Analysts' Belief Formation in Their Own Words

Shikun (Barry) Ke*

This draft: November 16, 2024
Preliminary draft. Comments are welcome.

Abstract

I study the formation of analysts' subjective belief about stock earnings using their own written text from over 1.2 million equity research reports. Text in analyst reports strongly predicts analysts' forecast errors. Using a Large Language Model, I differentiate between factual and subjective content in analyst reports, and distill them into interpretable components of a firm's fundamentals. I document four novel findings regarding analysts' subjective beliefs. First, I show that analysts pay pro-cyclical attention to earnings growth information and counter-cyclical attention to financial distress information. They also reduce attention to profitability information when making long-term forecasts. Second, I find evidence that analysts under-react to information about firm's revenue, earnings and margin in their short-term earnings forecasts. Third, I uncover patterns in analysts' reasoning when forming their subjective outlooks, and identify meaningful misalignment with the ex-post realizations in firm's fundamentals. Finally, I find that both asymmetric information and differences in subjective outlooks contribute to disagreement in earnings forecasts. Together, these results offer new insights into the formation of subjective beliefs about stock earnings in reality.

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

*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 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? How do they combine different types of information to form beliefs? Which types of information do agents misreact to? And what explains disagreement in agents' beliefs? Answering these questions is essential for understanding agents' beliefs in reality, yet typical 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 1.2 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 an ideal 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 with, I demonstrate that the text in analyst reports is informative about analysts' subjective earnings beliefs. I represent the texts using high-dimensional embedding vectors and use them to predict analysts' forecasts, 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. In particular, I find that the textual embedding of analyst reports has stronger predictive power for forecast errors compared to analysts' forecast revisions——a common proxy for the information analysts observe. This

[1] A non-exhaustive list of paper that study survey-based measures of subjective expectation 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. (2024). See Adam and Nagel (2023) for a review of the use of subjective expectation data in asset pricing.

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

suggests that we can study analysts' systematic deviations from rationality by examining the granular information contained in analyst reports. Finally, 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 extract and summarize detailed content from analyst reports. Compared to traditional textual analysis methods, such as dictionary-based searching and topic modeling, LLMs are superior tools thanks to their ability to understand semantics and comprehend natural language in a manner similar to humans. By using LLMs in a RAG framework, where the model generates answers based on the retrieved context, I am able to extract information with great flexibility.

I proceed with information extraction 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 identify financial concepts (e.g., nouns) in both factual and subjective content, along with adjectives associated with those financial concepts. This step results in 8,456,795 occurrences of 401,715 unique financial concepts and 5,528,993 occurrences of 228,326 adjectives. Finally, to further distill the information, I use zero-shot classification to categorize the most common financial concepts into 10 topics related to different components of a firm's fundamentals, and to classify adjectives as positive, negative, or neutral.[3]

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

First, I study analysts' attention to different fundamental information. I measure analysts' attention to a particular fundamental component by simply counting the frequency of related nouns in the factual content, and examine how their attention varies across time, stocks, analysts, and forecast horizons. I find that aggregate attention varies significantly

[3]These components are "Assets", "Cash", "Expense", "Debt", "Earnings", "Equity", "Growth", "Margin", "Revenue", and "Others". See Section 4.2 for details.

over time and across business cycles: analysts' attention to profitability and growth-related information (topics such as Revenue, Earnings, and Growth) is pro-cyclical, while their attention to information related to financial distress (topics such as Debt, Cash, and Expense) is counter-cyclical.

I also document significant variation in analysts' attention in the cross-section of stocks: they pay more attention to Expense for value firms and to Growth for growth firms. Additionally, analysts focus more on Debt and Cash when a firm has high debt levels or high cash flow volatility. Analysts' stock and industry coverage explain their attention to different topics, suggesting that cognitive burden is a potential driver of analysts' choice to focus on different types of information. Finally, analysts pay more attention to a firm's profitability when making short-term forecasts, while focusing more on Growth and non-accounting information when making long-term forecasts.

Second, I revisit the previous literature's findings on analysts' systematic mis-reaction in their earnings forecasts by investigating what information analysts mis-react to. I capture the directional information in each fundamental component using the adjectives associated with it, and regress analysts' forecast errors on these directional pieces of information. Since these pieces of information are extracted from analysts' own reports, nonzero regression coefficients will indicate irrational reactions to information about certain fundamental components. This approach is analogous to the seminal Coibion-Gorodnichenko regression (Coibion and Gorodnichenko (2015), CG-regression henceforth), but replaces forecast revisions with granular information directly extracted from analyst reports.

I find positive regression coefficients for the directional information about Revenue, Earnings, and Margin, and a negative regression coefficient for Expense in short-term forecast error prediction. These results imply that analysts' short-term forecasts are not high enough when they observe good news about a firm's profitability, and not low enough when they receive bad news. This pattern is consistent with sticky beliefs about profitability as the main driver of under-reaction in short-term earnings forecasts (e.g., Bouchaud et al. (2019)).

Third, I relate the extracted factual information to analysts' subjective outlooks, measured using the subjective content in analyst reports. By doing so, I shed light on how analysts process the different information they receive to form subjective outlooks about various aspects of a firm's fundamentals. Similar to the directional factual information, I measure analysts' directional outlook for each fundamental component using adjectives in the subjective content summarized by the LLM. I then regress the directional outlook on directional factual information. I find that analysts tend to believe that different aspects of a firm's fundamentals will continue to change in the same direction. For example, when analysts note that a firm has an increasing margin as a piece of factual information, they

tend to express a subjective outlook that the firm will have a higher or increasing margin in the future. I also uncover analysts' reasoning across fundamental components: when they receive factual information that a firm has high debt, they tend to believe the firm will have lower cash in the future.

Do analysts have the correct reasoning? To shed light on this question, I predict future realizations of firm fundamentals using their current values, which serve as a benchmark for analysts' subjective reasoning. I find that while analysts' beliefs about trend continuation in fundamental components are mostly consistent with the persistence observed in realized firm fundamentals, there are noticeable misalignments between analysts' subjective reasoning and the relationships in ex-post realizations. For example, analysts tend to believe that a firm will experience higher growth in the future when they receive news about high growth in factual information, whereas a firm's realized growth (measured by the 1-year change in net income) negatively predicts future growth. I also discover discrepancies in analysts' subjective outlooks for a firm's cash. Regressions using subjective outlooks show that analysts believe cash will be low if a firm has higher or increasing debt, expenses, and assets, whereas these relationships do not seem to exist in the realized data. These results provide suggestive evidence of the mistakes analysts make when forming their subjective views.

Finally, I investigate the drivers of analyst disagreement. Conceptually, disagreement can arise from either asymmetric information or different interpretations and outlooks. I capture these two channels by examining the (dis)similarity of factual information and subjective outlooks between two analyst reports whose authors forecast earnings for the same firm on the same date. I find that pairwise differences in short-term EPS forecasts are greater when there is more dissimilarity in both factual information and subjective content, suggesting that both the information and belief channels 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. (2024)), learning with fading memory (Nagel and Xu (2022)), noisy cognitive defaults (De Silva and Thesmar (2024)), and many 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 directly 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 their subjective outlooks.

The findings in this paper connect to various aspects of subjective beliefs. First, it adds 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 mis-reaction in analysts' forecasts. The granular information signals extracted from analyst reports provide a new avenue to examine what information analysts mis-react to, an aspect that traditional tests like CG-regression cannot address. 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.

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 results of embedding-based prediction that demonstrate the informativeness of the textual content in

[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.

analyst reports. Section 4 details the implementation of information extraction using RAG. Section 5 studies the variation in analyst attention. Section 6 studies how analysts mis-react to different types of information. Section 7 studies analysts' subjective outlooks and the drivers of disagreement. 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 conflict of interests 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]

I obtain sell-side equity research analyst reports from Investext, an aggregator provided by

[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 lead to fines or suspensions of analyst certification, including cases involving major investment banks like Deutsche Bank and Barclays. Andrew J. Ceresney, former SEC's 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 an important part of their research process for generating trade ideas.

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 and 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). The matched sample includes 2,217,615 equity research reports covering 8,597 stocks, written by 10,766 sell-side analysts from 543 brokerage firms, including 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.[8] Summary statistics for the sample can be found in A.4.1.

Finally, I match analyst reports to the announcements in the IBES Detail File. Specifically, for each announcement 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.5 gives more details about the matching process. As a result, I am able to match 777,625 EPS announcements made by 119 institutions with 1,330,92 analyst reports.[9]

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.

[8] Reports from some major investment banks–such as Morgan Stanley and UBS–are not available through Investext. I am currently in the process of obtaining those reports from Refinitive LSEG.

[9] 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 Estimated 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) and their long-term EPS growth forecasts. I also incorporate analysts' 12-month ahead price targets (PTG). I obtain the consensus forecasts from the IBES Summary file and individual forecasts from the IBES Detail file. The sample period spans from 01/01/1998 to 11/30/2023, which aligns with the sample period for the Investext analyst report data.

Other data used in this paper include stock price and return data from the Center for Research in Security Prices (CRSP), as well as 153 stock-level characteristics from Jensen et al. (2023). These characteristics cover a broad spectrum of stock-level variables that have been studied in the asset pricing 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. Existing 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.[10] 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-

[10] Examples of textual analyses that uses 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), Ke et al. (2024), among others. See Gentzkow et al. (2019) for a recent review for these traditional approaches in textual analysis.

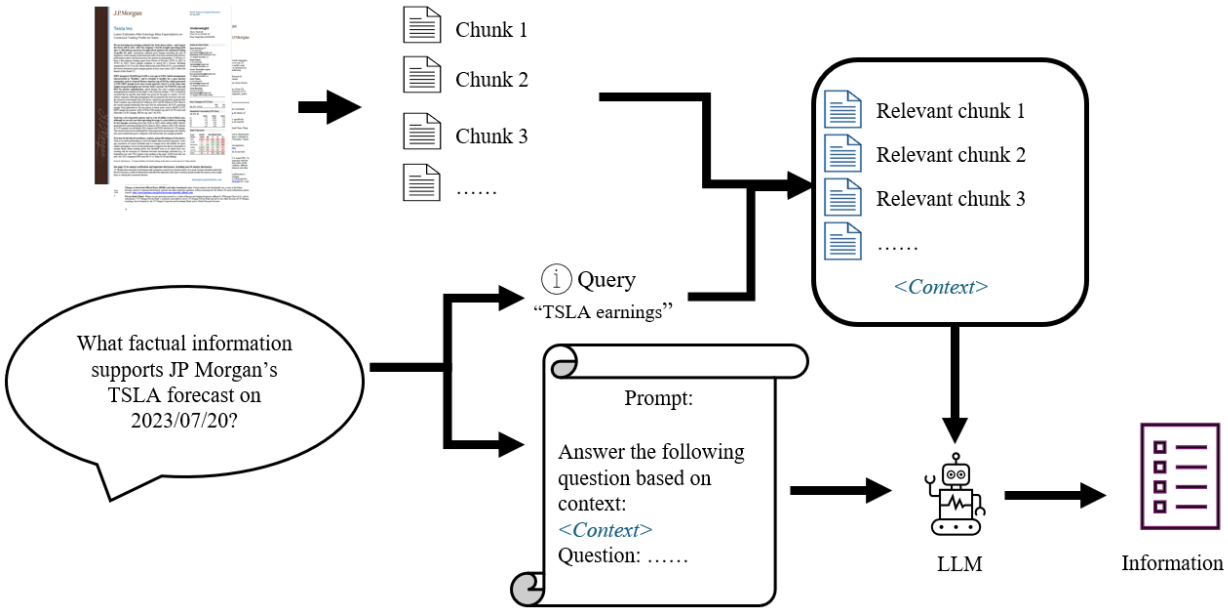


Figure 1: Illustration of the Retrieval-Augmented Generation (RAG) framework.

based NLP tasks, including text summarization, zero-shot classification, translation, and question answering, reaching performance levels comparable to human benchmarks.[11]

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 1 illustrates the RAG workflow.

In this paper, I use `Llama-3-8b-Instruct` as the generator in RAG. `Llama-3-8b-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,[12] while the open-source nature of `Llama-3-8b-Instruct` allows me to flexibly deploy the model locally, run inferences without limits, and gain full control of the generation process. In Section 4 and Appendix C, I provide further details on the RAG implementation, including the retrieval technique and the various prompts used.

[11] See Brown et al. (2020), Wei et al. (2022) for surveys of recent developments in large language models.
[12] See https://ai.meta.com/blog/meta-llama-3/ for description of `Llama-3-8b-Instruct`.

Remarks on LLM Implementation

While LLMs have shown astonishing capabilities comparing to traditional textual analysis method and researchers have become more open to adopt LLMs in their research pipelines, many others have raised concerns regarding the validity of LLM outputs. In this section I briefly address some of the concerns.

One major concern of using LLM is that, given the stochastic nature in 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 an inputs. 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) is 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)). This, however, does not apply to open-source LLM models like `Llama3` used in this paper, for which the researcher has full control of the model weights.

Finally, since most LLMs are trained on mass corpus available on the internet, researchers raise concerns of look-ahead bias (e.g. Glasserman and Lin (2023), Sarkar and Vafa (2024)). However, this concern is unlikely to apply to analyst reports, since these reports are not freely available from the internet and therefore not in the training corpus of LLMs. I also note that the text-generation capability of the LLM is mostly used as information extraction and summarization, rather than performing out-of-sample prediction.

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 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-3-8b-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.[13] In `Llama-3-8b-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.

3.1 Analyst forecast errors are predictable by their own 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.[14]

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

$$Error^h_{i,j,t} = \theta_h x_{i,j,t} + \epsilon_{i,j,t} \tag{1}$$

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. (2024).

The parameters θ in (1) 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 (1), which is optimally tuned using 5-fold cross-validation in the training

[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-3-8b-Instruct`.

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

data.[15] I evaluate the model predictions based on their out-of-sample R^2, defined as

$$R^2 = 1 - \frac{\sum_{j,t \in OOS}(Error^h_{i,j,t} - \widehat{\theta}_h x_{i,j,t})^2}{\sum_{j,t \in OOS}(Error^h_{i,j,t})^2}. \tag{3}$$

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 1 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.

	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 1: Out-of-sample performance of individual forecast error prediction.

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 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

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

$$\min_{\theta} ||Error^h_{i,j,t} - \theta_h x_{i,j,t}||_2 + \lambda ||\theta_h||_2 \tag{2}$$

where λ is the hyperparameter controlling the ridge shrinkage.

Revision to have the same standard deviation as text-based prediction.

Table 2 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.

	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 2: 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.

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 (1) 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.

3.2 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)); under-reaction in short-term forecasts (Bouchaud et al. (2019)); and over-reaction in long-term forecasts (Bordalo et al. (2024)). 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

$$E_t^{i,\text{text-adj.}}[\pi_{j,t+h}] = E_t^i[\pi_{j,t+h}] + \widehat{Error}^h_{i,j,t} \tag{4}$$

where $E_t^i[\pi_{j,t+h}]$ is the original forecast from IBES, and $\widehat{Error}^h_{i,j,t}$ 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 3 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.

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

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;

$$Error^h_{i,j,t} = \alpha + \beta \cdot Revision^h_{i,j,t} + \epsilon_{j,t}, \tag{5}$$

where revision is defined as the change in the forecast between two periods: $Revision^h_{i,j,t} = E_t^i[\pi_{j,t+h}] - E_{t-1}^i[\pi_{j,t+h}]$. 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 4 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.[16] Panel B in Table 4 shows the CG-regression results if text-adjusted forecasts and revisions are used instead. A striking observation is that the behavioral patterns docu-

[16]I focus on the forecasts that can be matched with research reports. Table 23 in Appendix B.2 shows that the regression results are qualitatively the same if I run the regression using all IBES forecasts, thus confirming the representativeness of my sample. Outliers below 2.5 and above 97.5 percentiles are dropped.

mented 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 4: 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.

Remark on institutional incentive distortion While this paper proposes misreaction to qualitative information as the explanation for the results in Sections 3.1 and 3.2, an alternative explanation is that analysts' numerical forecasts are distorted by institutional incentives in the equity research industry.[17] 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 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.

[17]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)).

3.3 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;

$$\pi_{j,t+h} = \theta \overline{x}_{j,t} + \epsilon_{j,t+h}. \tag{6}$$

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 5 reports the out-of-sample R^2 of realized earnings prediction using average embeddings.[18] 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 5 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 5: Out-of-sample R^2 of future earnings prediction for embedding-based prediction, IBES consensus forecast prediction, and JKP characteristics prediction.

[18] Similar to the R^2 formula in (3), 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.

How does the predictive power of embedding-based forecasts compare with the predictive power 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 5 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.[19] 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.[20] 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:

$$\pi_{j,t+h} = \beta_{\text{text}}\widehat{\pi}^{\text{text}}_{j,t+h} + \beta_{\text{IBES}}E^{\text{IBES}}_t[\pi_{j,t+h}] + \beta_{\text{JKP}}\widehat{\pi}^{\text{JKP}}_{j,t+h} + \epsilon_{j,t+h}. \tag{7}$$

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.[21] Table 6 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.

3.4 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

[19] The IBES consensus forecast is obtained as the median estimate from `statsum_epsus` table. The characteristics-based prediction model is the same as (6); 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.

[20] 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.

[21] 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 B.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 6: Regression result for (7). Numbers in parentheses are t-statistics based on bootstrapped standard errors of parameter estimates to account for sampling error in the generated regressor.

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 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[22]

$$\widetilde{E}_t[r_{i,t+12}] = \frac{\texttt{Price Target}_{i,t+1} - \texttt{Current Price}_{i,t}}{\texttt{Current Price}_{i,t}}. \tag{8}$$

Figure 2 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 7 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

[22]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.

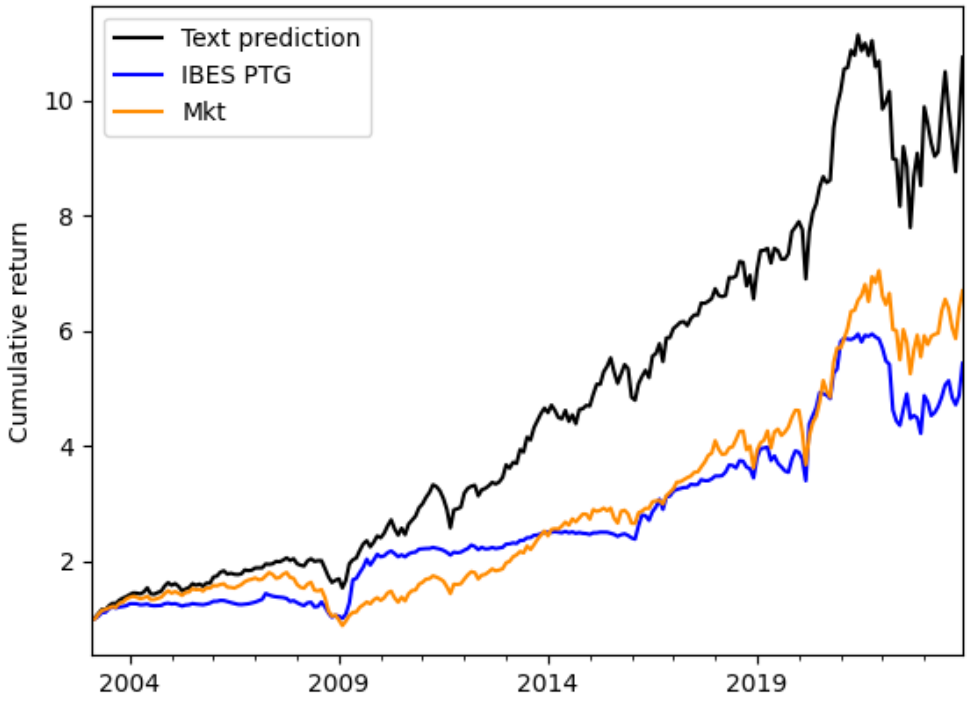


Figure 2: This figure shows the cumulative return of a trading strategy that uses an embedding-based prediction model ("Text pred") and the implied expected return based on consensus price targets from IBES ("IBES PTG-implied return"), as well as the market portfolio ("Mkt"). Each strategy is based on an equal-weighted decile long-short spread and is scaled based on ex-post volatility to have an annualized volatility of 10%.

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.[23] 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.

4 Information Extraction from Analyst Reports

Section 3 shows that the textual content written by analysts is informative about their own forecast errors, and that it contains information about future realized earnings and returns. But what exactly do analysts write about in their reports? This question cannot

[23] Another noticeable finding is that the text-based long-short strategy has large significant loadings on the 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)
R_MKT		0.94***	0.80***	0.77***	0.84***	0.80***
		(12.14)	(14.49)	(14.57)	(7.95)	(9.05)
FF_SMB			0.64***	0.59***		
			(11.38)	(9.86)		
FF_HML			-0.17**	-0.15*		
			(-2.24)	(-1.91)		
FF_RMW				-0.22***		
				(-2.73)		
FF_CMA				-0.06		
				(-0.53)		
HXZ_IA					-0.01	
					(-0.08)	
HXZ_ROE					-0.35***	
					(-2.47)	
SY_PERF						-0.28***
						(-3.11)
SY_MGMT						-0.53***
						(6.75)

Table 7: 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.

be answered by just looking at the embeddings of text. Although the embedding encodes the overall context, it transforms text into a numerical vector, making it hard to interpret and study.

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.

Before diving into the details of information extraction, Figure 3 provides a conceptual framework for the textual content in analyst reports that guides my information extraction later. From the reports, researchers can discern the factual information the analysts pay attention to. This factual content includes statements about objective facts, such as a firm's recent earnings or balance sheet information, and reveals the information set that analysts focus on. Researchers can also extract analysts' subjective outlook on a firm's future performance—such as their narratives about the firm's future margin—which form the basis for their numerical earnings forecasts. Finally, analysts derive their price target and recommendation using a specific valuation model, which is also stated in the written reports. As the focus of this paper is on analyst earnings forecasts, I use an LLM to extract factual information and analysts' outlook about firms' fundamentals that matter for future earnings, while not focusing on the valuation methods in the reports.[24]

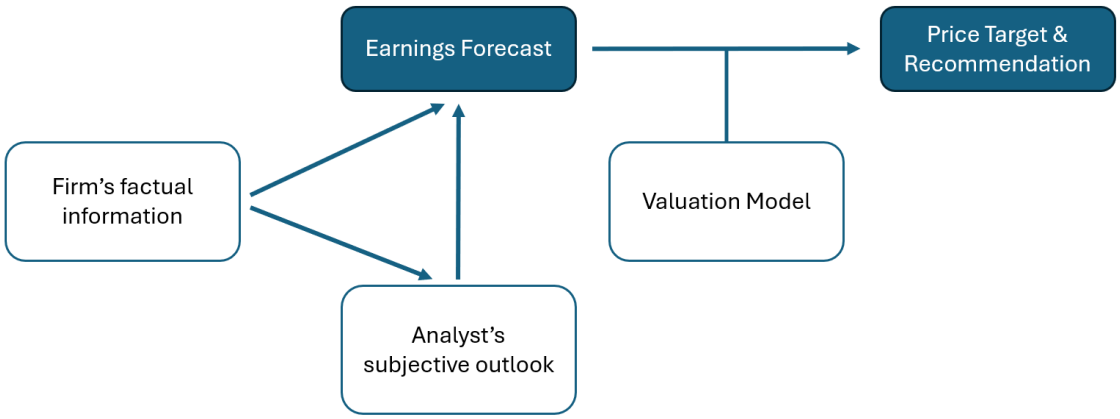


Figure 3: Conceptual framework for the textual content in sell-side analyst reports

[24]See Decaire and Guenzel (2023) and Decaire and Graham (2024) for studies of the valuation models used by analysts.

4.1 RAG Implementation

As introduced in Section 2.2, RAG is a framework that combines generative AI with retrieved relevant information. In this section, I describe the RAG implementation for information extraction. The procedure can be summarized in 4 steps: 1) retrieve relevant context, 2) identify correct stock, 3) separate factual versus subjective content, and 4) extract financial concepts with their associated adjectives.

4.1.1 Relevant Context Retrieval

The first step 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 use each paragraph in the report as a chunk. I further split each paragraph into chunks with at most 128 tokens, if the paragraph's length exceeds 128 tokens. This step results in the context chunk set, from which retrieval will be performed.

For each context chunk n, I generate a 768-dimensional embedding x_n using `FinBERT`.[25] I then generate another embedding y for the query "{TICKER} EPS forecast", where {TICKER} is the stock ticker being studied. I extract the relevant context chunks as the 10 most similar context chunks with the query, measured by cosine distance of the context embedding with the query embedding:

$$\text{Relevant Context Set} = \{n \text{ for } n \in \text{ Top 10 of } \{\frac{x_n \cdot y}{||x_n||||y||}\}\} \tag{9}$$

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

4.2 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 for better extraction and generation quality. The steps are described below.

Step 0: Remove Boilerplate Content

Before the actual information extraction, I first use LLM to remove the boilerplate contents. These boilerplate contents are usually for disclosure purposes and only contains information

[25] I don't use `Llama-3` embedding for runtime consideration. Manual inspection shows that the retrieval results using `FinBERT` embedding are similar to those when using `Llama-3` embedding.

about analyst certificate, brokerage firm's relationship with the public company, rating systems used and so forth. Figure 15 in Appendix C describes the prompt used. I remove the boilerplate segments of the reports for which the LLM identifies with self-reported confidence greater than 90. This step turns out to be effective upon manual inspection.

Step 1: Identify Stock Information

After boilerplate contents are removed, the LLM is asked to identify the stock in the context, and only extract information if the stock can be correctly identified. This step is important to guarantee that the information extracted is specifically about the target stock, especially there could be multiple stocks mentioned in the content. Figure 16 in Appendix C shows the exact prompt used for identifying stock information.

As a proof of validation, figure 4 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.[26] The "Actual" column shows the main results. 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.

In order to show that LLM doesn't hallucinate, I also conduct two placebo tests. 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 4, 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 shows that the identifying step makes LLM correctly report "I don't know" when the information is in fact not present in the context.

Step 2: Separate Factual vs. Subjective Content

After validating the efficacy of context retrieval, I use LLM to separately summarize the factual information and analyst's subjective statements presented in the context. The sell-

[26]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.

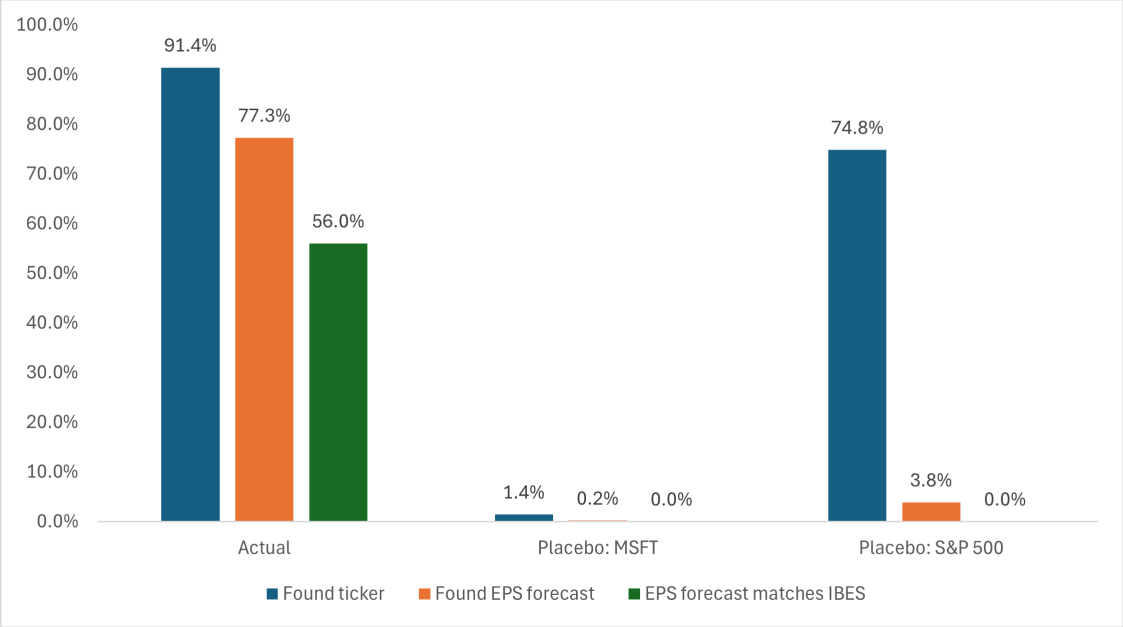


Figure 4: RAG retrieval validity of finding stock EPS information.

side 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 states his own subjective opinion about the firm's future outlook. As a concrete 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 in the report, while the second sentence is a subjective statement made by the analyst that reflects his 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 naturally maps to the analyst's narrative that reflects his belief about the future.

I leverage 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 17 in Appendix C describes the actual 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. I define narrative as a tuple of {"Noun": "Description"}, where "Noun" is a noun or a noun-phrase for an business or economic related concept or object, and "Description" is an adjective or short phrase that describes the "Noun".[27] 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. Figure 18 shows the actual prompt used.

Table 8 reports an example of the factual versus subjective information 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 D reports more examples of LLM responses, which further validate that the LLM gives satisfactory results that separate factual information from subjective statements.

4.3 Narrative Classification

The LLM inference step extracts 8,456,795 occurrences of 401,715 unique financial concepts and 5,528,993 occurrences of 228,326 adjectives. I focus on the 1500 most frequent financial concepts (covers 73% of total occurrences) and 500 most frequent adjectives (covers 72% of total occurrences). To further distill the information for better interpretability, I categorize financial concepts and adjectives using zero-shot classification. 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 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` based on BART, which is a transformer model with a bidirectional (BERT-like) encoder and an autoregressive (GPT-like) decoder developed by Meta AI.[28]

[27] 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.

[28] It is the most downloaded zero-shot classification model on Huggingface.

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). 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)."

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." "

Financial economic concepts extracted by LLM: factual content:

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

Financial economic concepts extracted by LLM: subjective content:

- "SG&A": "difficulty"

Table 8: Example of factual and subjective content summarization

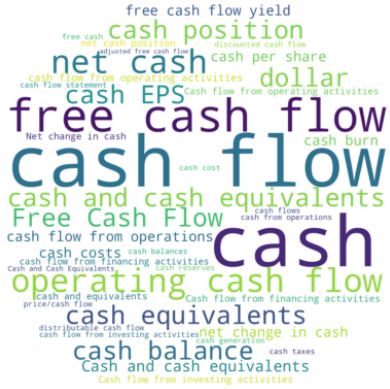
Specifically, for each financial concept, I use the zero-shot classifier to predict the probability that this concept belongs to one of the nine common labels that I manually specify: "Asset", "Cash", "Debt", "Expense", "Earnings", "Equity", "Growth", "Margin", and "Revenue", as well as a null class that is labeled by an empty string ("Others"). Each phrase

is assigned to the class based on the highest predicted probability. Out of the 1,500 concepts, 1,127 can be classified into one of the 9 topics, covering 62.5% of total occurrences. Figure 5 plots the word cloud for each topic.

Similarly, for each adjective, I use the classifier to predict the probability that it belongs to either a "positive" class, a "negative" class, or a "neutral" class labeled by an empty string. Figure 6 shows the word cloud for positive, negative, and neutral adjectives.



(a) Assets



(b) Cash



(c) Expense



(d) Debt



(e) Earnings



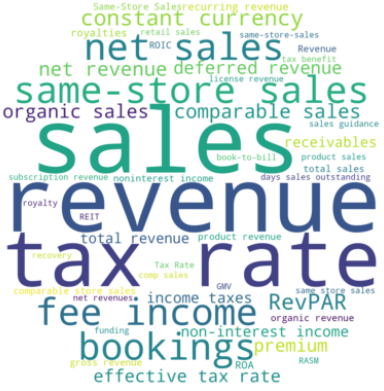
(f) Equity



(g) Growth



(h) Margin



(i) Revenue

Figure 5: Word cloud of financial concepts. The size of a phrase is proportional to the phrase's frequency multiplied by its probability in a given class.



(a) Positive (b) Negative (c) Neutral

Figure 6: Word cloud of adjectives. The size of a phrase is proportional to the phrase's frequency multiplied by its probability in a given class.

5 An Anatomy of Analysts' Attention

The factual information and subjective content extracted by the LLM in Section 4 provides details that justify an analyst's forecasts. This provides a novel avenue for studying the subjective belief formation process. In this section, I study the attention the analyst pays to different types of information using the occurrence of financial concepts in their written reports.

Specifically, I measure analyst i's attention to topic k when making an EPS forecast announcement for firm j at time t as the fraction of financial concepts in the factual content in the analyst report that belongs to topic k:

$$Attn^k_{i,j,t} = \frac{\sum_{c \in A^{factual}_{i,j,t}} \mathbf{I}(c \in C^k)}{\sum_{c \in A^{factual}_{i,j,t}} \mathbf{I}}, \tag{10}$$

where $c \in A^{factual}_{i,j,t}$ is a financial concept mentioned in the concept list in the factual content $A^{factual}_{i,j,t}$ written by analyst i for firm j at time t, $\mathbf{I}$ is an indicator function and $\mathbf{I}(c \in C^k)$ denotes the indicator function that concept c belongs to the topic of k. Note that with the presence of the "other" class, the attention doesn't sum up to 1 across the nine financial topics.

Figure 7 shows the full sample average attention to different topics in factual content. Analysts pay most attention to a firm's earnings and revenue when writing the reports, and also pay attention to a firm's margin and expenses. On the other hand, analysts pay less attention to a firm's balance sheet information, including its assets, debt and cash.

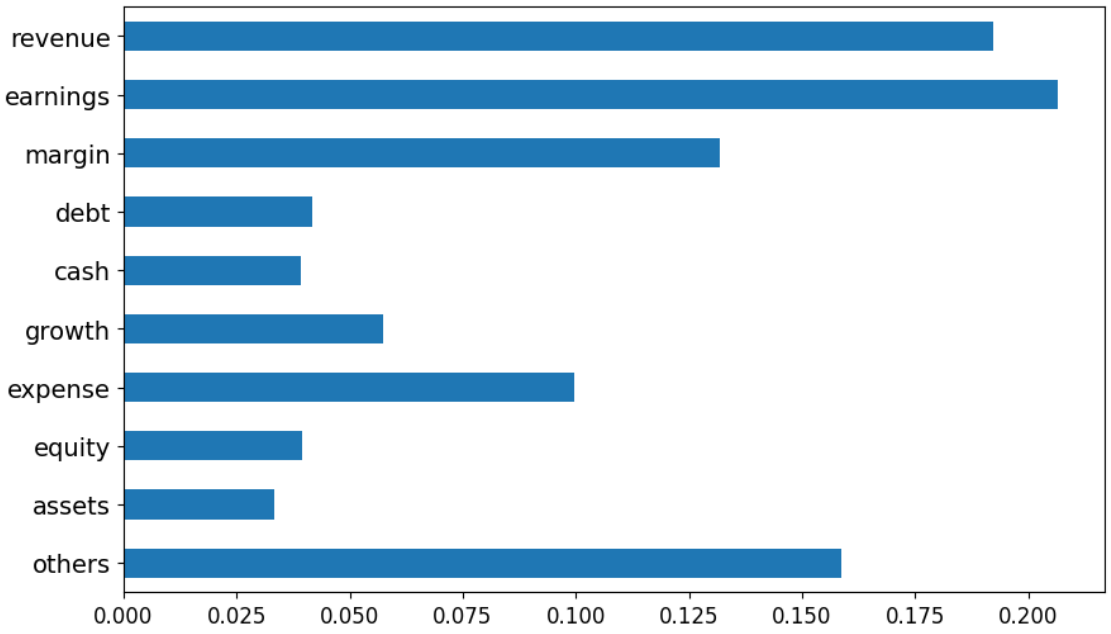


Figure 7: Full sample average of analyst attention to different factual information.

5.1 Time variation in attention to firm-level information

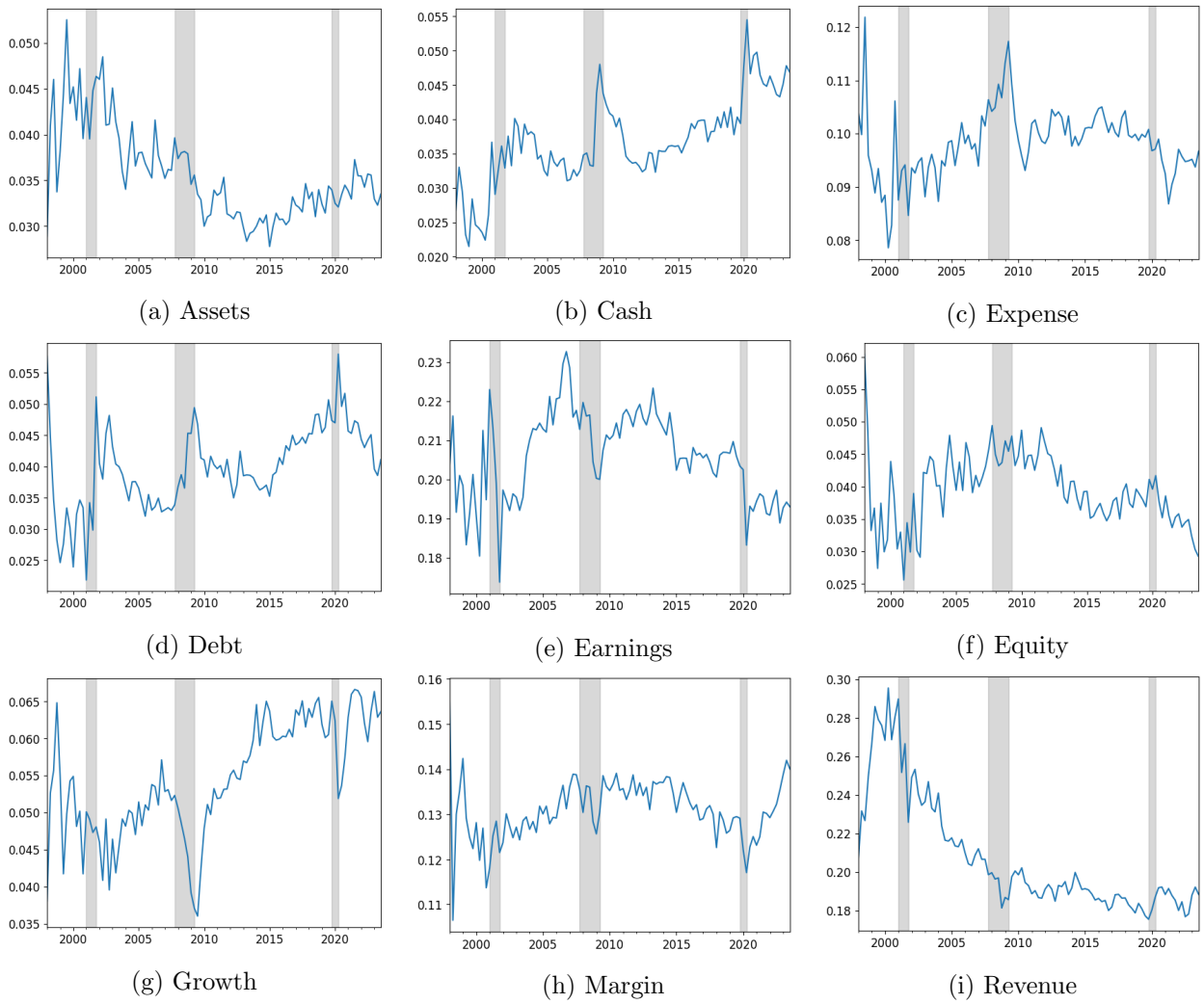
Analysts' attention exhibits significant variation over time and over business cycles. I construct an aggregate measure of attention to topic k information as simply the average $Attn^k_{i,j,t}$ measure across analysts and firms at time t. Figure 8 plots analysts' attention to different topics of factual information over time. Comparing to the 2000s, analysts pay more attention to information regarding a firm's cash, debt, and growth, while reducing their focus on revenues and assets. One prominent pattern in analyst's attention is how it changes during recessions. During a recession, analysts pay more attention to a firm's debt and cash, while reducing their attention to revenue, earnings and growth. This result highlights that financial stress becomes the dominant consideration for analysts during a recession.

More formally, I regress the aggregate attention to topic k information in a quarter on log industrial production, the 10-Year minus 2-Year Treasury spread, the VIX, and a trend indicator:[29]

$$\overline{Attn}^k_t = \alpha Trend_t + \beta_1 IP_t + \beta_2 Spread_t + \beta_3 VIX_t + \epsilon_t. \tag{11}$$

The results, in Table 9, show that analysts' attention to a firm's margin, debt, cash, growth and expenses has increased over time, while their attention to revenue and assets has decreased. Analysts' attention to a firm's revenue, earnings and assets is pro-cyclical, while attention to debt and cash are counter-cyclical (as captured by industrial production). At the same time, analysts pay more attention to a firm's equity while reducing attentions to

[29]Data are obtained from FRED.



(a) Assets (b) Cash (c) Expense

(d) Debt (e) Earnings (f) Equity

(g) Growth (h) Margin (i) Revenue

Figure 8: Analysts' quarterly average attention to different topics. Shaded region denotes NBER recessions.

revenue, growth and assets when financial conditions are loose (as captured by the high term spread). Finally, when risk in financial market increases, analysts pay more attention to a firm's assets, debt, and cash, while reducing their attention to earnings, growth and margin. Overall, these results show that analysts' attention to information related to firms/ earnings is pro-cyclical, while attention to information related to firms' financial condition is counter-cyclical.

	Revenue	Earnings	Margin	Debt	Cash	Growth	Expense	Equity	Assets
Trend	-2.91***	-0.83*	0.34	2.12***	2.48***	1.67***	0.44	-0.51	-2.23***
	(-6.81)	(-1.65)	(0.61)	(4.87)	(6.56)	(4.59)	(0.73)	(-0.87)	(-5.36)
IP	0.35***	0.28**	0.05	-0.38***	-0.50***	-0.03	-0.14	-0.06	0.18***
	(3.40)	(2.15)	(0.50)	(-5.38)	(-6.82)	(-0.67)	(-1.16)	(-0.45)	(2.64)
Spread	-0.23**	0.26*	0.20	0.07	0.19*	-0.23***	0.2	0.35***	-0.21***
	(-2.19)	(1.78)	(1.33)	(0.61)	(1.91)	(-2.49)	(1.30)	(2.60)	(-2.36)
VIX	0.01	-0.06***	-0.03**	0.03***	0.04***	-0.02***	0.01	0.01	0.03***
	(0.41)	(-2.98)	(-2.09)	(3.22)	(4.01)	(-2.90)	(0.54)	(0.50)	(2.84)
Adj-R^2	0.71	0.21	0.05	0.33	0.50	0.37	0.02	0.13	0.48

Table 9: Regression results of regressing quarterly topic attention on macroeconomic variables as in (11). The attention is 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. Standard errors are HAC-adjusted with 4 lags. *, **, *** denote significance level at 10%, 5% and 1%.

5.2 Attention in short-term 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 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 the attention to different components of firms' fundamentals when an analyst makes an announcement to IBES that only updates his FY1 forecast (short-term forecast), versus the analyst's attention when he makes an announcement to IBES that only updates his FY3-FY5 or LTG forecast (long-term forecast). Figure 9 shows the difference between average long-term forecast attention and short-term forecast attention. When analysts make long-term forecasts, they pay less attention to a firm's revenue, expense, and margin, and pay more attention to a firm's past growth as well as "other" information not belonging to the nine financial information.

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:

$$Attn^k_{i,j,t} = \delta \cdot \mathbf{1}\{\text{LT Forecast}\} + \eta_j + \eta_i + \epsilon_{i,j,t}. \tag{12}$$

Here, $\widehat{\delta}$ can be interpreted as the difference of attention to topic k in short-term versus long-term forecasts. Table 10 reports the regression results. I find that analysts decrease their

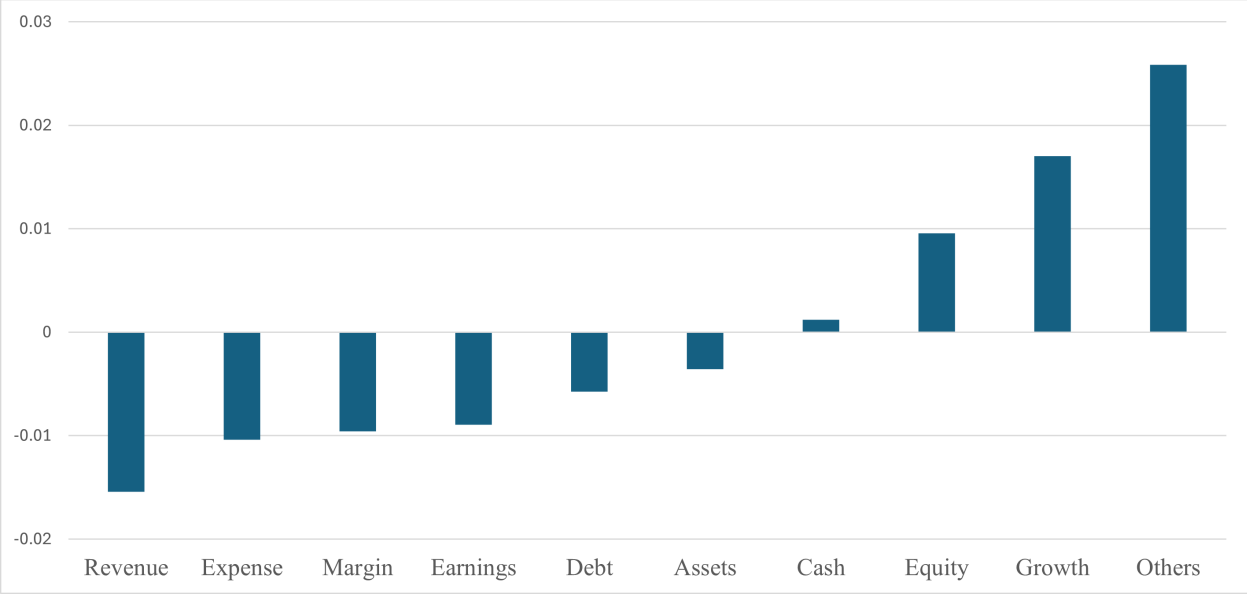


Figure 9: Differences between average long-term forecast attention and average short-term forecast attention.

attention to a firm's revenue, earnings, and margin when making a short-term forecast. The estimated $\widehat{\delta}$'s for these topics are negative and significant, translating to a -7.3% to -8.6% reduction in attention compared to average attention to those topics. On the other hand, analysts significantly increase their attention to the growth, equity and "others" topics when making long-term forecasts, with the magnitudes of attention increase ranging from 14% to 27% relative to the sample average.

	Revenue	Earnings	Margin	Debt	Cash	Growth	Expense	Equity	Assets	Others
$\widehat{\delta}$	-0.017	-0.015	-0.011	0.004	0.000	0.008	-0.006	0.011	0.003	0.023
	(-8.34)	(-6.73)	(-6.76)	(3.24)	(0.53)	(5.61)	(-3.74)	(8.31)	(3.04)	(10.07)
$\widehat{\delta}/\overline{Attn}$	-8.6%	-7.3%	-8.3%	8.5%	1.3%	14.1%	-5.8%	27.0%	9.1%	14.5%

Table 10: Regression result of (12). The third row reports the estimated $\widehat{\delta}$ as a percentage of average attention to each topic.

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' standard accounting information (particularly related to the income statement) while paying more attention to growth and other non-standard information. This result echoes Daniel and Titman (2006) who argue that reaction to intangible information drives stock's long-term return reversal.

5.3 Cross-sectional variation in attention to firm-level information

In addition to the aggregate time-series variation, I also find that there is substantial variation in analysts' attention across stocks. In the first exercise, 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:

$$\overline{Attn}^{k}_{j,t} = \alpha + \beta \mathbf{I}\{j \in G_s\} + \epsilon_{j,t}, \tag{13}$$

where $\overline{Attn}^{k}_{j,t}$ 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. Standard errors are double-clustered by firm and quarter.

	Public Admin	Manufacturing	Finance	Wholesale	Service	Transportation	Retail	Mining	Agriculture	Construction
Revenue	***	***	***		***	***	***	***		
Earnings	***		***	***		***		***		
Margin	***	***	***	***	***	***	***	***		***
Debt		***	***	***	***		***	***	***	***
Cash	***	***	***	***	***		***	***	***	***
Growth	***	***	***		***		***	***		***
Expense	***	***	***		***	***		***		***
Equity		***	***	***	***		***			
Assets		***	***	***	***	***			***	***
Others	***	***	***	***	***	***	***			

Figure 10: Differences in analyst attention to topics across industries. A stronger red (green) color denotes a more negative (positive) β estimate from regression (13). *** denotes significance at 1% level.

Figure 10 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 analyzing finance firms, analysts pay more attention to information regarding the firm's debt and assets and pay less attention to the firm's revenue. This makes sense as financial firms operate using financial assets and debt and are usually highly levered. On the other hand, when analyzing firms in the service and manufacturing industries, information about revenue becomes more important. For the retail industry, analysts pay more attention to information related to the margin. Overall, these results suggests that industry is an important determinant in analysts' attention to different types of information.

In addition to industry, I explore how analysts pay attention to different topics for different kinds of stocks. Specifically, I run a regression of attention on a set of stock characteristics, including size (market cap), value (book-to-market ratio), profitability (operating profit-to-book ratio), investment (change in total asset), debt level (total debt to market

	Revenue	Earnings	Margin	Debt	Cash	Growth	Expense	Equity	Assets
ME	0.03	0.33***	0.04	-0.27***	-0.52***	0.36***	-0.31***	0.13**	-0.14**
	(0.54)	(5.57)	(0.65)	(-4.78)	(-6.47)	(5.23)	(-4.91)	(2.36)	(-2.19)
BE/ME	-0.04	0.03	-0.04	-0.05*	0.02	-0.18***	0.23***	0.03	-0.18***
	(-1.36)	(1.03)	(-1.23)	(-1.85)	(0.58)	(-5.99)	(7.35)	(1.03)	(-6.03)
OP/BE	0.04	0.1***	0.03	-0.02	-0.05*	0.01	-0.12***	0.07***	-0.07***
	(1.38)	(3.77)	(1.00)	(-1.01)	(-1.81)	(0.27)	(-3.99)	(2.39)	(-3.20)
Δ Asset	-0.01	0.00	-0.01	-0.00	-0.03*	0.11***	-0.08***	0.02	0.02
	(-0.49)	(0.17)	(-1.09)	(-0.20)	(-1.68)	(6.25)	(-4.80)	(1.04)	(1.00)
Debt/ME	-0.02	0.10***	0.08***	0.24***	-0.13***	-0.11***	-0.02	-0.08**	0.10***
	(-0.54)	(2.99)	(2.48)	(7.49)	(-3.14)	(-3.40)	(-0.50)	(-2.02)	(2.68)
σ(OCF/Sale)	0.09***	-0.12***	-0.09***	0.08***	0.10***	-0.05	0.05	-0.0	-0.15***
	(3.31)	(-3.53)	(-3.04)	(2.54)	(3.53)	(-1.61)	(1.50)	(-0.15)	(-4.06)
Time FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Stock FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Total R^2	0.46	0.27	0.22	0.46	0.35	0.21	0.20	0.12	0.27

Table 11: Regression results of regressing quarterly topic attention on selected stock characteristics 14. 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 two-way clustered at firm and time level. *, **, *** denotes significance level at 10%, 5% and 1%.

cap, following Bhandari (1988)), and cash flow volatility (standard deviation of operating cash flow to sales, following Huang (2009)), together with time and stock fixed effects:

$$\begin{aligned}\overline{Attn}^k_{j,t} = \alpha + \beta_1 ME_{j,t} + \beta_2 BE/ME_{j,t} + \beta_3 OP/BE_{j,t} + \beta_4 \Delta Asset_{j,t} \\ + \beta_5 Debt/ME_{j,t} + \beta_6 \sigma(OCF/Sale)_{j,t} + \eta_t + \xi_i + \epsilon_{j,t}.\end{aligned} \tag{14}$$

Table 11 reports the results. It shows that analysts' attention to different topics differs significantly across different types of stocks. Specifically, analysts pay more attention to information related to a firm's earnings and growth for large-cap stocks, but more attention to debt, cash, and expense information for small-cap stocks. For value stocks (with high BE/ME), analysts pay more attention to expense information and for growth stocks (with low BE/ME), more attention to growth and assets. For profitable firms (with high OP/BE), analysts pay more attention to earnings and equity, but less attention to expense and assets. For firms with high investment rate (with high Δ Asset), analysts care more about growth. For firms with high debt level, analysts pay more attention to information regarding debt, earnings, assets and margin, while paying less attention to information regarding cash and growth. Finally, for firms with higher cash-flow volatility, analysts pay more attention to debt, cash, revenue while paying less attention to assets, earnings, and margin.

5.4 Cross-analyst variation in attention to firm-level information

Finally, I investigate the cross-analyst variation in attention to firm-level information. According to theory of optimal attention choice (e.g. Kohlhas and Walther (2021)), 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.

Specifically, I estimate the following fixed-effect model for each topic k:

$$Attn^k_{i,j,t} = \gamma_{j\times t} + \epsilon_{i,j,t} \tag{15}$$

$$Attn^k_{i,j,t} = \tau_{i,j,t} + \gamma_{j\times t} + \epsilon_{i,j,t}, \tag{16}$$

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 (16) should not help to capture additional variation in $Attn^k_{i,j,t}$, and that the model fits of regression (15) and (16) should be the same.

I first run regressions (15) and (16), where I use an analyst fixed effect in (16), i.e. $\tau_{i,j,t} = \mathbf{I}_i$. I then compute the R^2 of regression (16) relative to regression (15):

$$\text{Relative } R^2 = 1 - \frac{MSE(\text{regression (16)})}{MSE(\text{regression (15)})} \tag{17}$$

If analysts have homogeneous attention, then the relative R^2 should be 0. Table 12 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 12: Relative R^2 as calculated from formula (17). *** 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:

$$Attn^k_{i,j,t} = \beta_1 Exp_{i,t} + \beta_2 ExpFirm_{i,j,t} + \beta_3 ExpInd_{i,j,t} + \beta_4 FirmCover_{i,t} + \beta_5 IndCover_{i,j,t} + \gamma_{j\times t} + \tau_i + \epsilon_{i,j,t}. \tag{18}$$

Table 13 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 13: Regression results of regressing topic attention on analyst experience and coverage (18). 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.

6 Mis-reaction to Factual Information

Previous literature that studies individual-level analyst forecasts typically finds under-reaction at the short-horizon (Bouchaud et al. (2019)) and over-reaction at the long-horizon (Bordalo et al. (2024)).[30] However, an important open question still remains: What information do analysts under- or over-react to? To investigate this question, one essentially needs to condition on the information the analyst observes. The factual content extracted from analyst reports provides a good characterization of the information set the analyst observes, thus providing an opportunity to assess the origins of this mis-reaction to information.

I first use the adjectives associated with the financial concepts to measure the directional information for each fundamental component. I define a variable $Signal^k_{i,j,t}$ as the difference between the number of positive and negative adjective counts associated with financial concepts in topic k in the factual content

$$Signal^k_{i,j,t} = \frac{\text{positive adj. \#} - \text{negative adj. \#}}{\text{positive adj. \#} + \text{negative adj. \#}} \text{ for topic } k \text{ in factual content } i,j,t \quad (19)$$

This *Signal* variable therefore has the literal meaning of changes in fundamental component k that the analyst observes. Because it is constructed based on factual information, a higher $Signal^{debt}_{i,j,t}$ can be interpreted as the analyst i observing firm j having higher/increasing debt at time t.

After I construct the *Signal* variable, I directly regress forecast errors on the signals for the nine topics

$$\pi_{j,t+h} - E_{i,t}[\pi_{j,t+h}] = \sum_k \gamma_k Signal^k_{i,j,t} + \tau_j + \tau_t + \epsilon_{i,j,t} \quad (20)$$

This regression can be viewed as a *granular* version of a CG regression. Without a measure of the information the agent observes, the usual CG regression uses positive (negative) forecast revision as a proxy for good (bad) information. With the detailed factual signals the analyst observes extracted from analyst reports, regression (20) replaces the forecast revision with these more granular signals the analyst observes. Since these signals are extracted from an analyst's own reports, they are clearly within the analyst's information set, and a nonzero coefficient γ_k can be interpreted as mis-reaction to information in fundamental component k.Specifically, a positive (negative) γ_k implies that the analyst's forecast is too low (high) when they receive information that the firm has an increasing k.

Table 14 reports the result for regression (20). I find that the signals to Revenue, Earnings, and Margin have a positive and significant $\widehat{\gamma}$ coefficient for short-term forecasts, suggesting that analysts' forecasts are not high enough when they receive information that

[30] I confirm their finding in my sample in Panel (A) in Table 4 and Appendix B.2

	Forecast Horizon					
	All	1 Year	2 Years	3 Years	4 Years	5 Years
Revenue	0.04***	0.03***	0.05***	0.03	0.01	0.06
	(5.4)	(6.79)	(5.32)	(1.12)	(0.13)	(0.97)
Earnings	0.03***	0.03***	0.04***	-0.0	-0.01	-0.01
	(6.22)	(7.26)	(5.53)	(-0.2)	(-0.16)	(-0.23)
Margin	0.03***	0.03***	0.04***	0.01	-0.01	-0.01
	(6.8)	(7.55)	(5.65)	(0.36)	(-0.24)	(-0.17)
Debt	-0.02*	-0.01***	-0.03***	0.02	0.06	0.03
	(-1.96)	(-2.46)	(-2.61)	(0.81)	(1.14)	(0.33)
Cash	0.01*	0.01**	0.01	0.0	-0.02	0.02
	(1.84)	(2.3)	(0.74)	(0.18)	(-0.33)	(0.34)
Growth	0.0	0.0	0.01	-0.0	-0.01	-0.09
	(0.2)	(1.03)	(0.84)	(-0.17)	(-0.31)	(-1.16)
Expense	-0.01***	-0.0	-0.02***	-0.03*	-0.03	-0.02
	(-2.66)	(-0.56)	(-2.82)	(-1.89)	(-0.84)	(-0.52)
Equity	0.03***	0.02***	0.04***	0.02	-0.01	0.1
	(4.04)	(3.62)	(3.3)	(0.81)	(-0.13)	(0.96)
Assets	-0.03***	-0.01***	-0.04***	-0.03	0.02	0.05
	(-3.39)	(-2.56)	(-3.87)	(-1.19)	(0.25)	(0.36)
Total R^2	0.15	0.14	0.23	0.31	0.47	0.58
Time FE	✓	✓	✓	✓	✓	✓
Stock FE	✓	✓	✓	✓	✓	✓

Table 14: This table shows the regression result of (20). h denotes the number of years between current quarter and forecast target date. Standard errors are clustered at stock and quarter level. *, **, *** denotes significance level at 10%, 5% and 1%.

revenue, earnings, and margin are high/increasing. A related observation is that there is a negative coefficient on Expense, which means analysts' forecasts are not low enough when they receive information that the firm's expense is higher. All of these results are consistent with the idea of sticky belief in firm profitability as the main driver for under-reaction in short-term earnings forecasts (e.g. Bouchaud et al. (2019)). In addition, the negative coefficient on Debt and Assets suggests that under-reaction to news about firm's financing cost can be another important driver of the short-term earnings forecast error. Signals about common accounting information do not seem to drive the mis-reaction in long-term earnings forecasts.[31]

[31]In Appendix E.1, I show that these results are robust to controlling for analysts' attention. When augmenting regression (20) with *Attn*, I also find that analysts are too pessimistic when they pay more attention to a firm's debt and expense, while being too optimistic when they pay more attention to a firm's growth and assets.

7 An Anatomy of Analysts' Subjective Outlook

In Sections 5 and 6, I explored the factual information the analysts pay attention to, and how analysts' mis-reaction to different information translates into forecast error predictability. Analyst reports also contain rich content about analysts' own opinions about various aspects of a firm's fundamentals in the future. These subjective statements about the future provide pathway for studying analysts' belief formation in a more granular fashion. In this section, I explore two questions related to analysts' subjective outlook.

First, I examine how analysts' subjective outlooks on various aspects of a firm's fundamentals are influenced by the information they receive. I find that analysts generally expect trends in each fundamental component to persist. For instance, when analysts observe an increase in a firm's margin as factual information, they often express a subjective outlook that the firm's margin will continue to rise in the future. Additionally, I observe analysts reasoning across different components; for example, they tend to believe that higher debt levels indicate the firm will hold lower cash in the future. Interestingly, some of these relationships in their narratives diverge from predictive regressions using ex-post realizations, suggesting potential errors in analysts' reasoning.

Second, I analyze both factual and subjective content to study the sources of disagreement in analysts' forecasts. Conceptually, disagreement can arise if analysts focus on different information ("information channel") or if they interpret the same information differently, leading to varying opinions and outlooks ("belief channel"). To determine which of these effects drives disagreement in the data, I regress the difference in EPS forecasts on the semantic differences in factual and subjective content——where the former captures differences in analysts' information and the latter captures differences in their opinions. The results indicate that both the information channel and the belief channel significantly contribute to explaining disagreement in short-term forecasts.

7.1 Uncovering Analysts' Reasoning

In this section, I study how analysts use various kinds of information to reason and form a subjective outlook. Similar to the *Signal* variable constructed in (21), I define a variable $Outlook^k_{i,j,t}$ as the difference between the number of positive and negative adjective counts for fundamental component k in the subjective content extracted by the LLM:

$$Outlook^k_{i,j,t} = \frac{\text{positive adj. \#} - \text{negative adj. \#}}{\text{positive adj. \#} + \text{negative adj. \#}} \text{ for topic } k \text{ in subjective content } i,j,t \tag{21}$$

Like *Signal*, *Outlook* carries a literal interpretation. For example, higher $Outlook^{\text{debt}}_{i,j,t}$ can be interpreted as the analyst believing the firm will have higher/increasing debt in the future.

I then directly regress analysts' *Outlook* on *Signal* for different fundamental components

$$Outlook^{k}_{i,j,t} = \sum_{k'} \beta_{k \leftarrow k'} Signal^{k'}_{i,j,t} + \tau_j + \tau_t + \epsilon_{i,j,t} \tag{22}$$

The coefficient $\beta_{k \leftarrow k'}$ captures analyst's reasoning. A positive $\beta_{k \leftarrow k'}$ means the analyst will think that firm's k will be higher in the future when they receive a piece of information that firm's k' is higher today.

Figure 11 depicts $\widehat{\beta}_{k \leftarrow k'}$ estimates. Each column represents a list of $\widehat{\beta}_{k \leftarrow k'}$ for outlooks in a fundamental component. A stronger red (green) color denotes a more negative (positive) $\widehat{\beta}_{k \leftarrow k'}$. Two noticeable patterns stand out from the plot. First, there are strongly positive coefficients along the diagonal, which suggests that analysts tend to think firms' different fundamentals will continue their trends. For example, the positive $\widehat{\beta}$ in the "margin–margin" cell means that analysts tend to believe the firm will have a higher margin in the future when they receive information that firm's margin is higher today. Second, I also uncover analysts' reasoning across different aspects of firms' fundamentals. For example, the negative $\widehat{\beta}$ in the "cash–debt" cell means that analysts tend to believe that the firm will have lower cash in the future if they receive information that the firm's debt is higher today.[32]

		Subjective Outlook								
		Revenue	Earnings	Margin	Debt	Cash	Growth	Expense	Equity	Assets
Factual Info	Revenue	***	***	***		***	***	***	***	***
	Earnings	***	***	***		***	***	***	***	***
	Margin	***	***	***		***	***		***	
	Debt		***	***	***	***		***		***
	Cash		***	***	***	***			***	
	Growth	***	***	***		***	***	***	***	***
	Expense	***			***	***	***	***		***
	Equity	***	***			***	***	***	***	
	Assets	***		***		***		***	***	***

Figure 11: Heatmap of regression coefficient of regressing *Outlook* on *Signal* as in (22). Each column represents one regression (22) for a specific topic. A stronger red (green) color denotes a more negative (positive) $\widehat{\beta}$. *** denotes significance at the 1% level.

Are analysts reasoning correctly? Although it is impossible to make a definite statement about this, to shed some light on this question, I run a regression analogous to (22) using realizations of firm characteristics. Specifically, for each fundamental component k, I predict

[32] In Appendix E, I explore how analyst's attention in subjective content relates to their attention in factual information. I find that when analysts pay more attention to factual information of component k, they also discuss more about k in their subjective statement.

its one-year ahead realization using current fundamental components:

$$X^k_{j,t+1} = \sum_{k'} \beta^X_{k \leftarrow k'} X^{k'}_{j,t} + \tau_j + \tau_t + \epsilon_{j,t}, \tag{23}$$

where $X^k_{j,t}$ denotes the firm characteristic capturing fundamental component k at time t for firm j.[33] Therefore, $\beta^X_{k \leftarrow k'} X^{k'}_{j,t}$ can be interpreted as the realized association between the past fundamental component k' and the future k, which serves as a benchmark for the $\beta_{k \leftarrow k'}$ in (22).

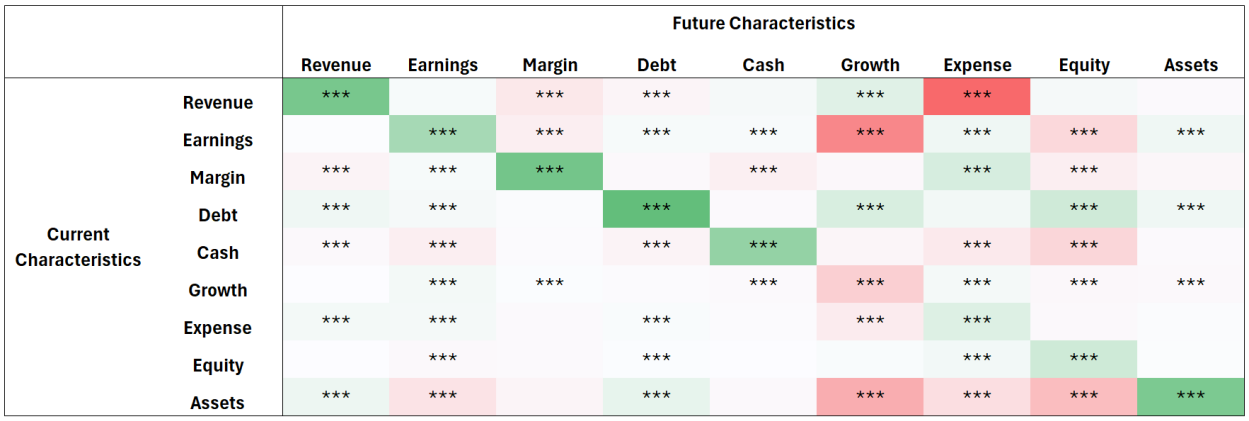


Figure 12: Heatmap of regression coefficient of regressing *Outlook* on *Signal* as in (23). Each column represents one regression (23) for a specific topic. More red (green) cell denotes more negative (positive) $\widehat{\beta}$. *** denotes significance at 1% level.

Figure 12 shows the $\widehat{\beta}^X_{k \leftarrow k'}$ estimates, analogous to Figure 11. A significant pattern is that there are again significant positive coefficients on the diagonal, indicating that these fundamental components do have persistence in their directional changes, and so analysts are right to pick up on this persistence. One noticeable exception is the growth topic: past growth (measured by 1-year change in net income) actually predicts future growth with a negative sign, but analysts make a mistake in thinking growth will continue. I also discover discrepancies in analysts' subjective outlook for firms' cash. Figure 11 shows that analysts believe cash will be low if a firm has higher/increasing debt, expense and assets, whereas these relationships do not appear to exist in the reality.

In summary, in this section I synthesize analysts' reasoning in firm's fundamental components. I find that the subjective outlook of each fundamental component continuing its own

[33] I use sales-to-market for Revenue, operating income-to-market for Earnings, gross profit-to-sales for Margin, debt-to-market for Debt, cash-to-market for Cash, 1-year net income growth for Growth, operating expense growth for Expense, equity net issuance for Equity, and asset-to-market for Assets. All characteristics are obtained from Jensen et al. (2023) and cross-sectionally rank-standardized to be within the [-0.5, 0.5] interval. $\widehat{\beta}^X$ is estimated from 1-year ahead characteristics prediction using end-of-year characteristics. The results are robust to use other similar characteristics.

trend is mostly consistent with the realizations, yet analysts make mistakes when making subjective statements about a firm's growth and cash.

7.2 Drivers of Disagreement

Finally, in this section, I combine both the factual and subjective content extracted using the LLM in Section 4 to study the drivers of disagreement. I 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,

$$Disagree^{h}_{i_1,i_2|j,t} = |E_{i_1,t}[\pi_{j,t+h}] - E_{i_2,t}[\pi_{j,t+h}]|. \tag{24}$$

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:

$$\begin{aligned} \text{Factual_Diff}_{i_1,i_2|j,t} &= 1 - \cos(x^{factual}_{i_1,j,t}, x^{factual}_{i_2,j,t}) \\ \text{Subjective_Diff}_{i_1,i_2|j,t} &= 1 - \cos(x^{subjective}_{i_1,j,t}, x^{subjective}_{i_2,j,t}). \end{aligned} \tag{25}$$

In the formula, $x^{factual}_{i,j,t}$ is the 768-dimensional vector embedding of the factual information 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 means i_1 and i_2 possess more distinct subjective outlooks in their written reports.[34]

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:

$$Disagree^{h}_{i_1,i_2|j,t} = \beta_1 \text{Factual_Diff}_{i_1,i_2|j,t} + \beta_2 \text{Subjective_Diff}_{i_1,i_2|j,t} + \tau_j + \tau_t + \epsilon_{i_{12},j,t}. \tag{26}$$

If disagreement is mainly driven by the "information channel", 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 the "belief channel", i.e. analysts having a different interpretation for the same factual information that then leads to different opinions and beliefs, we would expect a positive β_2.

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

	FY1	FY2	FY3	FY4	FY5
Factual_Diff	0.02**	-0.01	0.13**	0.70	0.34
	(2.02)	(-0.55)	(2.05)	(1.15)	(0.22)
Subjective_Diff	0.03**	0.11***	-0.07	-1.30	-0.26
	(2.25)	(3.50)	(-0.46)	(-1.56)	(-0.10)
Time FE	✓	✓	✓	✓	✓
Stock FE	✓	✓	✓	✓	✓
Total R^2	0.30	0.33	0.40	0.48	0.50

Table 15: Regression result of disagreement regression (26). Standard errors are clustered at stock and quarter level. *, **, *** denotes significance at the 10%, 5%, and 1% level.

Table 15 reports the result for regression (26). I find that the semantic differences in both the factual and subjective content significantly comoves with pairwise disagreement, suggesting that both information channel and belief channel are important drivers for analyst disagreement. The regression coefficients are significantly positive for 1-year to 3-year ahead forecast disagreement, which implies that the analyst reports contain information about short-term forecasts.

8 Conclusion

In this paper, I present a novel exploration of the belief formation processes of equity analysts through an analysis of their own written text in research reports. Using embedding techniques, I demonstrate that analyst reports contain substantial information about analysts' beliefs, especially regarding the behavioral biases that contribute to their earnings forecast errors. The textual content in these reports also predicts future realized earnings and stock returns, beyond what is captured by numerical consensus forecasts or common stock characteristics. I further leverage Large Language Models (LLMs) to extract both factual information and analysts' subjective outlooks about different aspects of firm fundamentals.

Using information extracted and summarized by LLM, I uncover novel insights into variation in analysts' attention, their mis-reaction to different types of information, reasoning in forming subjective outlooks, 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 mis-reactions to certain information in earnings forecasts, as well as systematic errors in analysts' subjective outlooks based on observed factual information. Finally, I emphasize the importance of both information and belief channels in

explaining forecast disagreements.

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.

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.

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*.

Bhandari, L. C. (1988). Debt/equity ratio and expected common stock returns: Empirical evidence. *The Journal of Finance*, 43(2):507–528.

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

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.

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. 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.

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*.

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.

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. G. (2009). The cross section of cashflow volatility and expected stock returns. *Journal of Empirical Finance*, 16(3):409–429.

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. 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.

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.

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.

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

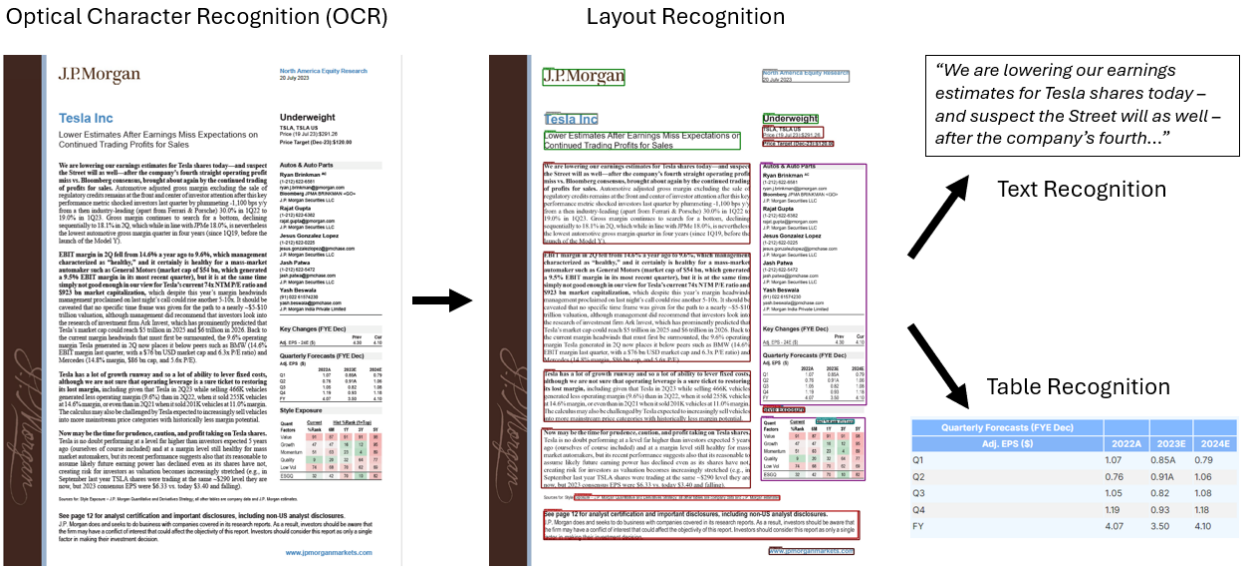


Figure 13: Illustration of parsing analyst reports

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[35]. The implementation combines Optical Character Recognition

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

(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 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 13 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.

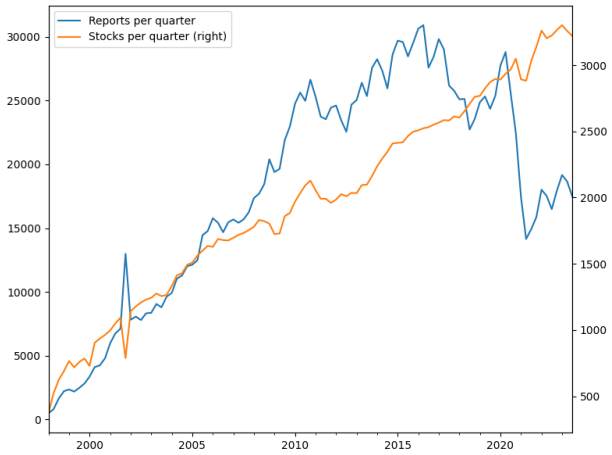
I then match contributors in Investext to brokerage firms in IBES. I only focus on brokerage firms that publishes more than 10 reports in a quarter on average. I obtain `estimid` from detail `recommendation` file from IBES. I first match the abbreviated broker names in `estimid` to full broker names (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 160 brokers in the Investext database.

A.4.1 Summary statistics

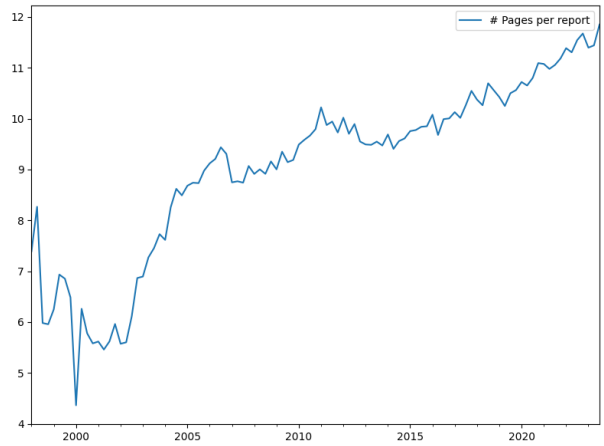
Table 16 reports summary statistics of the full Investext analyst reports sample used in this paper, and Figure 14 shows how the summary statistics of analyst reports evolve over time. The full sample contains more than 1.8 million analyst reports that covers 8,597 stocks written by 10,766 analysts employed by 543 brokerage firms. In each quarter, a broker on average issues 180 reports, and a firm on average gets 24 reports. Panel (a) of Figure 14 shows that the number of stocks covered in the full sample of analyst reports has been steadily increasing over time, and Panel (b) shows the analyst reports gets longer over time. This could reflect an increase in information production by sell-side analysts or heavier disclosure requirements. The blue line in Panel (a) shows a drop in the total number of analyst reports after 2020, and Panel (c) confirmed that there's a drop in the stock coverage and reports written per broker. These evidence might suggest a negative impact on sell-side research activities due to the Covid-19 pandemic, which started in the beginning of 2020.

Total number of reports	1855919
Number of unique stocks covered	8597
Number of unique brokerage firms	543
Number of unique research analysts	10776
Avg. # of pages per report	9.10
Quarterly avg. # of stock coverage per broker	84.4
Quarterly avg. # of reports per broker	179.8
Quarterly avg. # of brokers coverage per firm	3.94
Quarterly avg. # of reports per firm	24.3

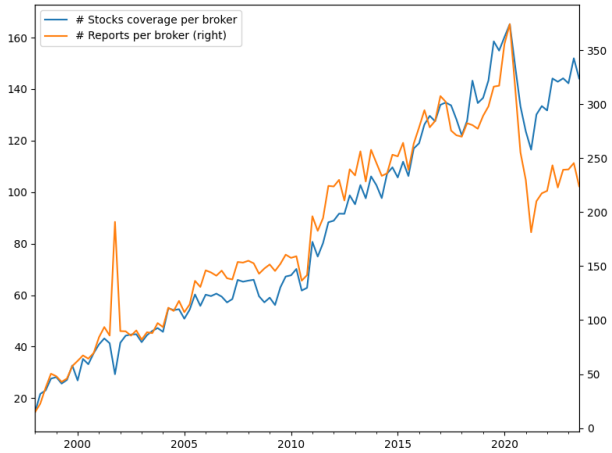
Table 16: Summary statistics of the full Investext sample.



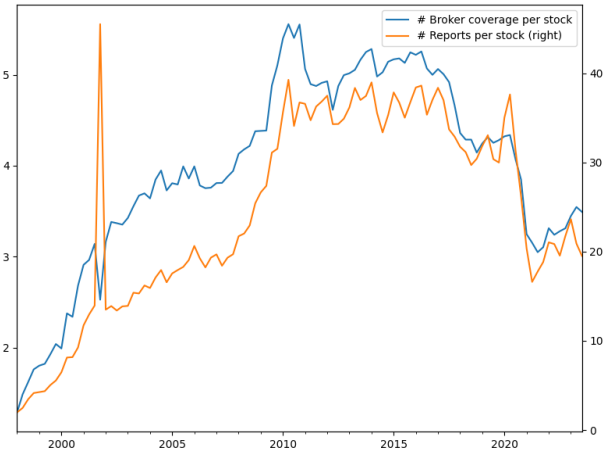
(a) Number of reports and number of stocks



(b) Average # of pages per report



(c) Reports and stock coverage per broker



(d) Reports and broker coverage per stock

Figure 14: Quarterly summary statistics of analyst reports

A.5 Matching reports with IBES records

In the final step, I download detailed EPS forecasts, price targets, and recommendations from Detailed IBES files. I match IBES records with Investext analyst reports by matching on the `estimid-cusip-anndats` combinations. Each identifies an announcement of a company from a broker. For each `estimid-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.

Table 17 reports the matching results. The first column shows the matching result for EPS forecasts. Of all the 5,522,326 unique `estimid-cusip-anndats` combinations coming from 1541 unique brokers in IBES Detailed EPS forecast file, the matched 160 brokers made 2,880,445 announcement, which represents 52.2% of the total announcements. Among those announcement, 777,625 of them can be matched with at least one report published by the broker that contains the cusip within the 3-business day window around the announcement day, representing a 26.9% match rate. These 777,625 are matched with 1,232498 reports, as there could be multiple reports issued within the 3-business day window about the same company.

The second and third column report the matching result for price targets and recommendations. For price targets, of all the 2,152,998 unique `estimid-cusip-anndats` combinations coming from 1085 brokers, the matched 160 brokers made 1,199,471 announcement, representing a coverage of 55.7%. Among those 1,199,471 announcement, the matching process matched 366,214 of them with at least one analyst reports, a match rate of 30.5%. For recommendations, of all the 709,840 unique `estimid-cusip-anndats` combinations coming from 1225 brokers, the matched 160 brokers made 357,781 announcement, representing a coverage of 50.4%. Among those 357,781 announcement, the matching process matched 71,312 of them with at least one analyst reports, a match rate of 19.9%.

	EPS forecast	Price target	Recommendation
Total announcements	5522326	2152998	709840
Total # of brokers	1541	1085	1225
From matched 160 brokers	2880445	1199471	357781
Ann. matched with reports	777625	366214	71312
Total # of matched report	1232498	588536	101163

Table 17: Summary results of matching IBES announcements with analyst reports.

B Additional Results for Section 3

B.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 3.1 and predict using text embeddings

$$\begin{aligned} E_t^i[\pi_{j,t+h}] &= \theta_h x_{i,j,t} + \epsilon_{i,j,t} \\ Revision_{i,j,t} &= \theta_h x_{i,j,t} + \epsilon_{i,j,t} \end{aligned} \tag{27}$$

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.

B.2 Original Coibion Gorodnichencko regression

I confirm that analysts under-react in short-term individual forecast while over-react in long-term individual forecast. I run the following regression at the individual forecast level:

$$\pi_{j,t+h} - E_{i,t}[\pi_{j,t+h}] = \alpha + \beta(E_{i,t-1}[\pi_{j,t+h}] - E_{i,t}[\pi_{j,t+h}]) + \epsilon_{i,j,t} \tag{28}$$

where $\pi_{j,t+h}$ is firm j's h-year ahead EPS at time t and $E_{i,t}[\pi_{j,t+h}]$ is analyst i's forecast of $\pi_{j,t+h}$. The left-hand side is the forecast error, and the main regressor on the right-hand side is the forecast revision, defined as the change of analyst's forecast of $\pi_{j,t+h}$ from $t-1$ to t. As the forecasting behavior can be different for recession versus boom, and for analysts with different experience levels (e.g. see Kelly et al. (2024)), I control for analyst fixed effect and time fixed effect in the regression specification. Panel (A) in Table 23 shows the results for the forecasts that can be matched with analysts reports, and Panel (B) reports the coefficients for all IBES forecasts. In addition, in Panel (C) I conduct consensus level CG regression

$$\pi_{j,t+h} - \overline{E}_{i,t}[\pi_{j,t+h}] = \alpha + \beta(\overline{E}_{i,t-1}[\pi_{j,t+h}] - \overline{E}_{i,t}[\pi_{j,t+h}]) + \epsilon_{i,j,t} \tag{29}$$

where $\overline{E}[\cdot]$ denotes the consensus belief obtained from IBES consensus `statsum` file. Outliers outside 2.5% and 97.5% percentiles are dropped.

	1 Year	2 Years	3 Years	4 Years	5 Years
Panel (A): Forecasts matched with analyst reports					
α	-0.00***	-0.01***	-0.01***	-0.02***	-0.04***
	(-5.10)	(-18.24)	(-17.70)	(-17.67)	(-16.85)
β	0.13***	0.22***	-0.03	-0.18***	-0.22***
	(8.63)	(7.26)	(-1.10)	(-4.71)	(-3.77)
Year FE	✓	✓	✓	✓	✓
Analyst FE	✓	✓	✓	✓	✓
R^2	0.00	0.001	0.00	0.01	0.01
Num Obs.	588403	524100	174533	36214	20241
Panel (B): All IBES individual forecasts					
α	-0.00***	-0.01***	-0.01***	-0.03***	-0.05***
	(-11.88)	(-27.72)	(-25.22)	(-25.08)	(-24.58)
β	0.12***	0.20***	0.02	-0.11***	-0.12***
	(10.63)	(8.25)	(0.69)	(-3.72)	(-2.89)
Time FE	✓	✓	✓	✓	✓
Analyst FE	✓	✓	✓	✓	✓
R^2	0.00	0.00	0.00	0.00	0.01
Num Obs.	3303642	2840700	954514	173583	93114
Panel (C): IBES consensus forecasts					
α	-0.01***	-0.02***	-0.02***	-0.04***	-0.07***
	(-20.13)	(-18.73)	(-19.25)	(-38.73)	(-42.54)
β	0.16***	0.30***	-0.08***	-0.26***	-0.22***
	(6.22)	(5.55)	(-2.40)	(-7.99)	(-5.16)
Time FE	✓	✓	✓	✓	✓
Stock FE	✓	✓	✓	✓	✓
R^2	0.00	0.00	0.00	0.00	0.01
Num Obs.	654878	559752	313838	125047	77740

Table 20: Baseline CG-Regression results of different IBES forecast horizons.

B.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

B.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

$$\widehat{\pi}^{\text{text}}_{j,t+h} = \gamma_{\text{IBES}} E^{\text{IBES}}_t[\pi_{j,t+h}] + \gamma_{\text{JKP}} \widehat{\pi}^{\text{JKP}}_{j,t+h} + \epsilon_{j,t+h} \tag{30}$$

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.

B.5 Regression of individual forecast error on current realized earnings

In this section, I conduct the KW-regression following Kohlhas and Walther (2021) by regressing individual analyst's forecast error on current realized fiscal year EPS (`fy0a` variable from IBES `statsum` file)

$$\pi_{j,t+h} - E_{i,t}[\pi_{j,t+h}] = \alpha + \beta^{KW} \pi_{j,t} + \epsilon_{i,j,t} \tag{31}$$

I continue to scale EPS numbers by current price and control for analyst and time fixed effects. I run this regression using all IBES individual forecasts. Outliers outside 2.5% and 97% percentiles are dropped. Table

	1 Year	2 Years	3 Years	4 Years	5 Years
Panel (A): Forecasts matched with analyst reports					
α	-0.00***	-0.01***	-0.01***	-0.03***	-0.05***
	(-12.21)	(-16.99)	(-22.94)	(-32.55)	(-34.98)
β	-0.00	0.02**	-0.04***	-0.15***	-0.33***
	(-1.41)	(2.05)	(-3.05)	(-9.13)	(-12.19)
Year FE	✓	✓	✓	✓	✓
Stock FE	✓	✓	✓	✓	✓
R^2	0.00	0.001	0.00	0.01	0.01
Num Obs.	588403	524100	174533	36214	20241

Table 23: Regression of individual forecast error on current realized earnings.

C 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 15: 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 16: Prompt #1: Identify company information

Prompt #2: Separating factual and subjective statements

SYSTEM PROMPT:

Forget previous prompts.
You are an expert in financial economics research. You will be given a context chunk from an equity research report written by an equity research analyst at {broker name} on {report date}.
Your goal is to answer user's questions below by extracting either equity analyst's opinion or factual information about a company in the context chunk.

Answer the question based on the context below only. Do not make up the answers.

USER PROMPT:

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

The above is the knowledge base.
Answer the following questions and respond in JSON format.

Questions:
Q1: Based on the context, extract the factual information about {ticker} or {company name} that is related to the company's operations, business, and financials.
Include as much details about the company's operations, business, and financials as possible.
Exclude analyst's opinions, estimates, or any subjective statements made by the analyst.
The answer should include all objective facts about {ticker} or {company name} mentioned in the text.
Write your as a passage in natural language.
Answer 'N/A' if there's no factual information related to the company's operations, business, and financials about {ticker} or {company name}. Do not make up your answer!

Q2: Based on the context, describe the analyst's view about the future of {ticker} or {company name} that is related to the company's operations, business, and financials.
Start your answer with "The analyst believes {ticker} will ...".
Include as much details about the analyst's view about the future of company's operations, business, and financials as possible.
Exclude any factual information that is already answered in Q1.
Answer 'N/A' if analyst doesn't provide a view in the context. Do not make up your answer!

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

Figure 17: Prompt #2: Separating factual and subjective statements

Prompt #3: Extracting financial economic concepts

SYSTEM PROMPT:

Forget previous prompts. You are an expert in financial economics research.
You will be given a summary paragraph from an equity research report written by an equity

research analyst.The summary paragraph contains some factual information about a public company. Your goal is to answer user's questions below by summarizing the financial or economic concepts in the paragraph.
Answer the question based on the context below only. Do not make up the answers.

USER PROMPT:

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

The above is the knowledge base.
Answer the following questions and respond in JSON format.

Questions:
Q1: Based on the context, what financial or economic concepts are mentioned in the context? Write your answer as short phrases less than 3 words, separated by comma.
Q2: For each concept you identified in Q1, is there an adjective describing the concept? If yes, add the corresponding adjective to the phrase. If no, answer N/A.

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

Figure 18: Prompt #3: Extracting financial economic concepts

D Additional Examples of LLM Responses

D.1 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."

D.2 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"

E Additional results for Section 6 and 7

E.1 Forecast error prediction with *Attn* and *Signal*

To investigate how analyst's forecast error depends on both the information content and analyst's attention to different fundamental components, I augment regression 20 with analyst's attention:

$$\pi_{j,t+h} - E_{i,t}[\pi_{j,t+h}] = \sum_k \gamma_k Signal^k_{i,j,t} + \sum_k \lambda_k Attn^k_{i,j,t} + \tau_j + \tau_t + \epsilon_{j,t} \tag{32}$$

Table 24 shows the result. The main regression coefficient $\widehat{\gamma}_k$ are unchanged as in Table 14. In addition, I find that analyst's attention to debt and expense have positive coefficients while analyst's attention to growth and assets have negative coefficients. Intuitively, this shows that analysts are too pessimistic when they pay more attention to firm's debt and expense, while being too optimistic when they pay more attention to firm's growth and assets.

		Forecast Horizon					
		All	1 Years	2 Years	3 Years	4 Years	5 Years
Signal for	Revenue	0.04***	0.03***	0.05***	0.03	0.01	0.06
		(5.44)	(6.76)	(5.4)	(1.13)	(0.27)	(1.05)
	Earnings	0.03***	0.03***	0.04***	-0.0	-0.0	-0.01
		(6.32)	(7.26)	(5.62)	(-0.14)	(-0.1)	(-0.18)
	Margin	0.03***	0.03***	0.04***	0.0	-0.01	-0.01
		(6.75)	(7.43)	(5.56)	(0.32)	(-0.24)	(-0.15)
	Debt	-0.02**	-0.01***	-0.03***	0.01	0.06	0.02
		(-2.16)	(-2.48)	(-2.75)	(0.54)	(1.15)	(0.28)
	Cash	0.01	0.01**	0.01	-0.01	-0.04	0.02
		(1.23)	(2.19)	(0.46)	(-0.34)	(-0.71)	(0.24)
	Growth	0.0	0.0	0.01	0.0	-0.01	-0.08
		(0.63)	(0.88)	(1.28)	(0.13)	(-0.34)	(-1.07)
	Expense	-0.01***	-0.0	-0.02***	-0.03**	-0.04	-0.03
		(-2.9)	(-0.63)	(-2.92)	(-2.14)	(-1.04)	(-0.59)
	Equity	0.03***	0.02***	0.04***	0.02	-0.01	0.1
		(4.14)	(3.48)	(3.42)	(0.99)	(-0.09)	(1.01)
	Assets	-0.02***	-0.01**	-0.04***	-0.03	0.02	0.07
		(-3.11)	(-2.34)	(-3.59)	(-1.1)	(0.22)	(0.57)
Attn for	Revenue	0.02	-0.01	0.01	-0.0	-0.13	-0.02
		(0.89)	(-0.74)	(0.32)	(-0.09)	(-1.04)	(-0.13)
	Earnings	-0.0	0.01	-0.01	-0.04	-0.16	0.02
		(-0.13)	(0.51)	(-0.43)	(-0.8)	(-1.56)	(0.12)
	Margin	0.03	0.02	0.03	-0.01	-0.01	-0.02
		(1.32)	(1.43)	(0.99)	(-0.24)	(-0.07)	(-0.1)
	Debt	0.09***	0.01	0.11***	0.21***	-0.06	-0.01
		(2.91)	(0.43)	(2.55)	(2.49)	(-0.25)	(-0.02)
	Cash	0.06*	0.0	0.03	0.13	0.22	0.16
		(1.75)	(0.0)	(0.67)	(1.4)	(1.42)	(0.48)
	Growth	-0.04**	0.02	-0.07**	-0.13***	-0.04	-0.11
		(-2.17)	(1.62)	(-2.19)	(-2.51)	(-0.36)	(-0.46)
	Expense	0.07***	0.02	0.06***	0.19***	0.24*	0.3
		(3.46)	(1.04)	(2.51)	(3.38)	(1.91)	(1.16)
	Equity	0.01	0.03*	0.0	-0.11*	-0.18	-0.08
		(0.49)	(1.72)	(0.03)	(-1.78)	(-1.18)	(-0.34)
	Assets	-0.06*	-0.04*	-0.09**	-0.09	0.0	-0.45
		(-1.95)	(-1.74)	(-2.08)	(-0.95)	(0.01)	(-1.06)
Total R^2		0.15	0.14	0.23	0.31	0.47	0.59
Time FE		✓	✓	✓	✓	✓	✓
Stock FE		✓	✓	✓	✓	✓	✓

Table 24: This table shows the regression result of (20). h denotes the number of years between current quarter and forecast target date. Standard errors are clustered at stock and quarter level. *, **, *** denotes significance level at 10%, 5% and 1%.

E.2 Drivers of attention in subjective statement

I measure the attention in analyst's subjective content in the same spirit as the attention measure in the factual content. I rename the attention measure in (10) as $Attn_{i,j,t}^{k,factual}$ to emphasize that it's computed from the factual content, and I define $Attn_{i,j,t}^{k,subjective}$ as

$$Attn_{i,j,t}^{k,subjective} = \frac{\sum_{c \in A_{i,j,t}^{subjective}} \mathbf{I}(c \in C^k)}{\sum_{c \in A_{i,j,t}^{subjective}} \mathbf{I}} \tag{33}$$

I then run the two regression as in Section 7.1:

$$Attn_{i,j,t}^{k,subjective} = \beta Attn_{i,j,t}^{k,factual} + \tau_j + \tau_t + \epsilon_{i,j,t} \tag{34}$$

$$Attn_{i,j,t}^{k,subjective} = \sum_{k'} \beta_{k'} Attn_{i,j,t}^{k',factual} + \tau_j + \tau_t + \epsilon_{i,j,t} \tag{35}$$

where β in (34) captures how attention in own factual information translates to attention in own subjective contents, and $\beta_{k'}$ captures how attention in subjective content depends on attention in factual information in other components. Table 25 and Figure 19 reports the results. I find that when analysts pay more attention to factual information of component k, they also discuss more about k in their subjective statements.

	Revenue	Earnings	Margin	Debt	Cash	Growth	Expense	Equity	Assets
Coef	0.39***	0.34***	0.28***	0.5***	0.47***	0.22***	0.41***	0.29***	0.28***
	(94.48)	(96.73)	(88.86)	(68.60)	(71.20)	(67.93)	(99.57)	(37.89)	(45.38)
Time FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
Stock FE	✓	✓	✓	✓	✓	✓	✓	✓	✓
R^2	0.07	0.07	0.07	0.1	0.13	0.06	0.11	0.02	0.04

Table 25: Regression result of belief transmission regression (34) for each component. Standard errors are clustered at stock and quarter level. *, **, *** denotes significance level at 10%, 5% and 1%.

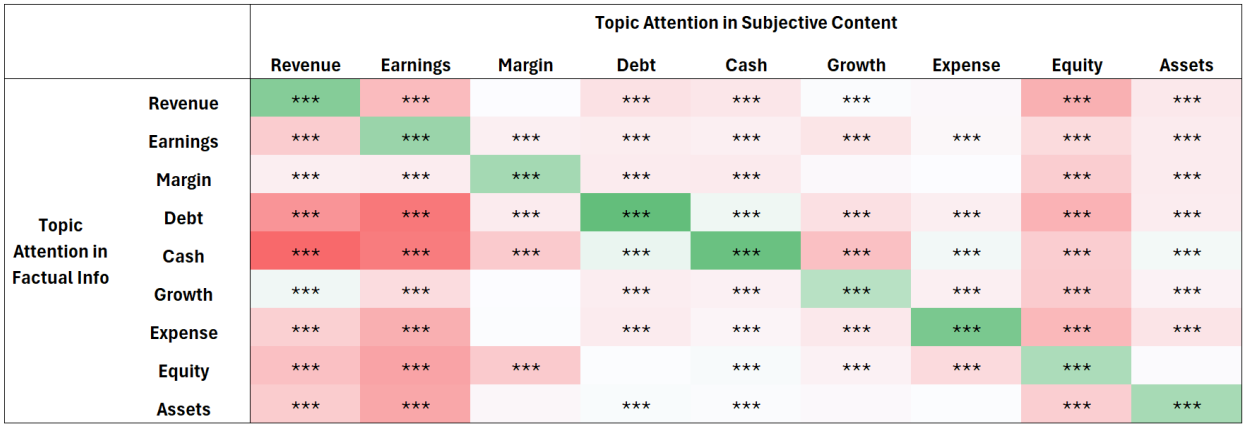


Figure 19: Heatmap of regression coefficient of regressing attention to topic k in subjective content on attention in factual content. Each column represents one regression (35) for a specific topic.More red (green) cell denotes more negative (positive) coefficient estimate from regression (35). *** denotes significance at 1% level.