant overreaction to Business Operations and Corporate Management seems to be consistent with the findings of Daniel and Titman (2006) and Jiang (2010), who argue that investors mainly overreact to intangible information rather than past tangible, i.e. accounting, information.

Attention-interacted CG regression In the similar spirit of testing misreaction to specific information, I also conduct an attention-interacted CG regression, where I regress forecast error on forecast revision, interacting with analysts' attention to different topics

$$y_{j,t+h} - F_{i,t}[y_{j,t+h}] = \sum_k \beta_k \times Attn^k_{i,j,t} \times Rev_{i,t}[y_{j,t+h}] + \epsilon_{i,j,t}. \tag{27}$$

This specification essentially parameterizes the original CG-coefficient as a function of analysts' attention, and a positive (negative) β_k can be interpreted as analysts underreact (overreact) to information when they pay more attention to topic k.

Table 31 in Appendix I reports the result. I find that the CG-regression coefficient is more positive for Profitability in short-term forecasts, indicating that analysts underreact more in their short-term forecasts when they pay more attention to Profitability. On the other hand, CG-regression coefficient is more negative when analysts pay more attention to Financial Conditions, Business Operations, and Corporate Management, broadly consistent with the text-instrumented CG regression results in Table 7. Note that this regression cannot be interpreted as misreaction to information about topic k, unless we make the assumption that analysts' revision due to topic k information is proportional to their attentions.

6.3 Why Do Analysts Overreact?

Why do analysts overreact more to intangible information compared to tangible information? Although it is difficult to provide a definite answer, in this section I provide evidence that this could potentially be explained by analysts over-reacting more to qualitative content, while over-reacting less to quantitative information. Recent work by Graeber et al. (2024) shows that there is a significant "story-statistics gap" in associative memory: agents' beliefs are more persistently influenced by information presented in a qualitative form (a story) than in a quantitative forms (statistics). As associative memory can induce larger overreaction (e.g. Bordalo et al. (2023) and Enke et al. (2024)), the "story-statistics gap" can potentially explain the heterogeneity in overreaction documented in text-instrumented CG regression.

I provide two pieces of evidence for this channel. First, I show that the information analysts overreact to the most—Business Operations, Corporate Management, and Macroeconomics—is less likely to be presented in the form of numerical statistics. To do so, I compute the fraction of *Description* in the $\{Noun\text{: } Description\}$ narrative tuples that contain numerical values.[30] A narrative tuple with numerical *Description* in the factual content can thus be viewed as describing a piece of information presented in statistical form.

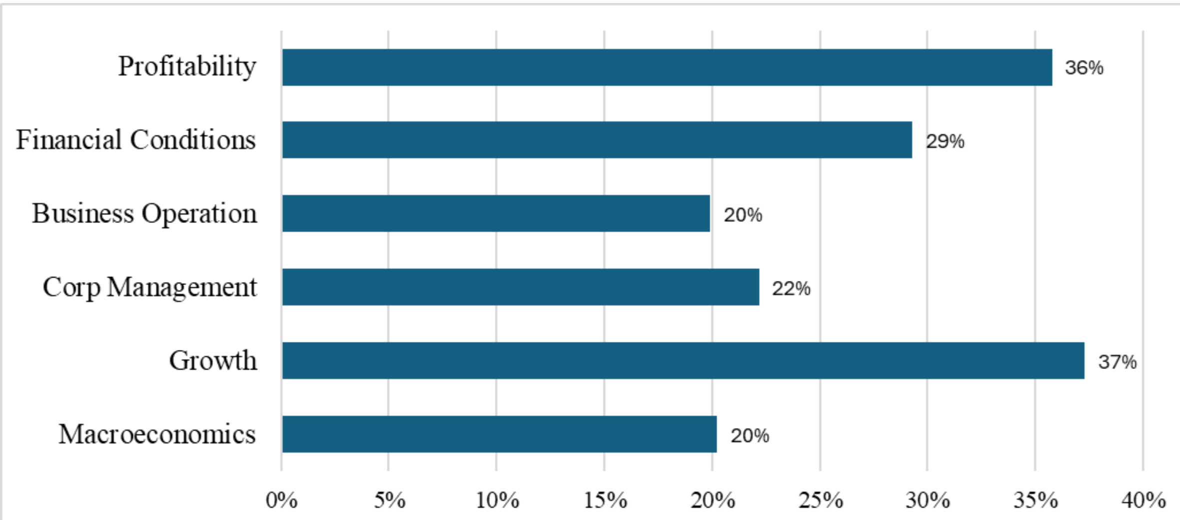


Figure 4: Fraction of *Description*s that contain numeric values

Figure 4 reports the fraction of *Description*s that contain numerical values for each topic. I find that only about 20% of the factual narratives about Business Operations, Corporate Management, and Macroeconomics contain numerical values, indicating that the majority of these types of information are presented in the form of "stories", or qualitative content. On the other hand, when discussing factual information about Profitability, Financial Conditions, and Growth, analysts tend to use more numerical values. The differences in the fraction of statistical descriptions align with the overreaction patterns in Table 7, suggesting analysts might be more prone to overreact to qualitative information.

In the second, more direct test of the story-statistics mechanism, I estimate the original CG-regression coefficient, conditional on the fraction of statistical information in the analyst report. Specifically, I run

$$\begin{aligned} y_{j,t+h} - E_{i,t}[y_{j,t+h}] = \beta_1(E_{i,t}[y_{j,t+h}] - E_{i,t-1}[y_{j,t+h}]) \\ + \beta_2(E_{i,t}[y_{j,t+h}] - E_{i,t-1}[y_{j,t+h}]) \times Qual\_Frac_{i,j,t} + \epsilon_{i,j,t} \end{aligned} \tag{28}$$

where $Qual\_Frac_{i,j,t}$ is the fraction of factual narrative tuples that contains no numerical *Descriptions*. The coefficient β_2 thus estimates the effect of misreaction when analysts are more exposed to information presented in qualitative form.

[30] Specifically, I use Python `regex` to check if a string contains numerical values.

Table 8 reports the results. The negative β_2 estimates imply that, when analysts are more exposed to information presented in qualitative form, their CG-coefficient becomes more negative, indicating more overreaction. Such effect is more significant for long-term forecasts with forecast horizons greater than 3 years.

Overall, the evidence in this section suggests that analysts' different reactions to quantitative versus qualitative information may be a driver of overreaction in long-term earnings forecasts.

	Forecast Horizon				
	1 Year	2 Years	3 Years	4 Years	5 Years
β_1	0.073***	-0.035	-0.059	0.151	0.066
	(2.89)	(-0.80)	(-0.89)	(0.87)	(0.31)
β_2	-0.002	0.086	-0.161*	-0.525**	-0.643**
	(-0.07)	(1.44)	(-1.72)	(-2.03)	(-2.11)
Quarter FE	✓	✓	✓	✓	✓
Analyst FE	✓	✓	✓	✓	✓
Firm FE	✓	✓	✓	✓	✓
Adj. R^2	0.003	0.000	0.003	0.008	0.019

Table 8: Original CG-regression results on individual earnings forecasts, conditional on fraction of quantitative information. Standard errors are double clustered at analyst-firm and firm-quarter level. **, *** denotes significance at 5% and 1% confidence level.

7 Drivers of Disagreement

The previous sections have documented new results about analysts' attention and misreaction to different information, which is based on the study of factual content in analyst reports. In this section, I use both factual and subjective content to study the drivers of disagreement in analysts' earnings forecasts.

Conceptually, disagreement can arise if analysts focus on different information ("asymmetric information channel") or if they interpret the same information differently, leading to varying opinions and outlooks ("differences of opinion channel"). Which channel is responsible for the observed disagreement in the data? The separation of factual and subjective content in analyst reports provides an ideal setting to shed light on this question.

I first measure pairwise disagreement as the absolute difference between the $h-$year ahead EPS forecast between analyst i_1 and i_2 for the same firm j when they both announce their forecast on the same date t, scaled by the stock's (1-month-lagged) price,

$$Disagree^h_{i_1,i_2|j,t} = |E_{i_1,t}[\pi_{j,t+h}] - E_{i_2,t}[\pi_{j,t+h}]|/P_{j,t}. \tag{29}$$

I then compute the semantic difference in the factual and subjective content in analyst i_1 and i_2's written reports. Specifically, I obtain the `FinBERT` embedding of the factual and subjective content of analyst i_1 and i_2's reports, and calculate the cosine distance between them:

$$\begin{aligned} \text{Factual\_Diff}_{i_1,i_2|j,t} &= 1 - \cos(x^{factual}_{i_1,j,t}, x^{factual}_{i_2,j,t}) \\ \text{Subjective\_Diff}_{i_1,i_2|j,t} &= 1 - \cos(x^{subjective}_{i_1,j,t}, x^{subjective}_{i_2,j,t}). \end{aligned} \tag{30}$$

In the formula, $x^{factual}_{i,j,t}$ is the embedding vector of the factual content about firm j written by analyst i at time t, and $\text{Factual\_Diff}_{i_1,i_2|j,t}$ measures how (dis-)similar the factual information in analyst i_1 and i_2's reports are. A high Factual_Diff measure means that i_1 and i_2 possess more distinct information when making the forecast. Similarly, $\text{Subjective\_Diff}_{i_1,i_2|j,t}$ measures how (dis-)similar i_1 and i_2's subjective outlook are; a high Subjective_Diff implies i_1 and i_2 possess more distinct subjective outlooks in their written reports.[31]

With the measures of pairwise disagreement and pairwise semantic difference in factual and subjective content, I investigate the drivers of disagreement. Specifically, I regress the measure of disagreement on the measure of semantic differences:

$$Disagree^h_{i_1,i_2|j,t} = \beta_1 \text{Factual\_Diff}_{i_1,i_2|j,t} + \beta_2 \text{Subjective\_Diff}_{i_1,i_2|j,t} + \tau_j + \tau_t + \epsilon_{i_{12},j,t}. \tag{31}$$

If disagreement is mainly driven by asymmetric information, i.e. analysts observing different information about a firm (due to different attention), then we would expect a positive β_1, meaning that analysts who write about more different factual information should have higher disagreement. On the other hand, if disagreement is mainly driven by differences of opinions, i.e. analysts having a different interpretation for the same factual information, we would expect a positive β_2.

Table 9 reports the result for regression (31). The positive and significant coefficients on Factual_Diff and Subjective_Diff for 4-Years and 5-Years ahead forecasts implies that the semantic differences in both the factual and subjective content comove with pairwise disagreement. This suggests that both asymmetric information and differences of opinions are important drivers for analyst disagreement, particular for the long-term forecasts.

[31] I merge factual and subjective content across all context chunks.

	1 Year	2 Years	3 Years	4 Years	5 Years
Factual_Diff	0.000	0.001	0.001	0.020*	0.033**
	(0.52)	(0.8)	(0.60)	(1.73)	(1.97)
Subjective_Diff	0.000	-0.001	-0.001	0.017*	0.036**
	(0.03)	(-0.90)	(-0.25)	(1.67)	(2.09)
Quarter FE	✓	✓	✓	✓	✓
Stock FE	✓	✓	✓	✓	✓
Total R^2	0.000	0.000	0.000	0.002	0.002

Table 9: Regression result of disagreement regression (31). Standard errors are clustered at analyst pair-firm and firm-quarter level. *, **, *** denotes significance at the 10%, 5%, and 1% level.

8 Conclusion

In this paper, I present a novel exploration of the belief formation process of equity analysts through an analysis of their own written text in research reports. Using information extracted and summarized by LLM, I uncover novel insights into the variation in analysts' attention, their misreaction to different types of information, and the drivers of disagreement in earnings forecasts. I find that analysts' aggregate attention varies significantly over the business cycle and across firms, suggesting that analysts focus more on information pertinent to firm fundamentals. My analysis also reveals distinct misreactions to certain information in earnings forecasts, provides evidence that the "story-statistics gap" can potentially explain the pattern of overreaction, and emphasizes the importance of both asymmetric information and differences of opinion in explaining forecast disagreement.

This paper demonstrates that information from unstructured data can provide novel insights into the behavior of economic agents. By 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 to illuminate the belief and decision-making processes of various economic agents. For example, future research can use the text from earnings call transcripts to study corporate managers' beliefs, analyze mutual fund managers' beliefs from their written annual reports, or examine narratives in FOMC members' speeches to understand their monetary policy decisions. Another fruitful research direction is to use the insights generated from these unstructured data to further discipline theories of belief formation, which should help researchers to navigate the "wilderness" of alternative models of expectation formation (Angeletos et al. (2021)).

References

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.

Anderson, E. W., Ghysels, E., and Juergens, J. L. (2005). Do heterogeneous beliefs matter for asset pricing? *The Review of Financial Studies*, 18(3):875–924.

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.

Ash, E., Gauthier, G., and Widmer, P. (2024). Relatio: Text semantics capture political and economic narratives. *Political Analysis*, 32(1):115–132.

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.

Bali, T. G., Kelly, B. T., Mörke, M., and Rahman, J. (2023). Machine forecast disagreement. Technical report, National Bureau of Economic Research.

Bastianello, F. (2022). Time-series and cross-section of risk premia expectations: A bottom-up approach. *Available at SSRN 4204968*.

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 Chinco, A. (2024). Expected eps x trailing p/e. *Working Paper*.

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., 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. (2024a). Long-term expectations and aggregate fluctuations. *NBER Macroeconomics Annual*, 38(1):311–347.

Bordalo, P., Gennaioli, N., Porta, R. L., and Shleifer, A. (2024b). 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*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bybee, L., Kelly, B., Manela, A., and Xiu, D. (2024). Business news and business cycles. *The Journal of Finance*, 79(5):3105–3147.

Cagan, P. (1956). The monetary dynamics of hyperinflation. *Studies in the Quantity Theory if Money*.

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, 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. (2023). How is chatgpt's behavior changing over time? *arXiv preprint arXiv:2307.09009*.

Chi, F., Hwang, B.-H., and Zheng, Y. (2024). The use and usefulness of big data in finance: Evidence from financial analysts. *Management Science*.

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.

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. and Titman, S. (2006). Market reactions to tangible and intangible information. *The Journal of Finance*, 61(4):1605–1643.

De La O, R. and Myers, S. (2021). Subjective cash flow and discount rate expectations. *The Journal of Finance*, 76(3):1339–1387.

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*.

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.

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.

Gentzkow, M., Kelly, B., and Taddy, M. (2019). Text as data. *Journal of Economic Literature*, 57(3):535–574.

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*.

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.

Groysberg, B., Healy, P. M., and Maber, D. A. (2011). What drives sell-side analyst compensation at high-status investment banks? *Journal of Accounting Research*, 49(4):969–1000.

Haaland, I. K., Roth, C., Stantcheva, S., and Wohlfart, J. (2024). Measuring what is top of mind. Technical report, National Bureau of Economic Research.

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.

Hong, H. and Kubik, J. D. (2003). Analyzing the analysts: Career concerns and biased earnings forecasts. *The Journal of Finance*, 58(1):313–351.

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.

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.

Jiang, H. (2010). Institutional investors, intangible information, and the book-to-market effect. *Journal of Financial Economics*, 96(1):98–126.

Jiang, H. and Sun, Z. (2014). Dispersion in beliefs among active mutual funds and the cross-section of stock returns. *Journal of Financial Economics*, 114(2):341–365.

Kacperczyk, M., Van Nieuwerburgh, S., and Veldkamp, L. (2016). A rational theory of mutual funds' attention allocation. *Econometrica*, 84(2):571–626.

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).

Kohlhas, A. N. and Walther, A. (2021). Asymmetric attention. *American Economic Review*, 111(9):2879–2925.

Kolasinski, A. C. and Kothari, S. (2008). Investment banking and analyst objectivity: Evidence from analysts affiliated with mergers and acquisitions advisors. *Journal of Financial and Quantitative Analysis*, 43(4):817–842.

Korinek, A. (2023). Generative ai for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4):1281–1317.

Kučinskas, S. and Peters, F. S. (2024). Measuring under-and overreaction in expectation formation. *Review of Economics and Statistics*, 106(6):1620–1637.

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*.

Liu, Q., Kusner, M. J., and Blunsom, P. (2020). A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*.

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.

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.

Mayew, W. J. (2008). Evidence of management discrimination among analysts during earnings conference calls. *Journal of Accounting Research*, 46(3):627–659.

MetaAI (2024). The llama 3 herd of models. *https://ai.meta.com/research/publications/the-llama-3-herd-of-models/*.

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.

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.

Nerlove, M. (1958). Adaptive expectations and cobweb phenomena. *The Quarterly Journal of Economics*, 72(2):227–240.

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

Sims, C. A. (2003). Implications of rational inattention. *Journal of monetary Economics*, 50(3):665–690.

Song, W. and Stern, S. (2021). Firm inattention and the transmission of monetary policy: A text-based approach. Technical report, Mimeo, University of Michigan.

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.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. (2023). Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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.

Vissing-Jorgensen, A. (2003). Perspectives on behavioral finance: Does" irrationality" disappear with wealth? evidence from expectations and actions. *NBER macroeconomics annual*, 18:139–194.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

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, the publisher, the authors of the report, the companies mentioned in the report 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 satisfies the filters `Language=English`, `Region=North America`, `Country=United States`, `Report Style=Equity` or `Company (Equity) Reports`, and `Category=Equity`. Then I download the PDF of the analyst research report using the file link associated with each report `ID`.

A.2 Parsing PDF and Preparing Content Chunks

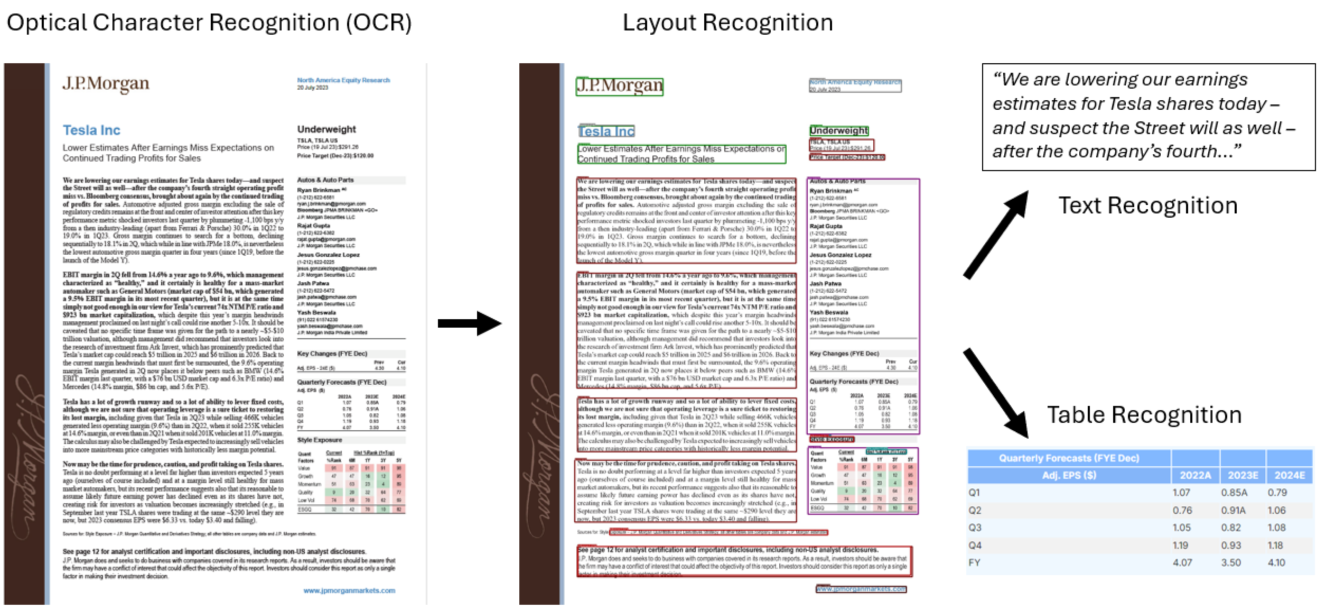


Figure 5: Illustration of parsing analyst reports

A PDF (Portable Document Format) file is an unstructured data format that can contain different layouts and different data formats such as text, pictures, or tables. These versatile features of a PDF requires sophisticated parsing techniques to represent the content into a workable format.

I implement a parsing procedure adopted from the `Deep Document` module in `RAGFlow`, developed by InfiniFlow[32]. The implementation combines Optical Character Recognition (OCR), 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

[32] More details can be found in https://github.com/infiniflow/ragflow/blob/main/deepdoc/README.md

model pre-trained using computer vision techniques to identify different components on a PDF page. For each component, another pre-trained model is used to identify if the component is a textual component that mainly consists of texts, or if it's a tabular component. I only focus on the textual components, and in the final step I extract the texts from the part of the image that corresponds to the textual components. Figure 5 shows the illustration of the parsing process.

A.3 Matching Investext companies

I match Investext companies with their CUSIP. To match the companies mentioned in the analyst research reports, I first remove all reports with no companies identified by Investext. I obtain the CRSP identity file (`stocknames` table) and the IBES identity file (`id` table) from WRDS, and merge them together to create a `ticker-companyname-CUSIP` linking table. For the company identifiers provided by Investext, if they are already in CUSIP format, I directly use it 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 match succeed, I try to match the company name using `fuzzywuzzy` package in Python, with a match cutoff of 90. All CUSIPs identified in this process are kept, as there could be multiple companies mentioned in one research report.

A.4 Matching Investext contributors and authors

I first match the authors of the research reports with the IBES analysts. IBES identifies with a unique analyst code that contains the last name and sometimes the first initial of the analyst (`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 `HumanName` package in Python. For each name that appears as an author of an analyst report in Investext in a given quarter, I try to match it with analysts that provide announcements, activation or review of recommendation or price target to IBES in 12 months before or after the quarter, by selecting on the `anndats`, `actdats`, or `revdats` variable. If there are multiple matches to the last name, I first check if the company CUSIPs for this report overlaps with the CUSIPs for which the analyst provides recommendation or price target for the previous or following 12 months, and I remove those matches with no overlapping CUSIPs. For the remaining multiple matches, if there are multiple authors associated to the report, I require all identified `amaskcd` 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 only focus on brokerage firms that publishes more than 10 reports in a quarter on average. I obtain `estimid` from detail `recommendation` file from IBES. I first match the abbreviated broker names in `estimid` to full broker names (or its subsidiaries) in Investext by resemblance, and check that the analysts associated with the brokerage firm in Investext has amaskcd code that are associated with the same estimid in the same quarter. For the remaining brokerages that doesn't 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 associated with it. As a sanity check, I obtain the brokerage firm matching table from Jared Flake, who constructed the matching between IBES estimid with brokerage firms in 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 checked the matching quality of the brokers in Investext (Law (2023)). At the end of the matching process, I am able to match 115 brokers in the Investext database.

A.5 Preprocessing IBES data

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

It has been noted in previous literature that EPS forecasts in IBES may mechanically differ from realizations if stock splits occur between forecast date and 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 keep the observations where the linkage holds when the forecast is announced. When IBES forecast announcement falls outside the trading dates in CRSP, I use the adjustment factor on the closest preceding trading date in CRSP.

I further keep only stocks with share codes 10 or 11 and with exchange codes 1, 2, or 3. I also drop the observations where the forecast's announcement date (anndats) is after the forecast period end date (fpedats), and drop the observations where the implied forecast horizon from forecast announcement date and forecast's target date aligns with FPI.

A.6 Matching reports with IBES records

In the final step, I match IBES records with Investext 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 published by the identified broker between 1 business day before and 1 business day after the announcement date that contains the identified cusip.

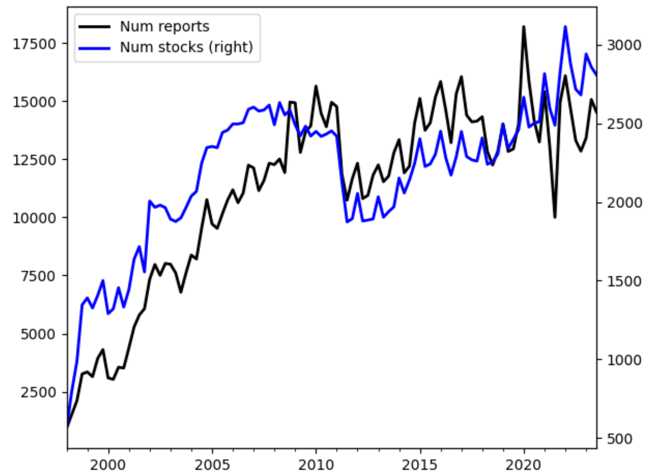
A.6.1 Summary statistics

Table 10 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 1201 brokers. Each report contains about 9 pages. In terms of analysts' research activities, each broker has close to 200 reports every quarter that gets mapped to the IBES records, and each firm on average gets 5 reports per quarter.

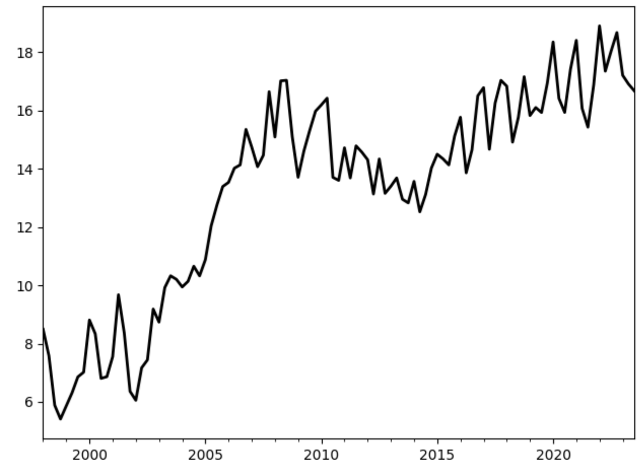
Figure 6 shows how the summary statistics of analyst reports evolve over time. Panel (a) shows that the total number of reports and the number of stocks covered in the matched sample steadily increasing over time, and Panel (b) shows the analyst reports gets longer over time. Both pattern indicates the growing importance of information production by sell-side analysts in the financial markets.

	Matched	Selected brokers	Full IBES
Total # of reports	1144952		
# of IBES announcements	818780	1644069	3963955
# of unique stocks (`cusip`)	6890	10646	13066
# of unique brokerage firms (`estimid`)	115	115	1201
# of unique research analysts (`analys`)	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 10: Summary statistics of the full Investext sample.



(a) Number of reports and number of stocks



(b) Average # of pages per report

Figure 6: Quarterly summary statistics of analyst reports

Table 11 reports the summary statistics of realized earnings yield, as well as 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 in magnitudes, indicating the institutions in the sample are less prone to optimism bias and incorporates new information more strongly in their forecasts.

	Matched	Selected brokers	Full IBES
EPS/P	0.0313	0.0277	0.0315
$Error\_1Y/P$	-0.0014	-0.0033	-0.0046
$Error\_2Y/P$	-0.0012	-0.0046	-0.0068
$Error\_3Y/P$	-0.0009	-0.0029	-0.0049
$Error\_4Y/P$	-0.0192	-0.0221	-0.0226
$Error\_5Y/P$	-0.0432	-0.0446	-0.0454
$Revision\_1Y/P$	0.0006	-0.0001	-0.0001
$Revision\_2Y/P$	0.0011	0.0004	0.0003
$Revision\_3Y/P$	0.0014	0.0003	-0.0001
$Revision\_4Y/P$	0.0026	0.0018	0.0008
$Revision\_5Y/P$	0.0043	0.0026	0.0014

Table 11: Summary statistics of analyst EPS forecasts.

A.7 Illustration of Retrieval-Augmented Generation (RAG) Framework

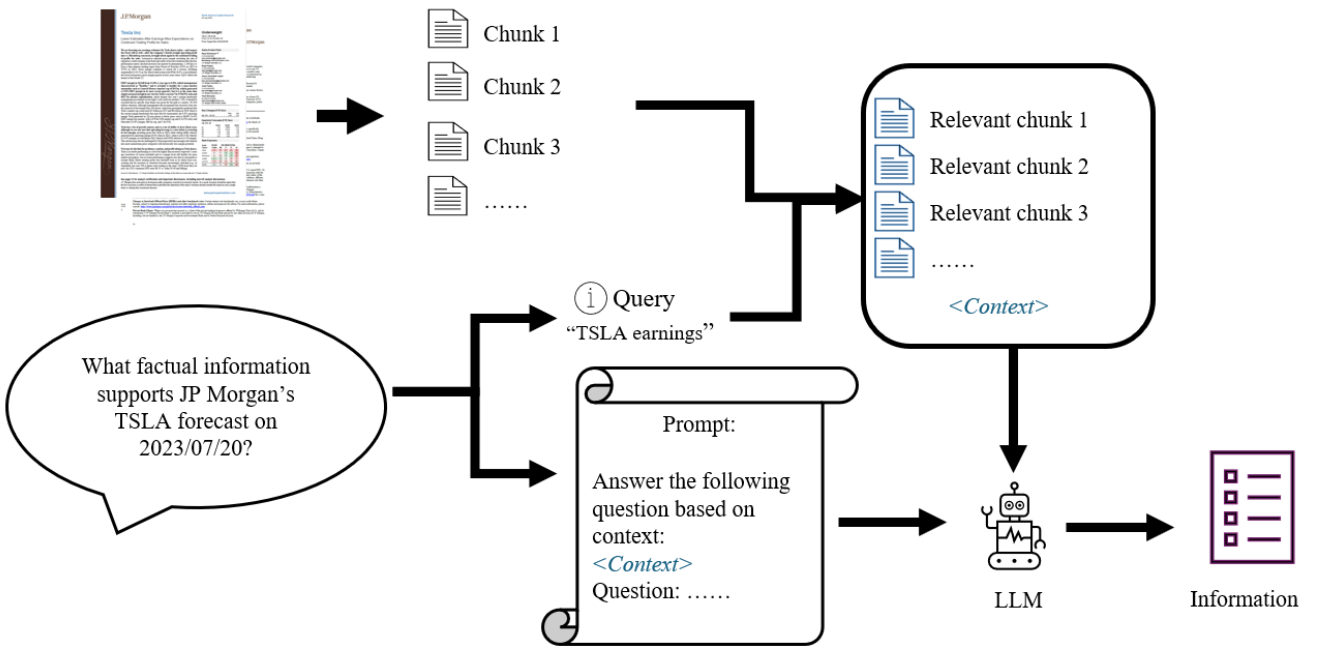


Figure 7: Illustration of the Retrieval-Augmented Generation (RAG) framework.

A.8 Remarks on LLM Implementation

While LLMs have shown astonishing capabilities compared to traditional textual analysis methods, and researchers have become more open to adopting LLMs in their research, others have raised concerns about the validity of LLM outputs. In this section, I briefly address some of the concerns.

One concern with using LLMs is that, given the stochastic nature of text generation, the outputs of LLMs are not reproducible. In a nutshell, the text generation feature of LLMs draws tokens from a distribution that captures the probability of the next token given the input. To address this concern, I set all LLM parameters that controls randomness (`temperature`, `top_k`, and `top_p`) to 0, which means, for a given input, the LLM always generates the same output.

Another form of irreproducibility comes from the fact that many closed-source LLM models (like ChatGPT) are constantly updating based on the newly available text, as well as text from user inputs. Therefore, future users will not reproduce the same results as the same model weights gets refreshed (e.g. Chen et al. (2023)). However, this concern does not apply to open-source LLM models like `Llama3.1` used in this paper, for which the researcher has full control of the model weights.

In addition, since most LLMs are trained on large corpora available on the internet, researchers have raised the concern of look-ahead bias (e.g. Glasserman and Lin (2023), Sarkar and Vafa

(2024)). However, this concern is unlikely to apply to analyst reports, as these reports are not freely available from the internet and therefore are not part of the training corpus of LLMs. I also note that the text-generation capability of the LLM is primarily used for information extraction and summarization, rather than for performing out-of-sample prediction.

B The Informativeness of Analyst Reports

The purpose of this paper is to study analysts' belief formation by examining their written reports. This study is based on the premise that analyst reports provide valuable insights into analysts' subjective beliefs. In this section, I demonstrate that the textual content in these reports is indeed informative about analysts' subjective expectations of future earnings and returns. Additionally, I show that the textual information in analyst reports possesses predictive power in forecasting realized future earnings and returns, beyond analysts' consensus forecasts and common stock characteristics. Notably, a striking finding indicates that the text within these reports can predict analyst-specific forecast errors, suggesting that analyst reports reveal systematic mistakes in analysts' expectations.

To represent the textual information in analyst reports, I obtain the embeddings of the context using `Llama-3.1-8b-Instruct`. An embedding is a (potentially high-dimensional) vector of numerical values designed to capture the contextual meaning of the text and forms the backbone of most state-of-the-art Large Language Models.[33] In `Llama-3.1-8b-Instruct`, each context provided to the model is embedded in a 4,096-dimensional space. Therefore, I obtain a 4,096-dimensional vector $x_{i,j,t}$ for each report written by analyst i about stock j at time t. To mitigate the concern that analyst's numerical forecasts are also included in the report, I remove all numerical symbols from the text before getting the embedding vectors.[34]

B.1 Predicting Individual Forecast Error by Analyst Reports

If an analyst's forecasts are fully rational given his own information and outlook, then his forecast errors should not be predictable. However, I find that individual analyst forecast errors are predictable by his *own reports*. This exercise tries to connect an analyst's forecast error directly to his own information and subjective outlook; the surprising predictability suggests that the systematic errors analysts make can be understood through their own words.[35]

Specifically, for each forecast horizon $h \in \{1, 2, 3, 4, 5\}$ years as well as the long-term-growth (LTG) forecast, I select those analysts who made more than 300 forecasts for any stocks. This step ensures that I have enough written reports from each analyst to train the model. For each selected analyst i (identified by `analys` code in IBES), I predict his forecast error $Error^h_{i,j,t}$ with

[33] See Liu et al. (2020) for a survey of contextual embedding techniques and Touvron et al. (2023) and MetaAI (2024) for specific training details of `Llama-3.1-8b-Instruct`.

[34] Most of the analyst reports can be fit into the 128k token context window in `Llama-3.1`. I truncate excess tokens if the report contains more than 128k tokens.

[35] In Appendix C.1, I also show that analyst reports are useful for predicting the level of earnings forecasts as well as forecast revisions at individual level.

the associated embedding of his report $x_{i,j,t}$

$$Error^h_{i,j,t} = \theta_h x_{i,j,t} + \epsilon_{i,j,t} \tag{32}$$

where $Error^h_{i,j,t}$ is defined as the difference between realized EPS and analyst i's forecast. Following Bouchaud et al. (2019) and De Silva and Thesmar (2024), I scale the EPS numbers by price at the time of the forecast, i.e. $\pi_{j,t+h} = \frac{EPS_{j,t+h}}{p_t}$, and compute $Error^h_{i,j,t} = \pi_{j,t+h} - E^i_t[\pi_{j,t+h}]$, where $E^i_t[\cdot]$ denote analyst i's forecast at time t. I define the realized LTG as $(EPS_{t+5} - EPS_t)/5$, following Bordalo et al. (2024a).

The parameters θ in (32) are estimated in-sample and the model fit is evaluated out-of-sample. I randomly assign 60% of analyst i's forecasts into the training sample and use the remaining forecasts as the test sample. To mitigate the problem of rank-deficiency, I add an L_2 penalty to (32), which is optimally tuned using 5-fold cross-validation in the training data.[36] I evaluate the model predictions based on their out-of-sample R^2, defined as

$$R^2 = 1 - \frac{\sum_{j,t\in OOS}(Error^h_{i,j,t} - \widehat{\theta}_h x_{i,j,t})^2}{\sum_{j,t\in OOS}(Error^h_{i,j,t})^2}. \tag{34}$$

An uninformative model that always predicts zero forecast error yields an R^2 of 0, and positive R^2 means the forecast errors are predictable by an analyst's own reports out-of-sample.

Table 12 summarizes the out-of-sample performance of embedding-based forecast error prediction. The first row shows the average out-of-sample R^2 for predicting each individual analyst's forecast error using his own report. Analyst's own written reports are informative about analyst's own forecast errors, as the forecast error prediction achieves an average out-of-sample R^2 of 0.08 for the 1-year ahead forecast and increases with forecast horizon to 0.30 for 5-year ahead forecast. The second and third rows report the number of analysts with more than 300 forecasts at a specific horizon and the number of analysts with positive R^2. I find that a large share of analysts have positive forecast error prediction R^2, suggesting that the forecast error predictability is a widespread phenomenon among sell-side analysts. In Appendix C.2 I showed that the text-based predictability is not spanned by forecast revision, a usual predictor for forecast error predictability.

[36] That is, I solve the following optimization problem:

$$\min_{\theta} ||Error^h_{i,j,t} - \theta_h x_{i,j,t}||_2 + \lambda||\theta_h||_2 \tag{33}$$

where λ is the hyperparameter controlling the ridge shrinkage.

	Analyst Forecast Horizon					
	1 Year	2 Years	3 Years	4 Years	5 Years	LTG
Avg. R^2	0.079	0.198	0.205	0.314	0.303	0.394
# Analysts	247	197	66	16	7	17
# with Positive R^2	210	190	65	16	7	17

Table 12: Out-of-sample performance of individual forecast error prediction.

	Analyst Forecast Horizon					
	1 Year	2 Years	3 Years	4 Years	5 Years	LTG
Original MSE	0.00048	0.00118	0.00126	0.00137	0.00127	0.47070
Text-adj. MSE	0.00043	0.00089	0.00094	0.00088	0.00085	0.22154
Improvement	9.5%	23.9%	25.0%	35.7%	33.4%	52.9%

Table 13: Mean squared error of original IBES forecast and text-adjusted forecast

B.2 Text adjustment mitigates behavioral biases

The previous literature has documented several significant behavioral patterns in analyst forecasts, including optimism in that analyst forecasts are in general too high (Jensen (2024)); underreaction in short-term forecasts (Bouchaud et al. (2019)); and overreaction in long-term forecasts (Bordalo et al. (2024a)). Given that the text embedding predicts analyst forecast errors, one may naturally wonder: If we adjust analyst forecasts by the predicted forecast error using their own reports, will these behavioral patterns be mitigated, or will they still persist?

In this section, I investigate this question by computing the text-adjusted forecast as

$$E_t^{i,\text{text-adj.}}[\pi_{j,t+h}] = E_t^i[\pi_{j,t+h}] + \widehat{Error}_{i,j,t}^h \tag{35}$$

where $E_t^i[\pi_{j,t+h}]$ is the original forecast from IBES, and $\widehat{Error}_{i,j,t}^h$ is the out-of-sample predicted forecast error. If the model predicts a positive forecast error based on analyst's written report, it means the analyst's forecast is likely to be too low, and the text-adjusted forecast will be higher.

Table 13 shows the mean-squared error of original IBES forecasts as well as that of the text-adjusted forecast. I find that the text-adjusted forecasts are more accurate than analysts' original forecasts, and that the improvement increases with forecast horizon.

I then compare the seminal Coibion-Gorodnichenko regression (Coibion and Gorodnichenko (2015), henceforth CG-regression) results using original and text-adjusted forecasts. Specifically, the CG-regression regresses individual forecast error on individual forecast revision;

$$Error_{i,j,t}^h = \alpha + \beta \cdot Revision_{i,j,t}^h + \epsilon_{j,t}, \tag{36}$$

where revision is defined as the change in the forecast between two periods: $Revision^h_{i,j,t} = E^i_t[\pi_{j,t+h}] - E^i_{t-1}[\pi_{j,t+h}]$. Under rational expectations, the forecast error should be unpredictable and so α and β should be zero. A positive (negative) α indicates overall pessimism (optimism) bias, and a positive (negative) β corresponds to underreaction (overreaction).

Panel A in Table 14 shows the individual CG regression results. A significant negative $\widehat{\alpha}$ is consistent with optimism bias. I estimate positive $\widehat{\beta}$ for short-term forecasts and negative $\widehat{\beta}$ for long-term forecasts, consistent with short-term underreaction and long-term overreaction. Panel B in Table 14 shows the CG-regression results if text-adjusted forecasts and revisions are used instead. A striking observation is that the behavioral patterns documented using raw IBES forecasts are no longer significant. This result further suggests the informativeness of analyst reports in understanding the irrationality in analysts' forecasts.

	Analyst Forecast Horizon					
	1 Year	2 Years	3 Years	4 Years	5 Years	LTG
	Panel A: Original IBES Forecast Errors					
Const.	-0.001	-0.004***	-0.007***	-0.011***	-0.014***	-0.347***
	(-1.11)	(-2.52)	(-7.09)	(-11.54)	(-9.32)	(-12.11)
Original Revision	0.152***	0.165**	-0.150	-0.572***	-0.429**	-0.002***
	(2.59)	(2.24)	(-1.29)	(-5.33)	(-2.02)	(-2.56)
R^2	0.005	0.002	0.002	0.031	0.024	0.001
	Panel B: Text-adjusted Forecast Errors					
Const.	0.000*	0.001	0.000	-0.002	-0.000	0.009
	(1.71)	(1.08)	(0.81)	(-0.13)	(-1.32)	(0.54)
Text-pred. Revision	0.131	0.065	-0.353*	-0.995	-0.175	0.014
	(0.97)	(0.30)	(-1.68)	(-1.32)	(-0.24)	(0.73)
R^2	0.000	0.000	0.001	0.005	0.000	0.001
Stock FE	✓	✓	✓	✓	✓	✓
Analyst FE	✓	✓	✓	✓	✓	✓
# Obs.	46759	35634	10643	3043	950	4109

Table 14: Coibion-Gorodnichenko regression with original IBES earnings forecast and text-adjusted forecast. Numbers in parentheses are t-statistics based on bootstrapped standard errors of parameter estimates to account for sampling error in the generated regressor.

In summary, I document that analysts' written reports are predictive of forecast errors at the individual level. This result suggests that text in written reports is informative about analysts' behavioral biases. For example, if an analyst is usually too optimistic when a firm launches a new product, the predictive regression (32) will pick up a relationship between a negative forecast error and an embedding that captures text saying "new product launch". The discovery of association between textual content and systematic errors lead to the positive out-of-sample R^2.

B.3 Predicting future realized earnings

In this section, I demonstrate that the textual information in analyst reports is useful for predicting future realized earnings, and that the predictability gain is beyond what can be obtained from the consensus IBES earnings forecast or from common stock characteristics.

Specifically, I take the report-level embedding vector $x_{i,j,t}$ and compute the average embedding $\overline{x}_{j,t}$ for each stock j in month t. The average embedding vector $\overline{x}_{j,t}$ can thus be interpreted as the "consensus" textual content the sell-side analysts collectively produce. I then predict realized $h-$year ahead earnings by estimating a pooled linear prediction model;

$$\pi_{j,t+h} = \theta\overline{x}_{j,t} + \epsilon_{j,t+h}. \tag{37}$$

I re-estimate the parameter θ at the end of December each calendar year starting from 12/31/2008 using a rolling 10-year window. I continue to add an L_2 shrinkage for the θ estimates and optimally tune shrinkage using 5-fold cross-validation in each rolling window. After $\hat{\theta}$ is estimated, I use it to make an $h-$year ahead earnings prediction using new embeddings in the following year.

The first row in Table 15 reports the out-of-sample R^2 of realized earnings prediction using average embeddings.[37] I find that the qualitative information in analyst reports is indeed informative about future earnings, as the prediction achieves an out-of-sample R^2 from 0.20 in the 1-year ahead forecast to 0.40 in the 5-year ahead forecast. I also conduct a placebo exercise, where I replace the average embedding vector $\overline{x}_{j,t}$ with a randomly generated 4,096-dimensional vector. The second row in Table 15 shows that the placebo predictor has no predictability, as expected, which validates that the textual embeddings are truly predictive of future earnings.

	Forecast Horizon				
	1 Year	2 Years	3 Years	4 Years	5 Years
Text Pred.	0.20	0.33	0.41	0.39	0.40
Text Pred. (placebo)	0.00	0.00	0.00	0.00	0.00
IBES Cons.	0.77	0.52	0.41	0.22	0.02
JKP Pred.	0.43	0.47	0.45	0.43	0.38

Table 15: Out-of-sample R^2 of future earnings prediction for embedding-based prediction, IBES consensus forecast prediction, and JKP characteristics prediction.

How does the predictive power of embedding-based forecasts compare with the predictive power

[37] Similar to the R^2 formula in (34), I compute total R^2 as $R^2 = 1 - \frac{\sum_{i,t}(\pi_{i,t+h} - \widehat{\pi}_{i,t+h})^2}{\sum_{i,t}\pi_{i,t}^2}$. The denominator is uncentered, as there is no intercept in the embedding vector. De Silva and Thesmar (2024) also uses this R^2 definition and interpret this as the allocative efficiency relative to a perfect foresight optimizer.

analysts' numerical forecasts and of common stock characteristics? While the information in the text in analyst reports is qualitative, the informational content might overlap with the analyst's numerical forecasts or with other observable stock characteristics. To investigate this question, the third row of Table 15 reports the out-of-sample R^2 when using IBES consensus forecasts, and the fourth row reports the out-of-sample R^2 when using the full 153 stock characteristics from JKP.[38] The Analyst consensus forecast achieves high accuracy for short-term forecasts (with a 0.77 out-of-sample R^2 for the 1-year ahead forecast) but deteriorates as the forecast horizon increases, while the characteristics-based forecast achieves pretty stable out-of-sample R^2 about 0.45.[39] Based on these results, I conclude that text itself has predictive power comparable to that of the IBES consensus and of common stock characteristics.

For a formal investigation, I conduct a horse race between the text-based forecast $\widehat{\pi}^{\text{text}}_{j,t+h}$, IBES consensus forecast $E^{\text{IBES}}_t[\pi_{j,t+h}]$, and JKP-characteristics forecast $\widehat{\pi}^{\text{JKP}}_{j,t+h}$ of realized earnings. Specifically, I run the following regression:

$$\pi_{j,t+h} = \beta_{\text{text}}\widehat{\pi}^{\text{text}}_{j,t+h} + \beta_{\text{IBES}}E^{\text{IBES}}_t[\pi_{j,t+h}] + \beta_{\text{JKP}}\widehat{\pi}^{\text{JKP}}_{j,t+h} + \epsilon_{j,t+h}. \tag{38}$$

By controlling for $E^{\text{IBES}}_t[\pi_{j,t+h}]$ and $\widehat{\pi}^{\text{JKP}}_{j,t+h}$, the coefficient β_{text} captures the text-based forecast's predictive power for residual variation in earnings that the IBES consensus and stock characteristics cannot predict.[40] Table 16 reports the regression results. I find that $\widehat{\beta}_{\text{text}}$ is positive and significant and increases with forecast horizon, indicating that the qualitative information in analyst reports is distinctly useful for predicting future earnings beyond their numerical forecasts and common stock characteristics.

B.4 Predicting future returns

Finally, I explore whether the textual information in the analyst reports can be used for better return prediction, by constructing trading strategies based on a return prediction model using textual embeddings. Specifically, at the end of each year starting from 2009, I estimate a ridge regression that predicts the 12-month ahead total return $r_{j,t\to t+12}$ of a stock using the average embedding vector $\overline{x}_{j,t}$ in the rolling past 5 years. Then I construct return forecasts for all stocks

[38] The IBES consensus forecast is obtained as the median estimate from `statsum_epsus` table. The characteristics-based prediction model is the same as (37); I continue to impose L_2 shrinkage in the problem. The training follows the same rolling window and cross-validation procedure as for the embedding-based prediction. The R^2 numbers are reported based on common observations.

[39] The term structure of R^2 is consistent with Van Binsbergen et al. (2023) and De Silva and Thesmar (2024), who find that analyst forecasts have high accuracy and outperform ML-based earnings forecasts in the short-term, but deteriorate and underperform as the forecast horizon increases.

[40] Another way to evaluate the additional informativeness in the text data is to "orthogonalize" text-based predictions with the consensus IBES and stock characterstics-based forecasts and to see if the orthogonalized text-prediction still predicts future earnings. I show in Table 22 in Appendix C.4 that it is indeed predictive of future earnings.

	Forecast Horizon					
	1 Year	2 Years	3 Years	4 Years	5 Years	LTG
β_{text}	0.026***	0.143***	0.262***	0.310***	0.380***	0.965***
	(12.26)	(41.97)	(54.60)	(56.54)	(52.92)	(91.16)
β_{IBES}	0.788***	0.501***	0.375***	0.281***	0.240***	0.259***
	(236.26)	(185.03)	(144.43)	(106.14)	(82.81)	(30.41)
β_{JKP}	0.162***	0.342***	0.409***	0.498***	0.504***	0.015***
	(27.92)	(72.78)	(124.07)	(155.61)	(136.17)	(14.24)
R^2	0.794	0.631	0.596	0.573	0.566	0.212
# Obs.	67831	62191	47649	22818	14566	32991

Table 16: Regression result for (38). Numbers in parentheses are t-statistics based on bootstrapped standard errors of parameter estimates to account for sampling error in the generated regressor.

for the next 12 months using the estimated model with the current quarter embeddings. Finally, I sort stocks by their predicted return at the end of each year into 10 bins, and compute the long-short trading strategy return for the next 12 months by going long the top equal-weighted decile and shorting the bottom equal-weighted decile.

To compare with the return prediction performance of subjective return expectations, I also construct an equal-weighted decile spread using the 12-month-ahead price target implied expected return using the consensus price target from IBES. Following Jensen (2024) and Andrei et al. (2023), I obtain the consensus 12-month ahead price target from the IBES `statsum_ptgus` table and compute the price-target implied expected return for stock j at time t as[41]

$$\widetilde{E}_t[r_{i,t+12}] = \frac{\texttt{Price Target}_{i,t+1} - \texttt{Current Price}_{i,t}}{\texttt{Current Price}_{i,t}}. \tag{39}$$

Figure 8 plots the cumulative return of the trading strategy that uses embedding-based return prediction and of the top-minus-bottom decile strategy based on price targets. I find that the text-based strategy consistently outperforms the price-target based strategy, suggesting that there is qualitative information in analyst reports that is not incorporated into analyst's forecasts.

Table 17 reports the summary statistics of the two strategies and how the text-based strategy model compares with other asset pricing models. The text-based strategy achieves an annualized Sharpe ratio of 0.82—36.5% higher than the price-target based strategy. The text-based strategy achieves significant alphas with respect to the price-target strategy as well as relative to prominent asset pricing models, including the Fama-French 3-factor and 5-factor models, the Hou-Xue-Zhang q-factor model, and the Stambaugh-Yuan mispricing model.[42] These results suggest that there

[41] I do not include dividend forecast, as dividend forecasts in IBES are scarce and could lead to fewer data points. The expected return measure with and without dividend forecast have a correlation of 0.996.

[42] Another noticeable finding is that the text-based long-short strategy has large significant loadings on the

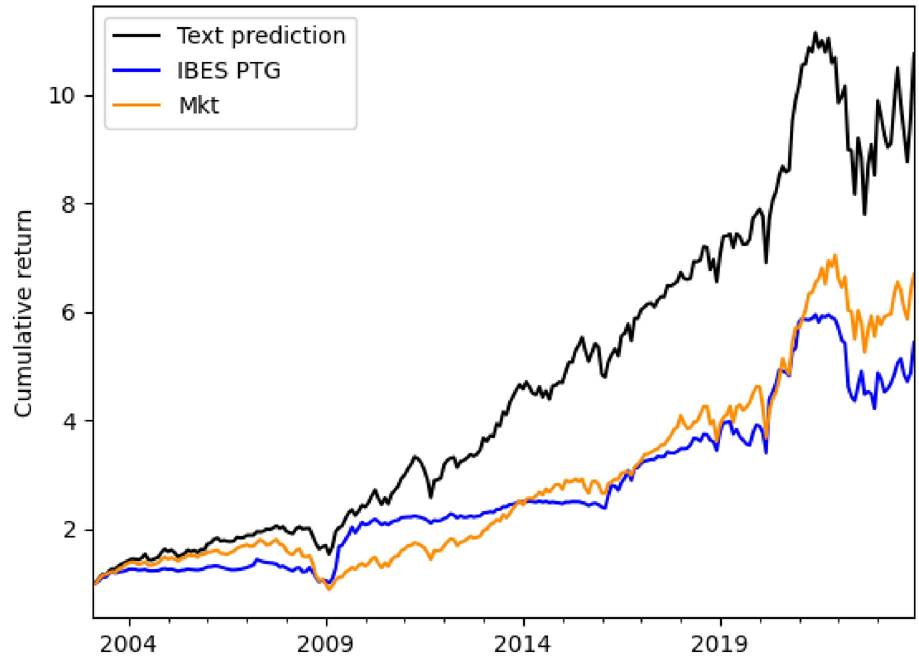


Figure 8: This figure shows the cumulative return of a trading strategy that uses an embedding-based prediction model ("Text pred") and the implied expected return based on consensus price targets from IBES ("IBES PTG-implied return"), as well as the market portfolio ("Mkt"). Each strategy is based on an equal-weighted decile long-short spread and is scaled based on ex-post volatility to have an annualized volatility of 10%.

is additional return predictability in the textual information in the analyst reports that adds significant economic value beyond the numerical price target predictions.

market portfolio. This is somewhat related to the finding in Bastianello (2022) who finds that analysts' expected return forecasts align with the CAPM prediction, which implies that a long-short portfolio based on analysts' return forecasts will have a positive loading on the market.

Panel (a): Summary Statistics	Text-based strategy	PTG-based strategy
Sharpe Ratio	0.82	0.60
Improvement	36.5%	

Panel (b): Model comparison	IBES-PTG	CAPM	FF3F	FF5F	HXZ	SY
Text-based Alpha	0.38%***	0.52%***	0.51%***	0.59%***	0.66%***	0.84%***
	(3.82)	(2.05)	(3.01)	(3.27)	(2.58)	(2.91)
IBES-PTG	0.19***					
	(6.03)					
PTG-based Alpha		0.21%	0.23%	0.33%	0.48%**	0.69%***
		(0.92)	(1.11)	(1.45)	(2.09)	(2.59)

Table 17: This table reports the performance of a return prediction model based on embeddings from analyst reports. Panel (a) reports the annualized out-of-sample average return and Sharpe ratio of trading strategies that uses an embedding-based prediction model ("Embedding") and the implied expected return based on consensus price target from IBES ("IBES-PTG"). Each strategy is based on equal-weighted decile long-short spread and is scaled based on ex-post volatility to have an annualized volatility of 10%. Panel (b) reports the alphas and betas of the embedding-based trading strategy with respect to various asset pricing models.

C Additional Results for Appendix B

C.1 Predicting earnings forecast level and revision

In this section, I show that analyst reports are highly predictive of analyst's earnings forecast level, as well as their forecast revisions. I adopt a similar procedure in Section B.1 and predict using text embeddings

$$\begin{aligned} E_t^i[\pi_{j,t+h}] &= \theta_h x_{i,j,t} + \epsilon_{i,j,t} \\ Revision_{i,j,t} &= \theta_h x_{i,j,t} + \epsilon_{i,j,t} \end{aligned} \tag{40}$$

Table 18 and 19 report the result for earnings forecast level prediction and forecast revision. Both earnings forecast and forecast revisions are strongly predicted by the text embeddings of the analysts reports.

	Analyst Forecast Horizon					
	1 Year	2 Years	3 Years	4 Years	5 Years	LTG
	Panel (A): individual out-of-sample R^2					
Avg. R^2	0.71	0.77	0.77	0.83	0.83	0.77
# Analysts	237	211	101	23	11	23
# with Positive R^2	237	211	101	23	11	23
	Panel (B): pooling regression					
Text Prediction	0.895***	0.869***	0.895***	0.884***	0.774***	0.950***
	(88.80)	(68.26)	(57.49)	(26.09)	(19.42)	(46.76)
Analyst FE	✓	✓	✓	✓	✓	✓
R^2	0.736	0.791	0.799	0.825	0.798	0.807
nobs	55235	49659	23863	4915	2354	6283

Table 18: Out-of-sample performance of individual earnings level prediction.

	Analyst Forecast Horizon					
	1 Year	2 Years	3 Years	4 Years	5 Years	LTG
	Panel (A): individual out-of-sample R^2					
Avg. R^2	0.0395	0.066	0.037	0.049	0.073	0.025
# Analysts	210	182	61	16	8	30
# with Positive R^2	166	157	46	15	6	23
	Panel (B): pooling regression					
Text Prediction	0.734***	0.782***	0.722***	0.85***	0.593***	0.869***
	(65.95)	(69.17)	(42.06)	(32.84)	(15.83)	(46.90)
Analyst FE	✓	✓	✓	✓	✓	✓
R^2	0.061	0.081	0.061	0.049	0.045	0.042
nobs	55235	49659	23863	4915	2354	6283

Table 19: Out-of-sample performance of individual forecast revision prediction.

C.2 Forecast error predictability: comparing with forecast revision

The literature has long been using forecast revisions to proxy for the information perceived by the agent, and predictability of the forecast error by the forecast revision demonstrates a deviation from the rational reaction. Is the text of analyst reports more informative about analysts' forecast

	Analyst Forecast Horizon					
	1 Year	2 Years	3 Years	4 Years	5 Years	LTG
Text Prediction	0.745***	0.857***	0.865***	0.691***	0.778***	0.917***
	(68.16)	(68.14)	(43.99)	(19.04)	(14.00)	(34.00)
Revision	0.236***	0.303***	0.149***	-0.191***	-0.281***	-0.001
	(36.37)	(52.88)	(15.70)	(-13.91)	(-6.48)	(-0.07)
R^2	0.07	0.17	0.13	0.09	0.10	0.25
Analyst FE	✓	✓	✓	✓	✓	✓
Obs.	53070	40070	11795	3025	963	3214

Table 20: Comparing text-based forecast error prediction with prediction using forecast revision. Numbers in parentheses are t-statistics based on bootstrapped standard errors of parameter estimates to account for sampling error in the generated regressor.

error than forecast revisions? To investigate this, I regress

$$Error_{i,j,t} = \beta \cdot \widehat{Error}_{i,j,t} + Revision_{i,j,t} + \epsilon_{i,j,t},$$

where $Error$ is the realized forecast error, $\widehat{Error}_{i,j,t}$ is the text-predicted forecast error out-of-sample, and $Revision_{i,j,t}$ is an analyst's forecast revision. For better comparison, I scale $Revision$ to have the same standard deviation as text-based prediction.

Table 20 shows the result. I find that the coefficient on text-based prediction is significantly larger than the coefficient on revision, suggesting that the text of analyst reports captures more information that analysts misreact to.

C.3 Forecast error predictability by different analysts and brokerages

In this section, I show that the forecast error predictability by analyst's own written text does not differ significantly for different types of analysts and brokerage firms. Specifically, for each analyst making a forecast, I compute the following variables at the announcement time:

- Analyst experience: number of years since the analyst first appears in IBES.
- Stock coverage: number of stocks the analyst issues forecasts for in a given year.
- Brokerage analyst count: number of analysts associated with the brokerage firm that issues forecasts in a given year.
- Broker stock coverage: number of unique firms the brokerage house issue foreccasts for in a given year.

For variable, I compute for each analyst the average value across the forecasts the analyst made. I then look at the average out-of-sample R^2 for analysts above or below the median across all analysts. Table 21 shows the result. For each column, $\overline{R^2}_{\text{above}}$ ($\overline{R^2}_{\text{below}}$) reports the average out-of-sample forecast error prediction R^2 for analysts in the top/bottom half of the distribution. $\overline{R^2}_{\text{above}} - \overline{R^2}_{\text{below}}$ reports the differences in the mean of R^2, and $p(\overline{R^2}_{\text{above}} = \overline{R^2}_{\text{below}})$ reports the associated $p-$value.

		Experience	Stock coverage	Broker analyst count	Broker stock coverage
1 Year	$\overline{R^2}_{\text{above}}$	0.081	0.072	0.086	0.081
	$\overline{R^2}_{\text{below}}$	0.079	0.088	0.074	0.079
	$\overline{R^2}_{\text{above}} - \overline{R^2}_{\text{below}}$	0.002	-0.016	0.012	0.002
	$p(\overline{R^2}_{\text{above}} = \overline{R^2}_{\text{below}})$	(0.86)	(0.20)	(0.35)	(0.89)
2 Years	$\overline{R^2}_{\text{above}}$	0.206	0.197	0.199	0.187
	$\overline{R^2}_{\text{below}}$	0.190	0.199	0.196	0.209
	$\overline{R^2}_{\text{above}} - \overline{R^2}_{\text{below}}$	0.017	-0.002	0.003	-0.021
	$p(\overline{R^2}_{\text{above}} = \overline{R^2}_{\text{below}})$	(0.41)	(0.90)	(0.87)	(0.29)
3 Years	$\overline{R^2}_{\text{above}}$	0.218	0.199	0.176	0.168
	$\overline{R^2}_{\text{below}}$	0.197	0.216	0.239	0.247
	$\overline{R^2}_{\text{above}} - \overline{R^2}_{\text{below}}$	0.021	-0.016	-0.063*	-0.080**
	$p(\overline{R^2}_{\text{above}} = \overline{R^2}_{\text{below}})$	(0.56)	(0.66)	(0.08)	(0.03)

Table 21: Average out-of-sample R^2 for analysts above and below the median in different analyst characteristics

C.4 Earnings prediction with orthogonalized text-based forecasts

In this section, I show that the orthogonalized text-based future earnings forecast still predicts future earnings significantly. I first orthogonalize text-based forecasts by regressing it on IBES consensus forecast and characteristics-based forecasts

$$\widehat{\pi}^{\text{text}}_{j,t+h} = \gamma_{\text{IBES}} E^{\text{IBES}}_t[\pi_{j,t+h}] + \gamma_{\text{JKP}} \widehat{\pi}^{\text{JKP}}_{j,t+h} + \epsilon_{j,t+h} \tag{41}$$

Panel (A) in 22 reports the result. IBES consensus and JKP characteristics-based forecast only explain partially the variations in text-based predictions, with R^2 ranging from 0.26 to 0.55. I then predict realized $\pi_{j,t+h}$ by the residual $\widehat{\epsilon}_{j,t+h}$, and Panel (B) shows it still predicts future earnings significantly.

	Forecast Horizon					
	1 Years	2 Years	3 Years	4 Years	5 Years	LTG
	Panel (A): $\widehat{\pi}^{\text{text}}_{j,t+h} = \gamma_{IBES} \cdot E^{\text{IBES}}_t[\pi_{j,t+h}] + \gamma_{JKP} \cdot \widehat{\pi}^{\text{JKP}}_{j,t+h} + \epsilon_{j,t}$					
β_{IBES}	0.170***	0.293***	0.367***	0.301***	0.283***	0.483***
	(27.27)	(53.93)	(64.34)	(41.42)	(38.24)	(35.49)
β_{JKP}	0.162***	0.342***	0.409***	0.498***	0.504***	0.015***
	(17.53)	(31.76)	(40.45)	(32.23)	(32.07)	(9.49)
R^2	0.261	0.445	0.553	0.534	0.543	0.361
	Panel (B): $\pi_{j,t+h} = \beta_{\text{text resid.}} \cdot \widehat{\pi}^{\text{text resid.}}_{j,t+h} + \epsilon_{j,t}$					
$\beta_{\text{text resid.}}$	0.026***	0.143***	0.262***	0.31***	0.38***	0.965***
	(12.26)	(41.97)	(54.60)	(56.54)	(52.92)	(91.16)
R^2	0.000	0.004	0.014	0.017	0.023	0.083

Table 22: Realized earnings prediction with residualized text-based forecast. Numbers in parentheses are t-statistics based on bootstrapped standard error of parameter estimates to account for sampling error in generated regressor.

D Proofs and Additional Results for Section 3

D.1 A Rational Inattention Benchmark

In this section, I provide a rational benchmark for attention choice and belief formation in the model described in Section 3.1. When the analyst has rational expectations, $F_t[\cdot] = E_t[\cdot]$. I study how attention choice m_k and analyst's reaction A_k change with component importance a_k when the analyst is rational but faces cost in acquiring a higher precision signal. I also derive the implications of attention choice for earnings and return predictability.

The following lemma characterizes the rational agent's expectation revision, given the data generating process and signal process in Section 3.1.

Lemma 1 (Rational analyst's belief revision)
A rational analyst's expectation revision about future earnings y_{t+1} satisfies

$$Rev_t[y_{t+1}] = \sum_{k=1}^{K} a_k V_k(\tau_k)\tau_k(s_t - E_{t-1}[s_t]) \tag{42}$$

where $V_k(\tau_k)$ denotes analyst's conditional variance of $\theta_{k,t}$ in steady state.

Next, I consider the analyst's attention choice. I assume that the rational analyst makes a one-time attention allocation choice m_k in the initial period $t = 0$ to maximize his rationally expected lifetime utility, which is the discounted squared error

$$\max_{\tau} \quad \mathcal{U} = -\mathrm{E}_0[\sum_{t=1}^{\infty} \beta^t (y_{t+1} - \mathrm{E}_t[y_{t+1}])^2] - C(\tau) \tag{43}$$

where β denotes the discount factor and $C(\tau)$ denotes the cost of acquiring more signal precision, which I assume is positive, increasing in all τ_k, and convex. The following proposition characterizes analyst's rational attention choice

Proposition 3 (Optimal attention allocation for a rational analyst)
The optimal precision τ_k^* is characterized by

$$a_k^2 V_k'(\tau_k^*) + \frac{\partial C(\tau)}{\partial \tau_k^*} = 0 \tag{44}$$

where $V_k(\tau)$ is implicitly defined by $\sigma_\theta^2 = (1-\rho_k^2+\tau\sigma_\theta^2)V(\tau)+\tau\rho_k^2 V_k(\tau)^2$. Furthermore, the optimal attention is increasing in the importance of a structural component

$$\frac{\partial m_k^*}{\partial a_k} > 0 \tag{45}$$

Intuitively, the effect of higher attention increases signal precision about component k, and the rational agent will make more a accurate forecast if the analyst pays more attention to the component that matters more for future earnings. If volatility in the fundamental component is less volatile (lower σ_θ), reduction in posterior variance will be larger given the same increase in signal precision. Therefore, attention is more sensitive to a when σ_θ is low, as the marginal benefit of increasing attention is larger.

D.2 Micro-founded Belief Formation Models

In this section, I provide micro-founded belief formation models and discuss how to interpret the text-instrumented CG coefficient within each model.

Diagnostic Expectation Bordalo et al. (2024b) proposes the diagnostic expectation model where agents suffer from representative heuristic, which leads to overreaction. This model maps directly to the belief updating model in 7. In particular, upon observing the signal $s_{k,t}$, the analyst's perceived conditional probability distribution for $\theta_{k,t}$ follows

$$\widehat{f}(\theta_{k,t}|s_{k,t}) = f(\theta_{k,t}|s_{k,t}) \cdot \left(\frac{f(\theta_{k,t}|s_{k,t} - E_{t-1}[s_{k,t}])}{f(\theta_{k,t}|(s_{k,t} - E_{t-1}[s_{k,t}]) = 0)}\right)^{\gamma_k} Z \tag{46}$$

where $f(\cdot)$ denotes the objective conditional probability density function, $\gamma_k > 0$ captures the representativeness distortion associated with each component k, and Z is a normalizing constant. Notably, given now analysts observe multiple signals, I allow for γ_k to vary across fundamental components, which can generate different degrees of overreaction.

By the similar proof as in Bordalo et al. (2024b), one can show that the analyst's expectation of $\theta_{k,t}$ follows

$$F_t[\theta_{k,t}] = \mathrm{E}_t[\theta_{k,t}] + \gamma_k G_k (E_t[\theta_{k,t}] - E_{t-1}[\theta_{k,t}]) \tag{47}$$

where G_k is the Kalman gain solved in Lemma 1. This belief updating rule directly maps to (7).

Misestimation of Signal Precision In this setup, the analyst misestimates the precision of the signal. He believes that the precision of the signal is $\hat{\tau}_k$ instead of τ_k. In particular, overestmating the precision of new information ($\hat{\tau}_k > \tau_k$) is usually interpreted as overconfidence (e.g. Daniel et al. (1998)). Models where the analyst needs to estimate signal strength (e.g. Augenblick et al. (2025)) can also lead to analyst update using a signal precision different from the truth.

Denote $\hat{G}_k = \frac{\hat{\tau}_k}{\bar{\tau}_k + \hat{\tau}_k}$ the perceived Kalman gain. In this case, the analyst's belief updating follows

$$
\begin{aligned}
F_t[\theta_{k,t}] &= (1-\hat{G}_k)E_{t-1}[\theta_{k,t}] + \hat{G}_k s_{k,t} \\
&= E_t[\theta_{k,t}] + (\hat{G}_k - G_k)(s_{k,t} - E_{t-1}[\theta_{k,t}]) \\
&= E_t[\theta_{k,t}] + (\hat{G}_k/G_k - 1)(E_t[\theta_{k,t}] - E_{t-1}[\theta_{k,t}])
\end{aligned} \tag{48}
$$

which maps exactly to (7). In particular, if the analyst underestimates the signal precision ($\hat{\tau}_k < \tau_k$), he will underreact to new information, whereas if he overestimates the signal precision due to overconfidence ($\hat{\tau}_k > \tau_k$), he overreacts to new information.

Misperception about Component Importance In this setup, the analyst misperceives the component importance a_k. In particular, he believes that the k'th component matters for future earnings y_{t+1} by $\hat{a}_k$ instead of a_k. In this case, the analyst's belief about $\theta_{k,t}$ is rational, i.e. $F_t[\theta_{k,t}] = \mathrm{E}_t[\theta_{k,t}]$, but his forecast for future earnings is

$$
F_t[y_{k,t+1}] = E_t[y_{k,t+1}] + (\hat{a}_k - a_k)E_t[\theta_{k,t+1}] \tag{49}
$$

While this formula is different from (7), the text-instrumented CG regression still identifies over/underreaction due to misperception $\hat{a}_k$, because

$$
\begin{aligned}
Cov(y_{k,t+1} - F_t[y_{k,t+1}], F_t[y_{k,t+1}] - F_{t-1}[y_{k,t+1}]) &= -\frac{\hat{a}_k}{a_k} Cov((\hat{a}_k - a_k)E_t[\theta_{k,t}], E_t[\theta_{k,t}] - E_{t-1}[\theta_{k,t}]) \\
&\propto -\frac{\hat{a}_k}{a_k}(\hat{a}_k - a_k)
\end{aligned} \tag{50}
$$

When $Cov(F_t[y_{k,t+1}], E_t[y_{k,t+1}]) > 0$, $\hat{a}_k$ and a_k have the same sign, so the sign of text-instrumented CG regression coefficient is driven by $\hat{a}_k - a_k$: When the analyst overestimate the importance of component k ($\hat{a}_k > 0$), he overreacts to the information about component k and the text-instrumented CG coefficient is negative, and vice versa for underestimation of a_k leading to underestimation.

Sticky Belief Bouchaud et al. (2019) proposes a model of sticky belief in the form of

$$
F_t[\theta_{k,t}] = (1-\lambda)E_t[\theta_{k,t}] + \lambda F_{t-1}[\theta_{k,t}] \tag{51}
$$

with belief stickiness $\lambda \in (0,1)$. Rearranging terms, we have

$$
F_t[\theta_{k,t}] = E_t[\theta_{k,t}] - \lambda(E_t[\theta_{k,t}] - F_{t-1}[\theta_{k,t}]) \tag{52}
$$

which is the same as (7) but replacing $\mathrm{E}_{t-1}[\theta_{k,t}]$ with $F_{t-1}[\theta_{k,t}]$. The text-instrumented CG regression identifies the underreaction due to belief stickiness because

$$Cov(\theta_{k,t} - F_t[\theta_{k,t}], F_t[\theta_{k,t}] - F_{t-1}[\theta_{k,t}]) = \lambda(1-\lambda)Var(E_t[\theta_{k,t}] - F_{t-1}[\theta_{k,t}]) \tag{53}$$

which comes directly from rearranging terms in (52).

Adaptive Beliefs The model of adaptive beliefs of Cagan (1956) and Nerlove (1958) takes the form of

$$F_t[\theta_{k,t}] = \delta s_{k,t} + (1-\delta)F_{t-1}[\theta_{k,t}] \tag{54}$$

with $\delta \in (0,1)$. Rearranging terms, we can directly show that

$$Cov(\theta_{k,t} - F_t[\theta_{k,t}], F_t[\theta_{k,t}] - F_{t-1}[\theta_{k,t}]) = \delta(1-\delta)Var(s_{k,t} - F_{t-1}[\theta_{k,t}]) \tag{55}$$

which means the text-instrumented CG regression coefficient captures the underreaction due to adaptivity.

Extrapolative Belief I use the backward-looking extrapolation formation in Afrouzi et al. (2023), specifically

$$F_t[\theta_{k,t}] = s_{k,t} + \phi(s_{k,t} - s_{k,t-1}) \tag{56}$$

Since

$$\begin{aligned} s_{k,t-2} &= \theta_{k,t-2} + u_{k,t-2} \\ s_{k,t-1} &= \rho_k\theta_{k,t-2} + \sigma_\theta\epsilon_{k,t-1} + u_{k,t-1} \\ s_{k,t} &= \rho_k(\rho_k\theta_{k,t-2} + \sigma_\theta\epsilon_{k,t-1}) + \sigma_\theta\epsilon_{k,t} + u_{k,t} \end{aligned} \tag{57}$$

We have

$$\begin{aligned} Cov(FE, FR) &= Cov(-\phi(s_{k,t} - s_{k,t-1}) - u_{k,t}, (1+\phi)s_{k,t} - (1+2\phi)s_{k,t-1} + \phi s_{k,t-2}) \\ &= -\rho_k((1+\phi)\rho_k - \phi)(1-\rho_k)^2 V - \phi((1+\phi) - (1-\rho_k)((1+\phi)\rho_k - (1+2\phi)))\sigma_\theta^2 \\ &\quad - ((1+\phi)^2 + (1+2\phi)\phi)\tau_k^{-1} \end{aligned} \tag{58}$$

where V is the steady-state posterior variance of $\theta_{k,t-2}$. We can conclude that when $\rho_k \to 1$, i.e. when the fundamental component is persistent, $Cov(FE, FR)$ is always negative, which implies the analyst overreact to information $s_{k,t}$ due to extrapolating from the recent trend.

D.3 Proofs

Proof for Proposition 1 The result follows directly from the conditional independence assumption (CIA) $X_t^k \perp \epsilon_t$. $\square$

Proof of Proposition 2 The result follows from

$$\beta_k \propto Cov(y_{t+1} - F_t[y_{t+1}], F_t[y_{t+1}] - F_{t-1}[y_{t+1}]) \tag{59}$$

Denote $g_k \equiv a_k V_k(\tau_k)\tau_k > 0$ be the rational updating rule (Kalman gain) such that

$$E_t[y_{t+1}] = E_{t-1}[y_{t+1}] + g_k(s_{k,t} - E_{t-1}[s_{k,t}]) \tag{60}$$

and let $\hat{A}_k = A_k/g_k$, we have

$$\begin{aligned} y_{k,t+1} - F_t[y_{k,t+1}] &= y_{k,t+1} - E_t[y_{k,t+1}] - A_k(s_{k,t} - E_{t-1}[s_{k,t}]) \\ &= y_{k,t+1} - E_t[y_{k,t+1}] - \hat{A}_k(E_t[y_{k,t+1}] - E_{t-1}[y_{k,t+1}]) \end{aligned} \tag{61}$$

and

$$F_t[y_{t+1}] - F_{t-1}[y_{t+1}] = (1 + \hat{A}_k)(E_t[y_{t+1}] - E_{t-1}[y_{t+1}]) - \hat{A}_k(E_{t-1}[y_{t+1}] - E_{t-2}[y_{t+1}]) \tag{62}$$

Since $Cov((E_t[y_{t+1}] - E_{t-1}[y_{t+1}]), (E_{t-1}[y_{t+1}] - E_{t-2}[y_{t+1}])) = 0$ for rational expectation, we have

$$\beta_k \propto -\hat{A}_k(1 + \hat{A}_k) \tag{63}$$

When $Cov(F_t[y_{t+1}], E_t[y_{t+1}]) > 0$, we have $1 + \hat{A}_k > 0$, so $\beta_k \propto -\hat{A}_k$ and thus $\beta_k \propto -A_k$. □

Proof of Lemma 1 The standard formula for Gaussian updating implies that

$$E_t[\theta_{k,t}] = \frac{\bar{\tau}}{\bar{\tau}_k + \tau_k} E_{t-1}[\theta_{k,t}] + \frac{\tau_k}{\bar{\tau}_k + \tau_k} s_{k,t} \tag{64}$$

where $\bar{\tau}_k \equiv \text{var}[\theta_{k,t}|\Omega_{t-1}]$. Notice that $E_{t-1}[\theta_{k,t}] = E_{t-1}[s_{k,t}]$, $\text{var}[\theta_{k,t}|\Omega_t] = \bar{\tau}_k + \tau_k$, and $E_t[y_{k,t+1}] = a_k E_t[\theta_{k,t}]$, the result follows from rearranging terms. □

Proof of Proposition 3 Notice that in steady state, $E_0[(y_{t+1} - E_t[y_{t+1}])^2] = \text{var}[y_{t+1}|\Omega_t]$ and given the independence assumption of $\theta_{k,t}$ across k, we have

$$\text{var}[y_{t+1}|\Omega_t] = \sum_k a_k^2 \text{var}[\theta_{k,t}|\Omega_t] := \sum_k a_k^2 V_k(\tau_k) \tag{65}$$

The autoregressive nature of $\theta_{k,t}$ means that starting at date t, the conditional variance about next period's fundamental is

$$\text{var}_t[\theta_{k,t+1}] = \rho_k^2 \text{var}_t[\theta_{k,t}] + \sigma_\theta^2, \tag{66}$$

while the Bayesian updating formula implies

$$\text{var}_{t+1}[\theta_{k,t+1}]^{-1} = \text{var}_t[\theta_{k,t+1}] + \tau_k \tag{67}$$

Combining (66) and (67) and notice that in steady state $\text{var}_{t+1}[\theta_{k,t+1}] = \text{var}_t[\theta_{k,t}] \equiv V_k(\tau_k)$, we have $V_k(\tau_k)$ follows

$$\sigma_\theta^2 = (1 - \rho_k^2 + \tau\sigma_\theta^2)V(\tau) + \tau\rho_k^2 V_k(\tau)^2 \tag{68}$$

and $V_k'(\tau_k) < 0$, $V_k''(\tau_k) > 0$. The optimality condition of τ_k^* therefore follows the first order condition. To study the relationship between optimal attention and component importance, notice that

$$m_k^* = \frac{\tau_k^*}{\sum_{k'} \tau_{k'}^*} = \frac{1}{1 + \tau_k^{*-1} \sum_{k' \neq k} \tau_{k'}^*} \tag{69}$$

and since $\tau_{k'}^*$ doesn't depend on a_k for $k' \neq k$, it's equivalent to prove $\frac{\partial \tau_k^*}{\partial a_k} > 0$. Given the entropy cost and by implicit function theorem, we have

$$2a_k V_k'(\tau_k^*) + \left(a_k^2 V''(\tau_k^*) + \frac{\partial^2 C(\tau)}{\partial \tau_k^*}\right) \frac{\partial \tau_k^*}{\partial a_k} = 0 \tag{70}$$

We conclude $\frac{\partial \tau_k^*}{\partial a_k} > 0$ by noting $V_k''(\tau) > 0$ and the convexity of $C(\tau)$. □

D.4 Implications for Return Predictability

Assuming the stock is priced by a risk-neutral investor who has earnings expectation formed following Lemma 1. The stock price is therefore simply given by

$$P_t = \sum_{l>1} \frac{E_t[y_{t+h}]}{(1+r)^h} \tag{71}$$

Define $R_t := (P_{t+1} + y_{t+1}) - (1+r)P_t$ as the return of the stock in the model, and define $\lambda_k := Cov(R_t, \theta_{k,t})$ be the return predictability of component k, the following proposition characterizes these quantities in the model

Proposition 4 (Rational inattention and return predictability)
When the stock is priced by a risk-neutral agent whose expectation is given by (7), denote $\xi_k = \frac{1}{1+r-\rho_k}$ and τ_k^* characterized in Proposition 3, we have the stock return given by

$$R_t = \sum_{k=1}^{K} a_k\big((1 + \xi_k g_k \rho_k)(\theta_{k,t} - E_t[\theta_{k,t}]) + \xi_k g_k \epsilon_{k,t+1}\big) \tag{72}$$

and the predictability

$$\lambda_k = Cov(R_t, \theta_{k,t}) = a_k(1 + \underbrace{\xi_k \rho_k V(\tau_k^*)\tau_k^*}_{:=q_k})V(\tau_k^*) \tag{73}$$

where $V_k(\tau)$ is implicitly defined by $\sigma_\theta^2 = (1 - \rho_k^2 + \tau\sigma_\theta^2)V(\tau) + \tau\rho_k^2 V_k(\tau)^2$. Note that $\lambda_k = 0$ when $a_k = 0$ (component irrelevant to future earnings) or $\tau_k^* = \infty$ (no attention cost).

Proposition 4 shows that the return predictability is entirely driven by the predictability of forecast error using realized fundamental $\theta_{k,t}$. The overall effect of limited attention on return predictability is actually driven by three components. First, $V(\tau_k^*)$ captures the noisy information channel: as the realization of fundamental component $\theta_{k,t}$ is unobservable, it's not fully reflected in price and contributes to predictability. When the agent pays more attention to k (increase in τ_k^*), he learns more about $\theta_{k,t}$ and price becomes more informative. This channel implies a negative relationship between attention and return predictability.

Second, the term q_k captures the effect of fundamental persistence. When the agent pays more attention to component k, his belief will respond more his signal at time $t+1$, which means the price will be more responsive to $s_{k,t+1}$. As the fundamental realization is persistent, higher attention means $\theta_{k,t}$ will correlate more with price at $t + 1$. This channel implies a positive relationship between attention and return predictability.

Finally, the term a_k captures the implicit relationship between attention and component importance. As described in Proposition 3, higher attention is associated with higher component importance when agent is making optimal attention allocation. Therefore, fundamental component should be more predictive of future return when attention is high. This channel implies a positive relationship between attention and return predictability.

Taken together, the theoretical association between attention and return predictability can actually be non-monotonic. Figure 9 provides a numerical example, where return predictability is hump-shaped in optimal attention.

Proof of Proposition 4 We can define $P_t^k := \sum_{l>1} \frac{E_t[y_{t+h}]}{(1+r)^l}$ and $R_{t+1}^k := (P_{t+1}^k + y_{t+1}^k) - (1+r)P_t^k$. Given the independence of $\theta_{k,t}$ across k, we know $R_{t+1} = \sum_k R_{t+1}^k$ and thus $Cov(R_{t+1}, \theta_{k,t}) = Cov(R_{t+1}^k, \theta_{k,t})$. The price strip P_t^k can be computed as

$$\begin{aligned} P_t^k &= a_k \sum_{l>1} \frac{\rho_k^{l-1}}{(1+r)^l} E_t[\theta_{k,t}] \\ &= a_k \xi_k E_t[\theta_{k,t}] \end{aligned} \tag{74}$$

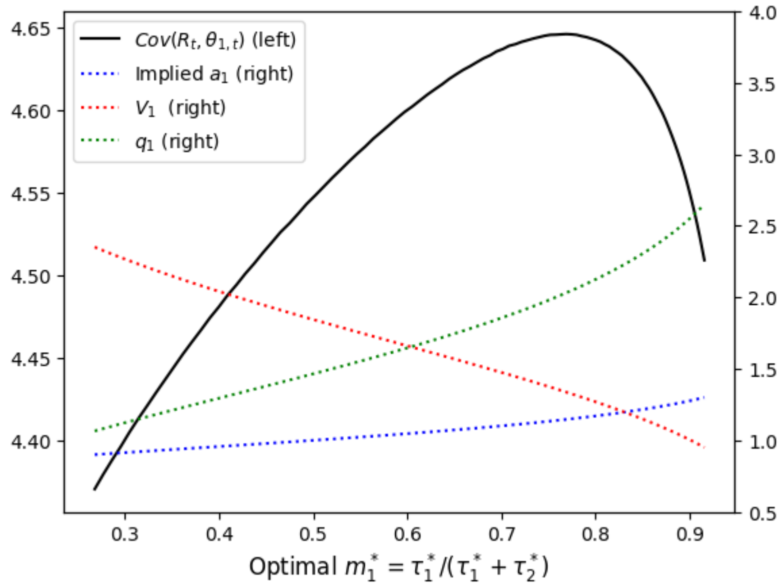


Figure 9: Numerical example of attention and return predictability. The cost function $C(\tau)$ is set to be an cost on entropy $C(\tau) = \mu \prod_k \tau_k$. The parameters not detailed in the figure are $\rho_1 = \rho_2 = 0.9$, $r = 0.05$, $\mu = 50$.

Thus, we have

$$R_{t+1}^k = y_{k,t+1} - a_k E_t[\theta_{k,t}] + a_k \xi_k (E_{t+1}[\theta_{k,t+1}] - \rho_k E_t[\theta_{k,t}]) \tag{75}$$

Also note that from (42) we have

$$\begin{aligned} E_{t+1}[\theta_{k,t+1}] - \rho_k E_t[\theta_{k,t}] &= E_{t+1}[\theta_{k,t+1}] - E_t[\theta_{k,t+1}] \\ &= g_k(s_{k,t+1} - E_t[\theta_{k,t+1}]) \\ &= g_k \rho_k (\theta_{k,t+1} - E_t[\theta_{k,t}]) + g_k \epsilon_{k,t} \end{aligned} \tag{76}$$

Combining terms, we have

$$R_{t+1}^k = a_k(1 + \xi_k g_k \rho_k)(\theta_{k,t} - E_t[\theta_{k,t}]) + a_k \xi_k g_k \epsilon_{k,t+1} \tag{77}$$

and the covariance follows

$$\begin{aligned} Cov(R_{t+1}^k, \theta_{k,t}) &= a_k(1 + \xi_k g_k \rho_k) Cov(\theta_{k,t} - E_t[\theta_{k,t}], \theta_{k,t}) \\ &= a_k(1 + \xi_k g_k \rho_k) V_k(\tau_k) \end{aligned} \tag{78}$$

D.5 Correlated Signals

[To be done]

E Prompts for RAG

In this section, I list the set of prompts used for Retrieval Augmented Generation (RAG). The RAG process is implemented by feeding relevant context to a large language model and prompting it to answer a set of questions based on the context provided. I use the Chat Completion functionality of `Llama-3-8b-instruct` model pre-trained by Meta AI. To reduce hallucination and enhance information retrieval quality, I implement a multi-step process in prompting.

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.
Q2: Q2: On the scale of 0-100, what is your confidence level of your answer? Answer a number from 0 to 100.
Output format:
{'Q1': ANSWER to Q1, 'Q2': ANSWER to Q2}

Figure 10: 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}

Figure 11: 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}

Figure 12: Prompt #2: Separating factual and subjective statements

Prompt #3: Extracting financial economic concepts

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 factual 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, ...}

Figure 13: Prompt #3: Extracting narratives

F Information Extraction Results

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."

Narrative tuple extracted by LLM: factual content:

- "SG&A": "-1% to +4%"
- "Corporate costs": "right size"
- "IT": "improving"
- "Marketing spend": "pre-COVID"

Narrative tuple extracted by LLM: subjective content:

- "SG&A": "difficulty"

Table 23: Example of factual and subjective content summarization

(a) Profitability

(b) Financial Condition

(c) Business Operation

(d) Corporate Management

(e) Growth

(f) Valuation

(g) Macroeconomics

Figure 14: Word cloud of topics. The size of a noun is proportional to the noun's (log) frequency within the topic.

G Additional Examples of LLM Responses

G.1 Validity of Identifying Stock Information

As a proof of validation, Figure 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.[43] 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 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.

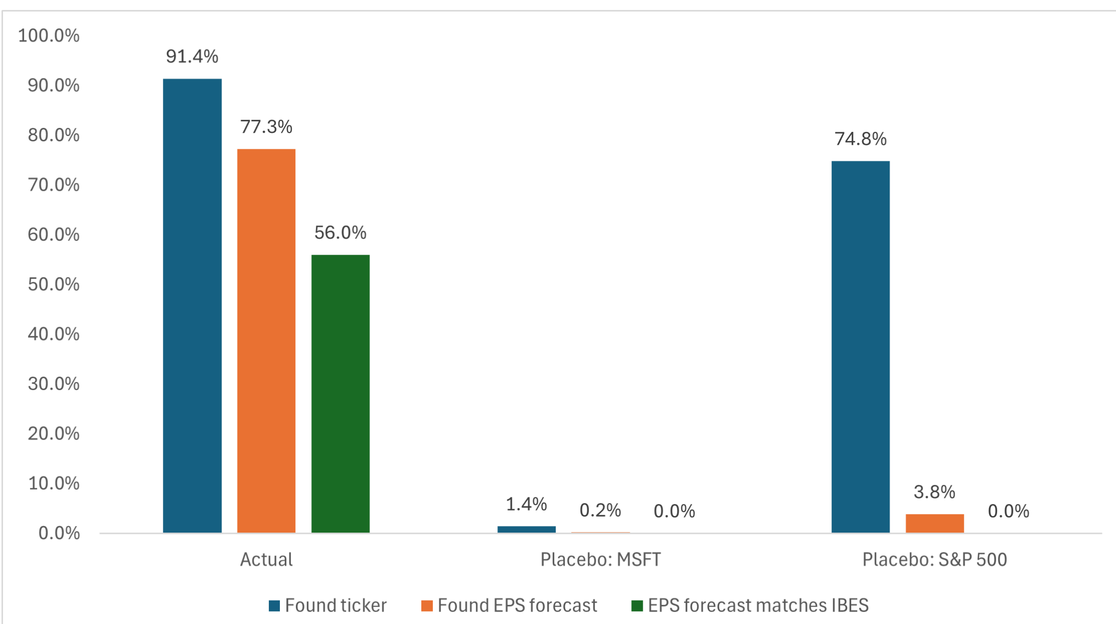


Figure 15: RAG retrieval validity of finding stock EPS information.

[43]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.

G.2 Separating Factual versus 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 Handing 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 adds $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. Asa 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."

G.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:

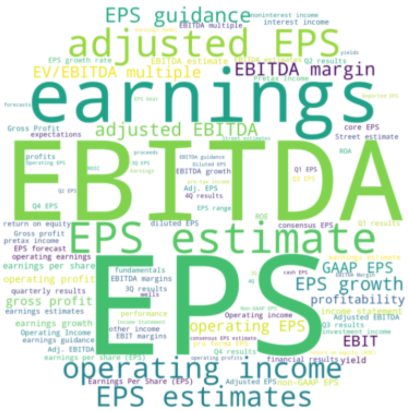
- "Debt": "little changed"
- "Orders": "N/A"
- "Operating income": "10% decrease"

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"

G.4 Word cloud for 25 labels

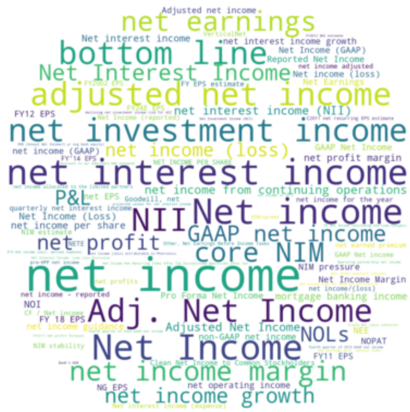


(a) “earnings”

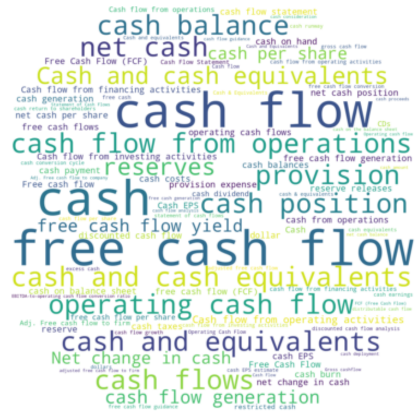
(b) “revenue”

(c) “expense”

(d) “margin”



(e) “net income”



(f) “cash”

(g) “debt”

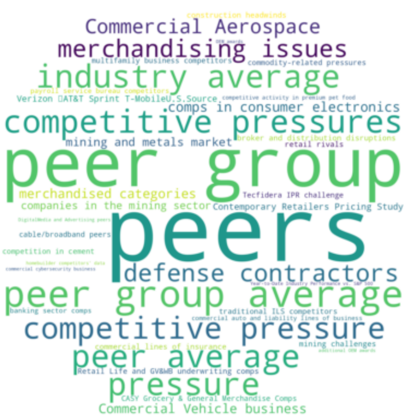
(h) “assets”

(i) “balance sheet”

(j) “inventory”

(k) “customer”

(l) “product”



(m) “industry competition”

(n) “management”

(o) “research and development”

(p) “merger and acquisition”

(q) “tax”

(r) “regulation”

(s) "growth"

(t) "valuation"

(u) "stock price"

(v) "macroeconomic condition"

(w) "international"

(x) "demand"

(y) "supply"

Figure 16: Word cloud for 25 labels from zero-shot classification. The size of a noun is proportional to the noun's (log) frequency within the label.

H Additional results for Section 5

H.1 Time-series Variation in Attention: Panel Regression

	Profitability	Financial conditions	Business operation	Corporate management	Growth	Macroeconomics
Log IP	0.087	-0.896***	-0.094*	0.832***	1.375***	-0.114*
	(1.48)	(-16.51)	(-1.78)	(14.01)	(22.51)	(-1.90)
VIX	-0.002***	0.002***	0.001**	-0.002***	-0.004***	0.004***
	(-6.94)	(7.95)	(2.23)	(-5.22)	(-14.66)	(14.59)
Trend	-0.151***	0.293***	-0.264***	-0.046***	-0.219***	0.253***
	(-8.12)	(14.86)	(-14.46)	(-2.51)	(-11.18)	(13.22)
R^2	0.001	0.003	0.002	0.001	0.005	0.002
Analyst FE	✓	✓	✓	✓	✓	✓
Firm FE	✓	✓	✓	✓	✓	✓
Nobs	392776	392776	392776	392776	392776	392776

Table 24: Panel regression of attention on business cycle variables. Standard errors are two-way clustered at analyst-firm and firm-quarter level. *, **, *** denotes significance level at 10%, 5% and 1%.

H.2 Cross-industry variation in attention

In this section, I explore how analysts' attention differs across industries. Specifically, I run a regression of average analysts' attention to firm j in quarter t on a dummy for each of the 10 SIC industry groups:

$$\overline{Attn}_{j,t}^{k} = \alpha + \beta \mathbf{I}\{j \in G_s\} + \epsilon_{j,t}, \tag{79}$$

where $\overline{Attn}_{j,t}^{k}$ is average analyst attention to topic k for firm j in quarter t, and $\mathbf{I}\{j \in G_s\}$ is an indicator that takes a value of 1 if stock j is in SIC industry S. The coefficient β thus measures the difference between analyst attention to topic k in industry s and its average in other industries. I control for quarter fixed effect. Standard errors are double-clustered by firm and quarter.

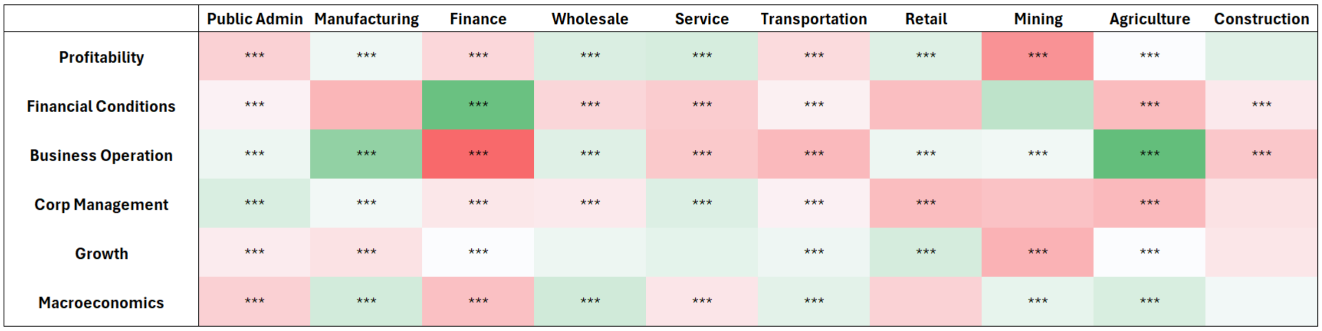


Figure 17: Differences in analyst attention to topics across industries. A stronger red (green) color denotes a more negative (positive) β estimate from regression (79). *** denotes significance at 1% level.

Figure 17 shows the regression results as a heatmap for better visualization. I find that there is significant heterogeneity in the information choice across industries. For example, when forecasting earnings for finance firms, analysts pay more attention to financial conditions and pay less attention to business operations. This makes sense as financial firms operate using financial assets and debt and are usually highly levered. Business Operation becomes more significantly more important when the analyst is forecasting manufacturing or agriculture firms. Overall, these results suggests that industry is an important determinant in analysts' attention to different types of information.

H.3 Attention in Short vs Long-term forecasts: Fixed Effect Regression

As a formal test, I run the following regression of attention to each topic k on a dummy variable indicating that the analyst is making a long-term forecast, while controlling for stock and analyst fixed effects:

$$Attn^k_{i,j,t} = \delta_0 + \delta_{LT} \cdot \mathbf{1}\{\text{LT Forecast}\} + \eta_j + \eta_i + \epsilon_{i,j,t}. \tag{80}$$

Here, $\widehat{\delta}_{LT}$ can be interpreted as the difference of attention to topic k in announcements when analysts make long-term forecasts relative to attention in announcements when analysts only short-term forecasts. Table 25 reports the regression results, which confirms that analysts decrease their attention to a firm's profitability information and increase attention to financial conditions, business operations, and growth information when making a long-term forecast, although due to the triple fixed effects at analyst, quarter and stock level, the magnitudes of attention differences are small.

	Profitability	Financial conditions	Business operation	Corporate management	Growth	Macroeconomics
$\widehat{\delta}_{LT}$	-0.0040***	0.0005**	0.0010***	-0.0001	0.0003***	-0.0001
	(-10.81)	(2.01)	(4.55)	(-0.762)	(2.45)	(-0.74)
$\widehat{\delta}_{LT}/\overline{Attn}^{ST}$	-1.36%	0.56%	0.98%	-0.18%	0.69%	-0.20%
Analyst FE	✓	✓	✓	✓	✓	✓
Quarter FE	✓	✓	✓	✓	✓	✓
Stock FE	✓	✓	✓	✓	✓	✓
Nobs	389181	389181	389181	389181	389181	389181

Table 25: Regression result of (80). $\widehat{\delta}_{LT}/\overline{Attn}^{ST}$ reports the estimated $\widehat{\delta}_{LT}$ as a percentage of average attention to each topic in short-term forecast announcements. *, **, *** denotes significance level at 10%, 5% and 1%. Standard errors are double clustered at analyst-firm and firm-quarter level.

H.4 Cross-analyst variation in attention to firm-level information

In this section, I investigate the cross-analyst variation in attention to firm-level information. First, I estimate the following fixed-effect model for each topic k:

$$Attn^k_{i,j,t} = \gamma_{j\times t} + \epsilon_{i,j,t} \tag{81}$$

$$Attn^k_{i,j,t} = \tau_{i,j,t} + \gamma_{j\times t} + \epsilon_{i,j,t}, \tag{82}$$

where $\gamma_{j\times t}$ captures the stock-quarter fixed effect and $\tau_{i,j,t}$ captures observable analyst-specific features. If analysts all follow the optimal attention choice, the choice of attention of different agents in the same information environment should be the same. This implies that adding $\tau_{i,j,t}$ in regression (82) should not help to capture additional variation in $Attn^k_{i,j,t}$, and that the model fits of regression (81) and (82) should be the same.

I first run regressions (81) and (82), where I use an analyst fixed effect in (82), i.e. $\tau_{i,j,t} = \mathbf{I}_i$. I then compute the R^2 of regression (82) relative to regression (81):

$$\text{Relative } R^2 = 1 - \frac{MSE(\text{regression (82)})}{MSE(\text{regression (81)})} \tag{83}$$

If analysts have homogeneous attention, then the relative R^2 should be 0. Table 26 shows however, that the relative R^2 are significantly positive for all topic attentions. This highlights that there is significant heterogeneity in the attention choice across analysts: different analysts choose to focus on different types of information even when they are forecasting the same firm at the same time.

Revenue	Earnings	Margin	Debt	Cash	Growth	Expense	Equity	Assets	Others
0.041***	0.042***	0.031***	0.037***	0.044***	0.030***	0.029***	0.047***	0.027***	0.025***

Table 26: Relative R^2 as calculated from formula (83). *** denotes significance at the 1% level using Diebold-Mariano test.

Which analyst attributes explain different attention to different kinds of information when analysts make forecasts for the same firm? I focus on two aspects of analyst attributes that the previous literature has shown to have an impact on forecasts: analysts' experience and analysts' coverage. I compute three measures of analyst experience:

- $Exp_{i,t}$: number of years between forecast announcement date and the first time the analyst issues an IBES forecast
- $ExpFirm_{i,j,t}$: number of years between forecast announcement date and the first time the analyst issues an IBES forecast for firm j
- $ExpInd_{i,j,t}$: number of years between forecast announcement date and the first time the analyst issues an IBES forecast for any firm in the same industry as j

and two measures of analyst coverage:

- $FirmCover_{i,t}$: number of firms analyst i issues forecasts for in the current year
- $IndCover_{i,t}$: number of industries analyst i issues forecasts for in the current year

I then run a regression of $Attn^k_{i,j,t}$ on the experience and coverage measures, controlling for stock-quarter and analyst fixed effects:

$$\begin{aligned} Attn^k_{i,j,t} = {} & \beta_1 Exp_{i,t} + \beta_2 ExpFirm_{i,j,t} + \beta_3 ExpInd_{i,j,t} + \beta_4 FirmCover_{i,t} + \beta_5 IndCover_{i,j,t} \\ & + \gamma_{j\times t} + \tau_i + \epsilon_{i,j,t}. \end{aligned} \tag{84}$$

Table 27 reports the results. I find that after controlling for an analyst fixed effect, analysts with different experience do not seem to exhibit significant difference in their attention choices. On the other hand, analysts with different firm or industry coverage do choose to focus on different topics, suggesting that the cognitive burden of information processing is a potential driver of analysts' choice of different information to focus on.

	Revenue	Earnings	Margin	Debt	Cash	Growth	Expense	Equity	Assets	Others
Exp	-0.0	0.0	-0.001	0.001	-0.001	-0.001*	0.001	-0.0	0.0	0.001
	(-0.234)	(0.339)	(-1.32)	(0.751)	(-0.927)	(-1.749)	(1.584)	(-0.117)	(0.321)	(0.776)
ExpFirm	-0.000	0.000	0.000	-0.001	0.000	0.001	-0.000	0.000	0.000	-0.000
	(-0.734)	(0.322)	(0.445)	(-0.962)	(0.049)	(1.117)	(-0.253)	(0.166)	(0.089)	(-0.007)
ExpInd	0.0	0.001	-0.0	-0.0	0.0	0.001**	-0.001	-0.001*	-0.0	-0.0
	(0.206)	(1.059)	(-0.133)	(-0.259)	(0.326)	(2.027)	(-1.064)	(-1.719)	(-0.319)	(-0.445)
FirmCover	-0.002*	0.001	-0.001	-0.002***	0.0	0.002***	0.001***	0.001	-0.001	-0.0
	(-1.702)	(0.901)	(-0.924)	(-3.055)	(0.256)	(3.874)	(2.392)	(1.424)	(-1.068)	(-0.84)
IndCover	0.008***	-0.002	0.0	0.005**	0.005***	-0.007***	-0.003*	-0.005***	-0.001	-0.001
	(2.877)	(-0.838)	(0.233)	(2.208)	(2.387)	(-3.764)	(-1.946)	(-2.435)	(-0.662)	(-0.498)
Analyst FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Stock×Quarter FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Relative R^2	0.043	0.043	0.032	0.037	0.045	0.03	0.03	0.049	0.028	0.026

Table 27: Regression results of regressing topic attention on analyst experience and coverage (84). The attention measure is normalized to have zero mean and unit standard deviation. Standard errors are two-way clustered at the firm and time level. *, **, *** denote significance level at 10%, 5% and 1% levels. Relative R^2 is measured against a model with only stock × quarter fixed effects.

H.5 Stock characteristics for Profitability and Financial Condition

Profitability characteristics		Financial Condition characteristics	
sales	Revenue	fnl_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	dltnetis_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	dltnetis_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_bel1	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 28: List of stock characteristics as proxies for Profitability and Financial Conditions

I Additional results for Section 6

	1 Year	2 Years	3 Years	4 Years	5 Years
Revision	0.063***	0.070***	-0.153***	-0.286***	-0.291***
	(15.20)	(9.65)	(-17.79)	(-18.60)	(-13.29)
Quarter FE	✓	✓	✓	✓	✓
Analyst FE	✓	✓	✓	✓	✓
Firm FE	✓	✓	✓	✓	✓
Adj. R^2	0.002	0.001	0.002	0.010	0.009
Num Obs.	2291998	2014373	644399	111282	58788

Table 29: Original CG-regression results on individual earnings forecasts: full IBES sample. Standard errors are double clustered at analyst-firm and firm-quarter level. *** denotes significance at 1% confidence level.

	Top 50 Principal Components					Top 80 Principal Components				
	1 Year	2 Years	3 Years	4 Years	5 Years	1 Year	2 Years	3 Years	4 Years	5 Years
Profitability	0.026***	0.032***	-0.009	-0.025	-0.036	0.029***	0.037***	-0.005	-0.026	-0.068
	(10.25)	(5.20)	(-0.80)	(-0.88)	(-0.91)	(11.04)	(5.80)	(-0.41)	(-0.90)	(-1.56)
Financial Condition	0.007***	0.013*	-0.022*	-0.013	-0.035	0.008***	0.012*	-0.026***	-0.003	-0.039
	(2.49)	(1.83)	(-1.83)	(-0.50)	(-0.87)	(2.79)	(1.77)	(-2.11)	(-0.11)	(-0.95)
Business Operation	0.012***	0.010	0.003	-0.039	-0.116***	0.013***	0.005	0.005	-0.043	-0.115***
	(4.01)	(1.36)	(0.21)	(-1.38)	(-2.95)	(4.17)	(0.71)	(0.42)	(-1.52)	(-2.86)
Corporate Management	0.002	-0.007	-0.007	-0.036	-0.046	0.003	-0.006	-0.010	-0.055***	-0.077*
	(0.72)	(-1.08)	(-0.60)	(-1.44)	(-1.07)	(1.20)	(-0.86)	(-0.83)	(-2.26)	(-1.76)
Growth	0.006***	0.008	-0.013	-0.046	-0.040	0.006***	0.013**	-0.016	-0.039	-0.024
	(2.44)	(1.33)	(-1.12)	(-1.74)	(-0.97)	(2.38)	(2.01)	(-1.37)	(-1.45)	(-0.57)
Macroeconomics	0.008***	0.020***	-0.006	-0.039	-0.088**	0.008***	0.015**	-0.009	-0.044	-0.107***
	(2.83)	(2.96)	(-0.51)	(-1.36)	(-2.20)	(2.80)	(2.27)	(-0.70)	(-1.55)	(-2.69)
Quarter FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Analyst FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Firm FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 30: Results of text-instrumented CG regression using top 50 and 80 principal components of `FinBERT` embedding vectors. Analyst, quarter and firm fixed effects are removed in the first-stage regression. Standard errors are estimated based on block bootstrap. *, **, *** denotes significance level at 10%, 5% and 1%.

	Forecast Horizon				
	1 Year	2 Years	3 Years	4 Years	5 Years
Profitability	0.228***	0.332***	0.049	0.394	-1.009
	(6.04)	(4.21)	(0.44)	(1.17)	(-1.59)
Financial Condition	0.041	0.122***	-0.586***	-1.793***	-0.986
	(0.55)	(0.77)	(-2.80)	(-2.97)	(-1.18)
Business Operation	0.323***	0.199	-0.334	-2.114***	-1.384***
	(4.00)	(1.25)	(-1.53)	(-3.86)	(-2.38)
Corporate Management	-0.489***	-1.105***	-1.071***	-1.388	-1.003
	(-3.39)	(-4.03)	(-2.54)	(-1.54)	(-0.80)
Growth	-0.129	0.183	-0.105	-0.881	2.271
	(0.85)	(0.64)	(-0.26)	(-0.86)	(0.95)
Macroeconomics	0.002	0.045	-0.276	1.868**	1.106
	(0.02)	(0.17)	(-0.86)	(1.92)	(0.95)
Quarter FE	✓	✓	✓	✓	✓
Analyst FE	✓	✓	✓	✓	✓
Firm FE	✓	✓	✓	✓	✓
Adj. R^2	0.004	0.001	0.003	0.018	0.023

Table 31: Results of attention-interacted CG regression

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

Standard errors are double clustered at analyst-firm and firm-quarter level. *, **, *** denotes significance level at 10%, 5% and 1%.