Economic Representations

Suproteem K. Sarkar

January 11, 2025

Please click here for the latest version

Valuations depend on how people categorize, perceive, or otherwise *represent* economic objects. This paper develops a measure of how the market represents firms, and uses this measure to study stock valuations. I train an algorithm to structure language from financial news into embeddings—vectors that quantify the economic features and themes in each firm's news coverage. I show that a firm's vector representation is informative of how the market perceives its business model. Representations explain cross-sectional variation in stock valuations, cash flow forecasts, and return correlations. Changes in representation help to explain changes in stock prices. Some changes in representations and prices are forecastable, and indicate that some of the explained variation in stock valuations stems from misperception. I find that misperception and misvaluation can intensify when a firm's news coverage includes attention-drawing features—like "internet" in the late 1990s or "AI" in the early 2020s.

JEL Codes: C55, E71, G12, G40, O31

Models: I trained a series of language models for social science research. They have been released to the research community: collection of foundation models, representation model.

I am especially grateful to Sendhil Mullainathan, Andrei Shleifer, Jeremy C. Stein, and Adi Sunderam for their guidance and support. I thank Malcolm Baker, John Y. Campbell, Jiafeng Chen, Sahil Chinoy, Lauren Cohen, John Conlon, Melissa Dell, Mark L. Egan, Xavier Gabaix, Edward L. Glaeser, Jeff Gortmaker, Shlok Goyal, Robin Greenwood, Samuel G. Hanson, Sebastian Hillenbrand, Spencer Yongwook Kwon, Max Miller, Dev Patel, Ashesh Rambachan, Dominic Russel, Marco Sammon, Kunal Sangani, Joshua Schwartzstein, Cassidy Shubatt, Emil Siriwardane, Erik Stafford, Johnny Tang, Keyon Vafa, Philippe van der Beck, Luis M. Viceira, and seminar participants at Harvard, MIT, and the University of Chicago. I gratefully acknowledge support from a National Science Foundation Graduate Research Fellowship under grant number DGE-2140743, a Two Sigma PhD Fellowship, and the Molly and Dominic Ferrante Economics Research Fund. First version: October 2024. Sarkar: Harvard University (suproteemsarkar@g.harvard.edu).

Valuations are challenging to explain. Empirical research cannot fully account for variation in stock valuations (Roll, 1988; Campbell et al., 2001, 2022), or valuations of startups, collectibles, and homes (Case and Shiller, 1989; Gompers and Lerner, 2001; Ashenfelter and Graddy, 2003).

To make progress on this explanatory challenge, it could help to measure how people perceive, categorize, associate—or otherwise *represent*—the objects of valuation. However, while economic theories often highlight the role of perception (Simon, 1956; Tversky and Kahneman, 1981; Mullainathan, 2002), it can be difficult to quantify perception using traditional structured data (Black, 1986; Shiller, 2019).

Language can be informative of perception. An investor who evaluates an electric car company, for example, might describe it as an "internet-of-cars company" or "just an automaker." Similarly, a homebuyer might describe a neighborhood as "up-and-coming" or "run down," and a consumer might describe a purchase as an "essential" or a "luxury." To measure this information, we must impose economically meaningful structure on language data.

This paper uses language data to measure how the market represents firms. I train an algorithm to transform financial news articles into *embeddings*—vectors that organize the language in each firm's news coverage. In this vector space, geometric proximity encodes linguistic similarity between firms. I use each firm's vector representation to proxy for how the market perceives its business model.

In my first set of empirical results, I show that representations help to explain valuations. Representations explain variation in stock prices, cash flow forecasts, return correlations, and analyst coverage. In addition, representations improve the explanatory power of cross-sectional asset pricing tests that use traditional stock characteristics and industry information. Changes in representation further explain variation in returns, which indicates that changes in a firm's perceived business model can contribute to changes in its valuation.

In my second set of empirical results, I show that some changes in representation are non-fundamental—the market misperceives some firms' business models. When a firm's representation deviates from a historical benchmark, the firm has predictable reversals in representation and predictable returns. Some of this misperception appears to stem from attention-drawing features of firms, relates to communication by managers, and accompanies technological transformations.

For this empirical procedure, I trained a series of language models using historical, time-indexed language data. This training strategy prevents lookahead bias from

pretraining, which is an issue with off-the-shelf language models that are trained on contemporary data. I have released these language models to the research community.

Measuring representations from language. Language from the financial news is detailed and dynamic. Variation in this language can reflect variation in how the market perceives firms. Consider news coverage of the electric vehicle maker Tesla from 2021 to 2023. In 2021, Tesla was called "an internet-of-cars company," and a firm that had "grown up as [a] software and tech company first, and automaker second." In 2023, Tesla was called "just an automaker ... with automaker problems and automaker cyclicality," as well as "a metal bender like everybody else." Over this period, mentions of "software" in Tesla's Wall Street Journal coverage declined by 24%.[1] These changes in language suggest a change in how the market perceived Tesla.

To analyze these kinds of language patterns across many firms, I use data from the Dow Jones Newswires. This financial news archive aggregates content from sources like the Wall Street Journal, Barron's, and MarketWatch. I develop an algorithm to transform each firm's coverage in each year into an embedding—a vector that structures information in language data (Jurafsky and Martin, 2024). To develop this algorithm, I train language models on millions of historical newspaper articles,[2] and fine-tune these models on financial news language.[3] Each embedding is a vector representation that quantifies the economic features and themes in a firm's news coverage.

Measured representations relate to market representations. To validate the representations measure, I show that it relates to how the market perceives firms. Linear projections in the representation vector space relate to features of firms like "platform economy," "upmarket," and "direct-to-consumer." In addition, higher representation similarity between a pair of firms is associated with higher shared analyst coverage and higher pairwise return correlation.

I further validate the measure by showing it helps to explain valuations. Representations help to explain 15–33% of the variation in valuation ratios and cash flow forecasts.

[1] Appendix A.1 includes links to these articles and further describes Tesla's news coverage.

[2] These historical language datasets were created through the excellent data curation and data transformation work by Dell et al. (2024, American Stories) and Silcock et al. (2024, Headlines). Researchers often analyze language using general-purpose foundation models, which can produce lookahead bias from pretraining (Glasserman and Lin, 2024; Sarkar and Vafa, 2024) and can generate results that do not replicate (Chen et al., 2023). I train new foundation models to avoid these issues in my analysis.

[3] I target the models to financial news language using contrastive learning. Contrastive learning is a procedure that can extract information from unstructured data, including information that general-purpose machine learning models may miss (Dell, 2024).

The measure adds incremental explanatory power over traditional stock characteristics and industry labels, which explain 11–26% of this variation. As these R^2 statistics are computed on a separate sample, explanatory power does not mechanically increase with the number of parameters—R^2 gains reflect additional information in the measure.

These validation results indicate that a firm's *embedding representation* is informative of its *market representation*. The market representation—how the market reasons about the firm's business—is not directly observable. The embedding representation—how the press writes about the firm's business—is quantifiable. Embedding representations could help to study how the market forms valuations.

Changes in representation help to explain changes in prices. In my first set of main results, I show that representations help to explain changes in stock prices. I characterize economic mechanisms that can explain these changes.

I first evaluate how well representations explain cross-sectional variation in annual returns. To benchmark these results, I also use characteristics traditionally studied in asset pricing, like industry, dividends, and profitability. Representations combined with traditional characteristics explain 19% of the variation in annual returns. On their own, representations explain 13% of the variation in returns, and traditional characteristics explain 11% of the variation in returns.

What are the economic mechanisms that could drive these returns? I decompose the estimated return on a stock into two components. The first component is from the change in an aggregate valuation function that maps representations into prices. The second component is from the change in the representation itself. Of the total variation in returns explained by representations, I estimate that two-thirds relates to changes in valuation functions, and one-third relates to changes in representations.

The aggregate valuation function can change if investors change how they value business models. Changes in valuation functions can drive changes in prices: For example, the stock prices of oil companies can change as commodity prices change, or as expectations about climate policy change. Accounting for this component increases explanatory power by a factor of 1.8 relative to traditional characteristics. These results indicate that the representations measure contains additional information about the perceived business model.

The representation of a firm can change if investors change how they perceive the firm's business model. Changes in representation can also drive changes in prices: For example, the stock price of an automaker can change if investors perceive it as a car

firm in one year, a software firm in the next, and a battery firm in the year after. I find that using changes in industry labels to proxy for changes in perception poorly explains returns. Unlike infrequent changes in industry labels, changes in representation can be measured at high frequency, and have additional explanatory power over returns. These results indicate that changes in perceived business models are also important drivers of changes in stock prices. If how Tesla is perceived changes from an "Internet-of-cars company" to a "metal bender like everybody else," its valuation may change as well.

Predictable changes in representations and prices indicate misperception. In my second set of main results, I show that some changes in representation are non-fundamental—the market sometimes misperceives firms' business models. I find that non-fundamental changes in representation can generate misvaluation.

I first show that when a firm's representation deviates from the historical mean, its representation predictably reverts. I construct a trading strategy that goes long firms whose deviation implies a lower price and goes short firms whose deviation implies a higher price. This strategy forecasts returns ($t = 3.4$), which indicates that price changes associated with deviations in representation predictably revert.

Why do these deviations in representation predictably revert? If a firm's representation changes because its true business model has changed, the representation is unlikely to revert unless true business models revert at short horizons. However, if the market is aware of such true reversion, the forward-looking stock price is unlikely to respond to such a transitory change in the business. Instead, predictable reversion in representations and prices suggest that the market predictably misperceives some firms' business models.

The market may misperceive a firm's business if some of the firm's features disproportionately draw attention and lead investors to neglect fundamentals. For example, although "Internet" drew attention in the late 1990s, and "AI" has drawn attention in the early 2020s, many firms' operations are not as related to these features as investors may believe (Cooper et al., 2001; Narayanan and Kapoor, 2024).

I measure whether a firm's representation incorporates features associated with trending news coverage, extreme profitability, or extreme returns. Betting against firms whose representations incorporate these features leads to additional return predictability over the raw deviation in representation. These results are consistent with a kind of "financial revisionism" (Hong et al., 2007)—investors predictably revert to old ways of thinking about a firm. Although Tesla was once an "internet-of-cars company," some of its shift to a "metal bender like everybody else" may have been predictable.

These results on representation and price predictability could be a consequence of valuation by analogy. Investors commonly use analogical reasoning in comparable companies analysis (Graham and Dodd, 1934; Bhojraj and Lee, 2002; Damodaran, 2012). Popular descriptions of firms often draw on analogies—the transportation firm Blade has been called "Uber for the air," while the food producer Beyond Meat has been called the "Tesla of meat." Analogical reasoning can help people adapt to changing environments, but it can also lead to predictable mistakes when the features that draw attention relate to "surface" similarities (Holyoak and Koh, 1987). When an analogy is extreme or loads heavily on attention-drawing features, investors may not fully appreciate a firm's fundamentals.

The role of communication by firms. To explore one potential influence on representations, I analyze communication by firms. I find that managerial communication sometimes adopts economic features that may attract the market's attention—including "AI," "virtual reality," and "internet-of-things." Across many features, these mentions in managerial communication eventually decrease. Beyond these individual features, I also find reversion in the overall content of managerial communication—deviations in embeddings of managers' speeches predictably revert to the historical mean. In addition, I find that changes in managerial communication are associated with changes in the representations measure. This evidence suggests that managerial communication may influence perceptions of firms. Managers may supply the market with analogies or other models of thinking about firms. These models may be consistent across the kinds of features that draw attention, but may not fully reflect firms' fundamentals.

Related literature. This paper contributes to literatures on characteristics-based empirical modeling, asset pricing, behavioral economics, and machine learning.

First, by developing vector representations of firms from language, this paper contributes to the literature on characteristics-based and feature-based empirical modeling. Empirical research often uses these models to explain valuations and other outcomes (Lancaster, 1966; Berry et al., 1995; Daniel and Titman, 1997; Koijen and Yogo, 2019). Many potentially relevant features can be challenging to measure. This paper uses contrastive learning to measure representations of firms, and finds that these representations help to explain outcomes.[4]

[4]Bajari et al. (2023), Magnolfi et al. (2024), Compiani et al. (2024), and Han et al. (2024) compute embedding representations of products, and study elasticities, prices, and product differentiation.

Second, by showing that representations help to explain valuation ratios and cash flow forecasts, this paper contributes to the asset pricing literature on explaining valuations.[5] Present value decompositions relate valuation ratios to future cash flows and returns (Campbell and Shiller, 1988; Cohen et al., 2003).[6] This paper shows that the contemporaneous information in representations can help to explain valuation ratios.

Third, by quantifying the price effects of changes in representation and changes in how representations are valued, this paper studies economic mechanisms for movements in stock prices. Empirical research cannot fully explain changes in the prices of individual stocks (e.g. Roll, 1988; Campbell et al., 2001, 2022).[7] This paper shows that representations computed from language data help to explain returns.[8] Representations add to the explanatory power of cross-sectional asset pricing tests that use traditional characteristics, and changes in representation further explain returns.

Fourth, by finding evidence of misrepresentation in observational data, this paper contributes to the literature in behavioral economics and finance on cognitive representation and misrepresentation. While many theories make predictions about misrepresentation (e.g., Mullainathan et al., 2008; Bordalo et al., 2024),[9] it has been difficult to test the full extent of these predictions in the field since it is challenging to measure

[5] In his presidential address, Cochrane (2011) writes that valuation ratios "should be our *left-hand* variable, the thing we're trying to *explain*."

[6] van Binsbergen et al. (2023) and Cho and Polk (2024) use future realizations of cash flows and returns to characterize variation in market prices. Rhodes–Kropf et al. (2005) and Golubov and Konstantinidi (2019) estimate contemporaneous components of valuation ratios to study mergers and returns.

[7] While many papers have focused on explaining the returns of aggregate portfolios of stocks (e.g., Fama and French, 1992; Jagannathan and Wang, 1996), it is more difficult to explain the returns of individual stocks (Lewellen et al., 2010). The unexplained component contributes to a large share of the variation in individual stock returns (Campbell et al., 2001, 2022). Changes in expectations help to explain some of the movements in portfolio returns and individual stock returns (Bordalo et al., 2025). Recently, asset pricing research has shown that language can help to characterize economic mechanisms behind price changes. Bybee et al. (2024) find that aggregate news topics help to explain returns on the market portfolio, and Bybee et al. (2023) find that return correlations with aggregate news topics help to explain returns on stock portfolios. This paper uses language to measure changes in the representations of individual stocks.

[8] The finance literature has developed methods to include rich conditioning information in asset pricing tests (e.g. Kelly et al., 2019; Freyberger et al., 2020; Kozak and Nagel, 2023; Bryzgalova et al., 2023; Didisheim et al., 2023). This paper shows that embeddings of financial language are effective sources of conditioning information.

[9] These theories include categorization (Mullainathan, 2002; Barberis and Shleifer, 2003; Fryer and Jackson, 2008), framing (Tversky and Kahneman, 1981), limited attention (Peng and Xiong, 2006; Hirshleifer et al., 2011; Kőszegi and Szeidl, 2013; Gabaix, 2014; Schwartzstein, 2014; Bordalo et al., 2023; Bohren et al., 2024), model selection (Hong et al., 2007; Ortoleva, 2012; Schwartzstein and Sunderam, 2021; Yang, 2023), analogy (Jehiel, 2005), and attenuation (Gabaix, 2019; Woodford, 2020; Enke, 2024), and relate to work in psychology on categorization (Rosch, 1973; Ashby and Maddox, 2005), attention (Treisman and Gelade, 1980; Nosofsky, 1986; Kahana, 2012) and analogy (Ross, 1987; Gentner, 2003).

very rich representations of firms. Therefore, empirical research in behavioral finance has often focused on changes in the valuation of individual business models, and has focused less on changes in the perceived business model itself.[10] This paper develops a measure of the perceived business model, and finds mispricing that is consistent with misrepresentation.[11]

Finally, by developing new tools for the research community, this paper contributes to research on machine learning for social science and credibility in machine learning. Finance research has computed embeddings of assets using language, returns, and holdings data (e.g., Chen and Sarkar, 2020; Dolphin et al., 2022; Chen et al., 2024; Gabaix et al., 2024).[12] This paper learns embeddings from language data, as news language is informative of how people reason about firms.[13] To avoid credibility issues with language models in empirical analysis, I train a series of foundation models for social science research.[14]

[10]Existing empirical studies of variation in perceived business models or asset categories often relate to individual dimensions like maturity (Shue et al., 2024), share price (Green and Hwang, 2009; Shue and Townsend, 2021), news coverage (e.g. Barber and Odean, 2008), index membership (e.g., Harris and Gurel, 1986; Shleifer, 1986; Barberis et al., 2005; Boyer, 2011), or other asset styles (Baker et al., 2022; Liu, 2022). In a multi-dimensional study, Chen et al. (2016) show that industry classifications of conglomerates can influence investor behavior, and find that firms may take advantage of this investor heuristic. This paper develops a detailed measure of changes in perceived business models using representations.

[11]By studying the relationship between managerial communication and representations, this paper builds on the literature on financial persuasion and managerial influence (e.g. Baker and Wurgler, 2004; Mullainathan and Shleifer, 2005b; Bergstresser and Philippon, 2006; Solomon, 2012; Schwartzstein and Sunderam, 2021). Managerial communication may relate to economic narratives (Shiller, 2019; Flynn and Sastry, 2022). For example, Cooper et al. (2001) finds that firms that adopted internet-related language during the dotcom boom experienced large stock price increases, regardless of the relevance of the internet to their operations. This paper finds that managerial communication highlights attention-drawing economic features, and that changes in communication are associated with changes in representations.

[12]Chen et al. (2024) use embeddings of financial news language to forecast stock returns. Dolphin et al. (2022) use embeddings of returns to learn sector classifications. Gabaix et al. (2024) use embeddings of stock holdings to fit valuations, comovement, and investor holdings. The studied horizon differs between this paper and Chen et al. (2024)—the authors compute embeddings of individual news articles to construct short-horizon trading strategies, while I compute mean embeddings of firms across longer horizons to measure economic representations. The main data source differs between this paper and Gabaix et al. (2024)—the authors focus primarily on investor holdings while I focus primarily on financial news language. This paper also has a strong focus on how changes in embeddings can help to quantify economic mechanisms behind price changes and mispricing, which is not a focus of the other papers.

[13]Given this empirical strategy, this paper also builds on the literature that uses language to study financial behavior, which includes Tetlock (2007, 2014), Tetlock et al. (2008), Li (2008), Hoberg and Phillips (2010, 2016), Loughran and McDonald (2011), Da et al. (2011), Jegadeesh and Wu (2013), Hassan et al. (2019), Ke et al. (2020), Fedyk and Hodson (2023), Flynn and Sastry (2022), Bybee et al. (2024), and van Binsbergen et al. (2024).

[14]The results on lookahead bias (Glasserman and Lin, 2024; Sarkar and Vafa, 2024) and non-replicability (Chen et al., 2023) relate to a larger literature on credibility in machine learning (e.g., Kapoor and Narayanan, 2022; Zhang et al., 2024). Using job sequence data, Vafa et al. (2024) also train a foundation model for social science research.

1. Organizing Framework

To motivate this paper's empirical analysis, I present an organizing framework to model valuation formation. I assume that a firm's stock price depends on its representation and an aggregate valuation function. The representation is a vector of loadings on features, and the valuation function maps feature loadings into prices. Prices can change if the valuation function changes and if the representation changes. If investors misrepresent a firm, changes in its representation may be predictable.

Appendix B includes full derivations of the equations presented in this framework.

1.1. Representations, Valuation Functions, and Prices

I assume the stock price of firm s at time t depends on its representation $\boldsymbol{x}_{s,t} \in \mathbb{R}^K$, an aggregate valuation function $\boldsymbol{v}_t \in \mathbb{R}^K$, and an idiosyncratic component $\eta_{s,t} \sim N(0, \sigma_\eta^2)$. The stock price can change if the aggregate valuation function changes, and if the representation of firm s changes.

Representations and prices. The stock price of firm s at time t is

$$P_{s,t} = \boldsymbol{v}_t \cdot \boldsymbol{x}_{s,t} + \eta_{s,t} \tag{1}$$

The **representation** $\boldsymbol{x}_{s,t}$ formalizes how the market reasons about the business model of firm s. It is a vector of intensities across economic features of firms, like *"software"* or *"cars."* A firm's representation can be influenced by how the market categorizes the firm (e.g. Mullainathan, 2002), and which of the firm's features the market pays greater attention to (e.g. Bordalo et al., 2024). As there are many features of firms, and the set of relevant features in the economy can evolve, I will not pre-impose features in my empirical analysis. Instead, I will learn the representation vector from language data.

The **valuation function** $\boldsymbol{v}_t$ formalizes how the market values a given business model. It maps representations to prices, and depends on cash flow and return expectations across features of firms. The market may value an "internet-of-cars company" differently from how it values a "metal bender like everybody else."

The idiosyncratic component $\eta_{s,t}$ corresponds to other drivers of the stock price that are independent of the representation and valuation function. For example, a firm's stock price could be affected by idiosyncratic demand shifters, like trading flows from rebalancing.

Appendix B.1 presents a simple microfoundation for this equation. Investors represent each firm using a collection of features. Prices depend on the representation, the distribution of payoffs from features of the representation, and investor preferences. The derivation in Appendix B.1 shows how representations can influence cash flow expectations and return expectations, which then influence prices.

Changes in prices. The change in stock price of firm s from time t to $t+1$ is

$$\Delta P_{s,t+1} = \overbrace{\Delta \boldsymbol{v}_{t+1} \cdot \boldsymbol{x}_{s,t}}^{\text{valuation function change}} + \overbrace{\boldsymbol{v}_{t+1} \cdot \Delta \boldsymbol{x}_{s,t+1}}^{\text{representation change}} + \varepsilon_{s,t+1} \tag{2}$$

where $\Delta \boldsymbol{v}_{t+1} \equiv \boldsymbol{v}_{t+1} - \boldsymbol{v}_t$ and $\Delta \boldsymbol{x}_{s,t+1} \equiv \boldsymbol{x}_{s,t+1} - \boldsymbol{x}_{s,t}$ are changes in the valuation function and representation.[15]

Valuation function change component. The stock price of firm s can change if the aggregate valuation function changes. For example, if there are changes in consumer demand for cars, or changes in productivity of airbag manufacturers, the stock prices of automakers can change. Moreover, if there are changes in commodity prices or changes in expectations about climate policy, the stock prices of oil drillers can change.

This component can be interpreted as the return on a vector of characteristics or factor loadings. Returns explained by this component correspond to changes in expected cash flows or discount rates from the features $\boldsymbol{x}_{s,t}$. The change in valuation function $\Delta \boldsymbol{v}_{t+1}$ can be interpreted as the time-varying compensation for these features.

Representation change component. The stock price of firm s can also change if its representation changes. A firm may be represented like a car company in one period, an software company in another, and an AI company in the next. Changes in representation can be fundamental—the market may update its representation one-

[15] The price change decomposition follows from

$$\begin{aligned} P_{s,t+1} - P_{s,t} &= \boldsymbol{v}_{t+1} \cdot \boldsymbol{x}_{s,t+1} + \eta_{s,t+1} - \boldsymbol{v}_t \cdot \boldsymbol{x}_{s,t} - \eta_{s,t} \\ &= \boldsymbol{v}_{t+1} \cdot \boldsymbol{x}_{s,t+1} + \eta_{s,t+1} + (\boldsymbol{v}_{t+1} \cdot \boldsymbol{x}_{s,t} - \boldsymbol{v}_{t+1} \cdot \boldsymbol{x}_{s,t}) - \boldsymbol{v}_t \cdot \boldsymbol{x}_{s,t} - \eta_{s,t} \\ &= (\boldsymbol{v}_{t+1} - \boldsymbol{v}_t) \cdot \boldsymbol{x}_{s,t} + \boldsymbol{v}_{t+1} \cdot (\boldsymbol{x}_{s,t+1} - \boldsymbol{x}_{s,t}) + (\eta_{s,t+1} - \eta_{s,t}) \\ &= \Delta \boldsymbol{v}_{t+1} \cdot \boldsymbol{x}_{s,t} + \boldsymbol{v}_{t+1} \cdot \Delta \boldsymbol{x}_{s,t+1} + \varepsilon_{s,t+1} \end{aligned}$$

This is a stock-level decomposition with similar structure to the Laspeyres and Paasche aggregate price level decompositions, and the Kitagawa–Oaxaca–Blinder group wage difference decomposition. Following the typical characteristics model specification in asset pricing, year t is used as the base year for the representation.

for-one as the business of firm s changes. Changes in representation can also be non-fundamental—the market may grow to misunderstand the business of firm s, or to correct its misunderstanding, which lead to changes in representation that are not one-for-one with changes in the true business.

This component can be interpreted as the return that coincides with changes in characteristics or factor loadings. As a firm's characterization changes, the market's expectations about the firm—and the firm's ensuing valuation—may also change. This component reflects an economic mechanism that further explains the change in price; it includes the time $t+1$ information in $\Delta \boldsymbol{x}_{s,t+1}$. To forecast the representation change $\Delta \boldsymbol{x}_{s,t+1}$ using information available at time t, we must use additional conditioning information z_t to construct the expected change in representation $\mathbb{E}_t[\Delta \boldsymbol{x}_{s,t+1} \mid z_t]$.

1.2. Misrepresentation and Predictable Changes in Representation

Suppose the true business model of firm s corresponds to a vector of **fundamental features** $\boldsymbol{x}^*_{s,t}$. Firm s is misrepresented when its representation $\boldsymbol{x}_{s,t}$ differs from its fundamental features $\boldsymbol{x}^*_{s,t}$. For example, a firm is misrepresented if investors represent it more intensely across the *"software"* feature than its fundamentals would imply. A firm that becomes misrepresented can become mispriced. As this misrepresentation is corrected, predictable changes in representation could coincide with predictable stock returns.

Misrepresentation and predictable reversion. If the market misrepresents a firm, the deviation of its representation from the historical mean can forecast the future change in its representation.

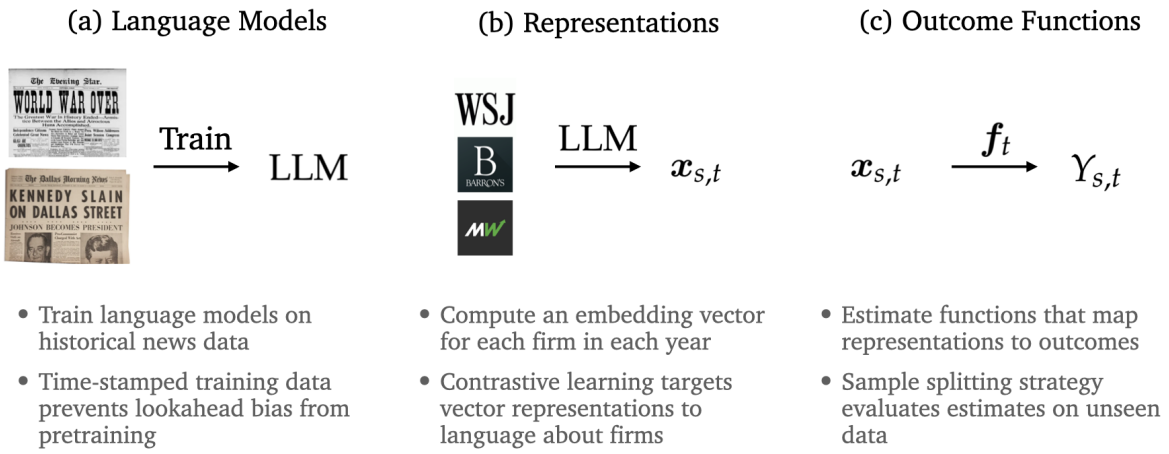
I assume that when the representation $\boldsymbol{x}_{s,t}$ deviates from the historical mean $\overline{\boldsymbol{x}}^h_{s,t} \equiv \frac{1}{h}\sum_{k=1}^{h} \boldsymbol{x}_{s,t-k}$, the expected future change in representation is

$$\mathbb{E}_t[\Delta \boldsymbol{x}_{s,t+1} \mid \boldsymbol{x}_{s,t} - \overline{\boldsymbol{x}}^h_{s,t}] = -\boldsymbol{\beta} \odot \overbrace{(\boldsymbol{x}_{s,t} - \overline{\boldsymbol{x}}^h_{s,t})}^{\text{deviation}} \tag{3}$$

where $\odot$ is the element-wise product, and a fixed vector $\boldsymbol{\beta} > \mathbf{0}$ describes the magnitude of the reversion.

Appendix B.2 presents a simple microfoundation for this expression. If fundamental features $\boldsymbol{x}^*_{s,t}$ follow a random walk, and misrepresentation is corrected over time, representation changes are predictable according to Equation (3). The optimal history

Figure 1: Summary of empirical procedures.



Notes: This figure summarizes the empirical procedures in this paper. First, I train language models on historical newspaper data—this avoids lookahead bias, which can be an issue with off-the-shelf language models. Second, I develop an algorithm to transform financial news language into vector representations for each firm in each year. Third, I estimate functions that relate representations to a series of outcome variables using a split-sample estimation strategy.

horizon h^* trades off the variance reduction in misrepresentation against the variance increase in fundamentals as the history horizon increases.

Appendix B.3 discusses the connections between misrepresentation and theories of economic behavior that make predictions about representations, including categorization, selective attention, and analogy. Appendix B.4 contrasts the predictions of theories of misrepresentation with the predictions from Bayesian learning theories. To empirically evaluate the predictions in this section, I compute representations of firms and study the relationships between representations and valuations.

2. Measuring Representations from Language

I structure the language in each firm's news coverage into an embedding vector. I use this vector representation to estimate a series of outcome variables.

First, I train a series of foundation models for social science research. A foundation model is a machine learning model trained on a large, general-purpose dataset, which can then be fine-tuned on data in a specialized domain. Training these foundation models allows me to avoid two issues with many off-the-shelf foundation models currently applied in social science research: Lookahead bias from pretraining, and non-replicability from model updates. I have released these models to the research community.

Second, I fine-tune the foundation models on financial news language using contrastive representation learning. Contrastive learning is a training procedure that can produce embedding vectors targeted to a specialized domain (Dell, 2024). I adapt a contrastive training objective to develop an embedding algorithm for firms. Using this embedding algorithm, I compute an embedding representation for each firm in each year.

Third, I estimate functions that relate representations to a series of outcome variables. I use a split-sample approach: I estimate parameters on one subset of firms, and apply each estimated function to a disjoint subset of firms. Because of this split-sample strategy, estimation using a high-dimensional embedding vector does not mechanically lead to higher explanatory power.[16]

2.1. Foundation Models for Social Science

Using historical news data, I train a series of general-purpose language models and a representation model. This subsection contains a high-level summary of the training procedures, and Appendix C.2 includes more details.

Training time-stamped language models to avoid lookahead bias. Many applications of off-the-shelf language models to empirical analysis can be subject to a form of lookahead bias (Glasserman and Lin, 2024; Sarkar and Vafa, 2024). This bias occurs when the language model's pretraining data and the analysis period overlap, and can lead information about the future to leak into analysis that should only use data from the past. It is crucial to avoid lookahead bias in this paper as my empirical analysis includes forecasting.

In addition, many providers of language models continually update these models. This can lead to replicability issues. For example, Chen et al. (2023) find that ChatGPT's performance on machine learning benchmarks changed from one month to the next. The structure of this change was not consistent—ChatGPT performed better on a visual reasoning benchmark but worse on a math benchmark. These results suggest that using continually-updated models like ChatGPT for research may not produce replicable

[16] In the cross section, sample splitting does not generate independent holdout samples. This is because returns and other outcome variables are correlated within each cross section. The goal of this estimation strategy is not to evaluate explanatory power on a holdout sample, but to evaluate how well these outcome variables can be explained using information in embeddings. With sample splitting, more parameters in the explanatory variable do not mechanically generate a higher R^2. The explanatory analyses in Section 3 and Section 4 are split-sample tests, and the forecasting analyses in Section 5 are out of sample tests.

results.

To avoid these issues, I train a new foundation model family for social science (StoriesLM). I train the family of models sequentially using masked language modeling (Devlin et al., 2019) on the American Stories dataset (Dell et al., 2024). Because I save a snapshot of the model in each year of the training data, the model has a "time subscript"—researchers can choose a snapshot of the model trained on data up to the time period of interest.

The StoriesLM model is a foundation language model that can be applied to English-language computational tasks. Because of each model snapshot's time subscript, any forecasting analysis that considers events after the snapshot's training data avoids lookahead bias from pretraining. In addition, because each model snapshot's weights are fixed, the model can be applied to produce replicable results.

I include more details on training in Appendix C.2.

Training a language model to produce embedding vectors. I further train the StoriesLM model so that it produces vector representations that encode English semantics. I train this model (RepresentLM) using contrastive learning (e.g., Wang and Isola, 2022; Reimers and Gurevych, 2019), which targets the embeddings to a semantic similarity objective.

I train the model on the Headlines dataset (Silcock et al., 2024), using language from 1920–1979. Over this period, local newspapers would source the content of many of their articles from newswires like the Associated Press, but would produce their own headlines for each article. The Headlines dataset contains multiple headlines that are matched to the same underlying article. As each matched pair of headlines refers to the same article, the headline pair can be used as a positive example for training a semantic similarity model. Given the dataset's structure, I train the model using a multiple negatives ranking loss function (Henderson et al., 2017) in which each matched headline pair is a positive example. The RepresentLM model produces embeddings that apply quantitative structure to language data.

I include more details on training in Appendix C.2. I have released both these models to the research community (Sarkar, 2024a,b).

2.2. Computing Vector Representations of Firms

The language used to discuss a firm can reflect how the market reasons about the firm. An article from 2021 that reads *"Tesla is not a car company—it's an Internet-of-cars company"*

conveys a different view of the firm from an article from 2023 that reads *"[Tesla is] a metal bender like everyone else."* To systematically measure patterns in language that discusses firms, I compute vector representations for each firm in each year using the Dow Jones Newswires, an archive of financial news that includes sources like the Wall Street Journal, Barron's, and MarketWatch.

A key assumption in this paper is that representations measured from financial news language can help to understand how the market represents firms. Should we expect how the news represents a firm to relate to how investors represent the firm? Since both journalists and investors follow the market, we may expect them to represent firms in common ways (Tetlock, 2015). In addition, the press has incentives to match how its readers think (Mullainathan and Shleifer, 2005a; Gentzkow and Shapiro, 2010). If investors represent Tesla similarly to how they represent technology firms, they may demand the press covers Tesla more like a technology firm. Moreover, the content of the Newswires is widely followed.[17] Many of its publications are written to be consumed by investors and may be an important source of information for these investors.

I develop an embedding algorithm targeted to language about firms. General-purpose language models may miss features of language in specific economic domains (Dell, 2024). For example, "platform," "sharing," and "two-sided" are more similar in financial language than they are in general English. Analogously, "return," "recur," and "recover" are more similar in general English than they are in financial language. To compute representations of firms, I use contrastive learning to fine-tune the RepresentLM on financial news language.

Financial news dataset. To train the representation model and compute embeddings, I use Newswires data from June 1979 to May 2022. For multiple articles that are "chained" to one news event, I use the first article in the chain. I filter out articles that contain mainly tabular data and firm filings. I use articles the Newswires codes as being about one stock, and match these articles to US common stocks and stock characteristics using the Jensen et al. (2023) dataset. I restrict the news dataset to articles from firms with non-missing market equity, book equity, 60-month market beta, asset growth, gross profits to assets, and dividends-to-assets in December of year t, and at least 10 articles in year t. The news dataset contains 4,504,121 articles covering 11,287 firms.

[17] For example, the Wall Street Journal reported 3.4 million digital subscribers in 2023. Link to filing.

Training a representation algorithm for firms. The RepresentLM model has been trained to produce semantic representations of English language sequences. The joint distribution of financial news language, however, is different from the joint distribution of general English. Some features may be important for distinguishing between English phrases, but are not key for distinguishing between firms. Other features may be particularly important for distinguishing between firms—for example a firm's focus on technological innovation—but not especially common in general English. To compute embedding vectors that are more targeted to features of firms, I develop an algorithm to learn representations of firms.[18]

In each year t, I learn an embedding algorithm $e(\cdot;\theta_t)$ that maps word sequences to vectors, parameterized by θ_t. The goal of the training procedure is to embed more similar economic language more closely in the representation space. Consequently, I adapt the multiple negatives contrastive training objective (Henderson et al., 2017) to treat each firm as a semantic object. I set the objective to generate embeddings that are more similar for a random pair of articles about the same firm than a random pair of articles about different firms. A firm is an economic unit—the market should on average find two sources of language about the same firm more similar than two sources of language about a random pair of firms.

Each input into the training procedure is the headline and body of an article about a firm.[19] Each training batch in year t takes as input a set of article pairs $\{(a_i, b_i)\}_{i=1}^{n}$ from that year. For every firm i, a_i and b_i are a "positive pair" of articles that refer to the same firm. For every $j \neq i$, a_i and b_j is a "negative pair" of articles that refer to different firms. The output of the procedure is an embedding algorithm $e(\cdot;\theta_t)$ for each year t.

In each year t in each training batch, the objective is

$$\min_{\theta\in\Theta} -\frac{1}{n}\sum_{i=1}^{n}\log\frac{\exp\left\{\psi\left(e(a_i;\theta),e(b_i;\theta)\right)/\tau\right\}}{\sum_{j=1}^{n}\exp\left\{\psi\left(e(a_i;\theta),e(b_j;\theta)\right)/\tau\right\}}$$

which optimizes a neural network $e(\cdot;\theta)$, and depends on the cosine similarity function $\psi(\boldsymbol{x}_1,\boldsymbol{x}_2) = \frac{\boldsymbol{x}_1\cdot\boldsymbol{x}_2}{||\boldsymbol{x}_1||\times||\boldsymbol{x}_2||}$ and a scaling parameter τ. Following Reimers and Gurevych (2019), I use $\tau = 20$. Given GPU memory limits, training proceeds in batches of size

[18] This procedure can also be applied to learn representations of other economic objects with appropriate matched language data.

[19] I separate the headline and body with two newline characters. To fit the context window of the model, I truncate each input to 512 tokens during the training process. A token is a unit of text input processed by the model—the model's tokenizer vocabulary splits text input into word, subword, and punctuation tokens.

$n = 16$, for one full pass across the firms in each year. The embedding dimension is $K = 768$, which the standard dimension for BERT (Devlin et al., 2019) models. I train an embedding algorithm for every year from 1979–2022.

Computing firm-level representations. The embedding algorithm is trained to produce article-level embeddings. For the analysis in this paper, I aggregate these article-level embeddings to compute a firm level representation $\boldsymbol{x}_{s,t}$ for firm s in year t. For every analysis indexed by t, I lag the embedding algorithm by year. This means that for every analysis indexed by time t, the representations $\{\boldsymbol{x}_{s,t-5}, \ldots, \boldsymbol{x}_{s,t-1}, \boldsymbol{x}_{s,t}, \boldsymbol{x}_{s,t+1}\}$ are all computed using the year $t-1$ embedding algorithm. This way, year t information from the training procedure is not used to make predictions about firms in year t.

I construct the firm-level representation using the following procedure. For each article about each firm, I first compute an article-level embedding, where an article is defined as the concatenation of the headline, two newline characters, and the body. If an article is longer than 512 input tokens, I compute embeddings across sequential 512-token chunks of the article and then compute the average embedding of the article. To construct firm-level embeddings, I aggregate article-level embeddings by averaging the article representations for each firm in each month. Averaging embeddings can aggregate multiple language sources about the same entity while preserving entity-specific information (Coleman, 2020). For the analysis in the main text, I then average the monthly embeddings for each firm in each year to produce a firm-by-year representation.

2.3. Estimating Outcomes Using Vector Representations

Throughout this paper, I use representations to estimate a series of outcome variables. I use a split-sample approach.

Sample construction. For the analysis dataset, I match news articles to data on stock returns and firm characteristics for US common stocks. I use the same filters on article chaining and article content—as well as the same non-missing stock characteristics—as the dataset used to train the embedding models and compute representations. As in the Jensen et al. (2023) dataset, the most recent accounting data is incorporated with a four month lag. To keep the sample uniform across analysis that conditions on past embeddings and prices, I further restrict the analysis dataset to firms with at least 10 articles and non-missing book-to-market information over years $t-5$ to $t-1$. I match to return data from CRSP. For some results, I additionally use the historical SIC code from

the Jensen et al. (2023) dataset, earnings as computed by De La O et al. (2024), and the median long-term growth forecast from IBES. I winsorize the long-term growth forecast at the 1% level at both tails. I use the codebook on Ken French's website to map SIC codes into FF12 industries and FF48 sub-industries.

The matched analysis dataset is a firm-by-year panel from 1984–2021 with 89,145 observations. In the return variance analysis in Section 4, I condition on survivorship, available news articles in year $t+1$, and non-missing December characteristics in year $t+1$, which reduces the number of observations to 81,708. I do not enforce this additional condition in the forecasting analysis in Section 5 so that each forecasting regression reflects an implementable trading strategy.

Explanatory variables. I evaluate fits using representations, industry information, characteristics, and combinations of these variables. The representation $\boldsymbol{x}_{s,t}$ is computed from Section 2.2. Industry vectors $\boldsymbol{x}_{s,t}^{\text{FF}}$ are a collection of indicator variables for whether a firm belongs to a Fama–French 12 industry or a Fama–French 48 sub-industry. Characteristics vectors $\boldsymbol{x}_{s,t}^{\text{char}}$ correspond to 60-month CAPM beta, log book equity, one-year asset growth, gross profits to assets, and dividends to assets. In each year, I winsorize the first four characteristics at the 1% level at both tails, and winsorize dividends to assets at 1% at the upper tail. For all analysis that involves estimated functions, I use $\boldsymbol{x}_{s,t}$ to refer to the representation concatenated with a constant, and $\Delta\boldsymbol{x}_{s,t}$ to refer to the change in representation concatenated with a constant. The same convention holds for industry and characteristics vectors. In some analyses, I form explanatory variables by concatenating multiple vectors.

Split-sample procedure for estimating outcome variables. I use the representation $\boldsymbol{x}_{s,t}$ for firm s in year t to estimate outcome variables $Y_{s,t}$. I form these estimates using a split-sample estimation strategy. This approach ensures that direct information about firm s is not used to fit the parameters or tune the hyperparameters that are used to estimate $\widehat{Y}_{s,t}$. I use the same strategy for every set of vectors, but will refer to the explanatory variable as $\boldsymbol{x}_{s,t}$ in this description for brevity.

I form estimates using five-fold cross validation. I construct the five folds by randomly sampling folds at the firm (permco) level—i.e. when a firm enters the dataset, it is "born into" a fold. I denote the set of firms in fold k as $\mathcal{K}_k$, and the set of firms in the fold that contains firm s as $\mathcal{K}_s$.

In each year t and for each firm s, I form an estimate $\widehat{Y}_{s,t} = \widehat{\boldsymbol{f}}_{t,k}^{\text{cv}} \cdot \boldsymbol{x}_{s,t}$ of the outcome

variable by applying a estimated linear function $\widehat{f}^{\text{cv}}_{t,k}$ to its representation $\boldsymbol{x}_{s,t}$. I fit the function using cross-sectional cross-validation—I estimate a function $\widehat{f}^{\text{cv}}_{t,k}$ for each year t and each fold k. For each target fold, I estimate parameters and optimize hyperparameters on the other four folds, and then estimate the outcome on the target fold. I parameterize each function using ridge regression.[20] Throughout this paper, I will use functions $\widehat{f}^{\text{cv}}_{t,k}$ estimated using the cross validation procedure. To economize on notation, I will omit the "cv" superscript and fold subscript—I will refer to each function as $\widehat{f}_t$.

Split-sample evaluation procedure. Given estimates $\widehat{Y}_{s,t}$, I evaluate average goodness-of-fit across the sample splits and years.

In each fold k in each year t, I compute

$$R^2_{\text{k,t}} = 1 - \frac{\sum_{s \in \mathcal{K}_k} (Y_{s,t} - \widehat{Y}_{s,t})^2}{\sum_{s \in \mathcal{K}_k} (Y_{s,t} - \overline{Y}_{s,t})^2}$$

To compute a summary R^2 statistic, I take the average of these R^2 statistics across folds and years

$$R^2 = \frac{1}{5T} \sum_t \sum_k R^2_{k,t}$$

This split-sample test is not a holdout sample test. Because returns and other outcome variables are correlated within a cross section, sample splitting does not generate independent splits of the data. The purpose of sample splitting is to avoid mechanical increases in explanatory power from higher dimensional explanatory variables. The explanatory results evaluate how much of the variation in outcome variables can be attributed to the representations measure. The explanatory analyses in Section 3 and Section 4 are split-sample tests, and the forecasting analyses in Section 5 are out of sample tests.

3. Representations Help to Explain Valuation Formation

I validate the representations measure by showing it contains information about how the market represents firms. I first show that representations embed features of firms and

[20] For each target fold in each year, I tune the regularization hyperparameter on the four training folds in the year. I use the group ridge penalty from Ignatiadis and Lolas (2021), which I describe further in Appendix C.3. Because the hyperparameter is selected on the training sample in each iteration of cross-validation, there is no leakage from hyperparameter tuning into the estimates.

similarity between firms. I then show that representations explain variation in stock prices and cash flow forecasts. Embedding representations relate to market representations, and could help to explain how the market forms valuations.

3.1. Representations Relate to Economic Features and Similarity

I find patterns in the geometry of representations that correspond to relationships between firms. Directions in representation space relate to features of firms, and geometric proximity in representation space relates to similarity between firms.

Properties of embeddings. I briefly summarize results in computer science and linguistics on properties of embeddings, and discuss how these properties can apply to representations of firms.

In language, a *feature* is an element of meaning (Fromkin et al., 1998, Chapter 4). Under the linear representation hypothesis (Park et al., 2023), some features correspond to directions in embedding space.[21] Mikolov et al. (2013b) analyze a word embedding algorithm $e(\cdot)$ and show it has the property $e(\text{king}) - e(\text{queen}) \approx e(\text{man}) - e(\text{woman})$. This example indicates that the features of *"gender"* (which distinguishes king from queen, and man from woman) and *"royalty"* (which distinguishes king from man, and queen from woman) relate to directions in this word embedding space.[22]

In the economy, *features* organize how the market represents firms. Firms' operations can be categorized and associated across several such features, including *"software," "platform economy," "upmarket."* A firm could be represented faithfully across many of these features—the representation could reflect fundamental information about the firm's business model. A firm could also be misrepresented along some features—for example, a firm could be represented more strongly the *"software"* feature than its fundamentals would imply.

Semantic *similarity* corresponds to similarity in meaning (Jurafsky and Martin, 2024). Semantic similarity is typically computed using geometric proximity in representation space. Bhatia (2017) shows that similarities in representation correspond to associative judgements across a series of scenarios studied in cognitive science research.

[21] The authors of Park et al. (2023) refer to features as concepts. Features can also be represented as polytopes (Park et al., 2024) or circles (Engels et al., 2024) in embedding space.

[22] Levy and Goldberg (2014), Arora et al. (2016) and Allen and Hospedales (2019) formally describe how the objective functions used for training word representation algorithms can lead semantic features to be encoded as directions in representation space. Park et al. (2023) discuss how transformer representations can linearly represent semantic features, and Bricken et al. (2023) and Yun et al. (2023) discuss how transformer representations combine information from different kinds of linguistic features.

The market's perceived *similarity* between firms corresponds to how similarly the market represents their business models. Market participants may believe that more similar firms have similar operations, and are exposed to similar kinds of economic shocks.

Examples of economic features. As is true for the individual elements of vectors learned from principal components analysis, the individual elements of vectors learned from the contrastive learning procedure cannot be interpreted as basic features of the embedded objects. However, under the linear representation hypothesis (Park et al., 2023), some features of the embedded objects can relate to linear projections in the representation space. It is not necessary for every economic feature of firms to be embedded linearly for the representation to be useful for studying valuations. However, exploring whether some features of firms correspond to directions in representation space could help to understand some of the structure of the representations measure.

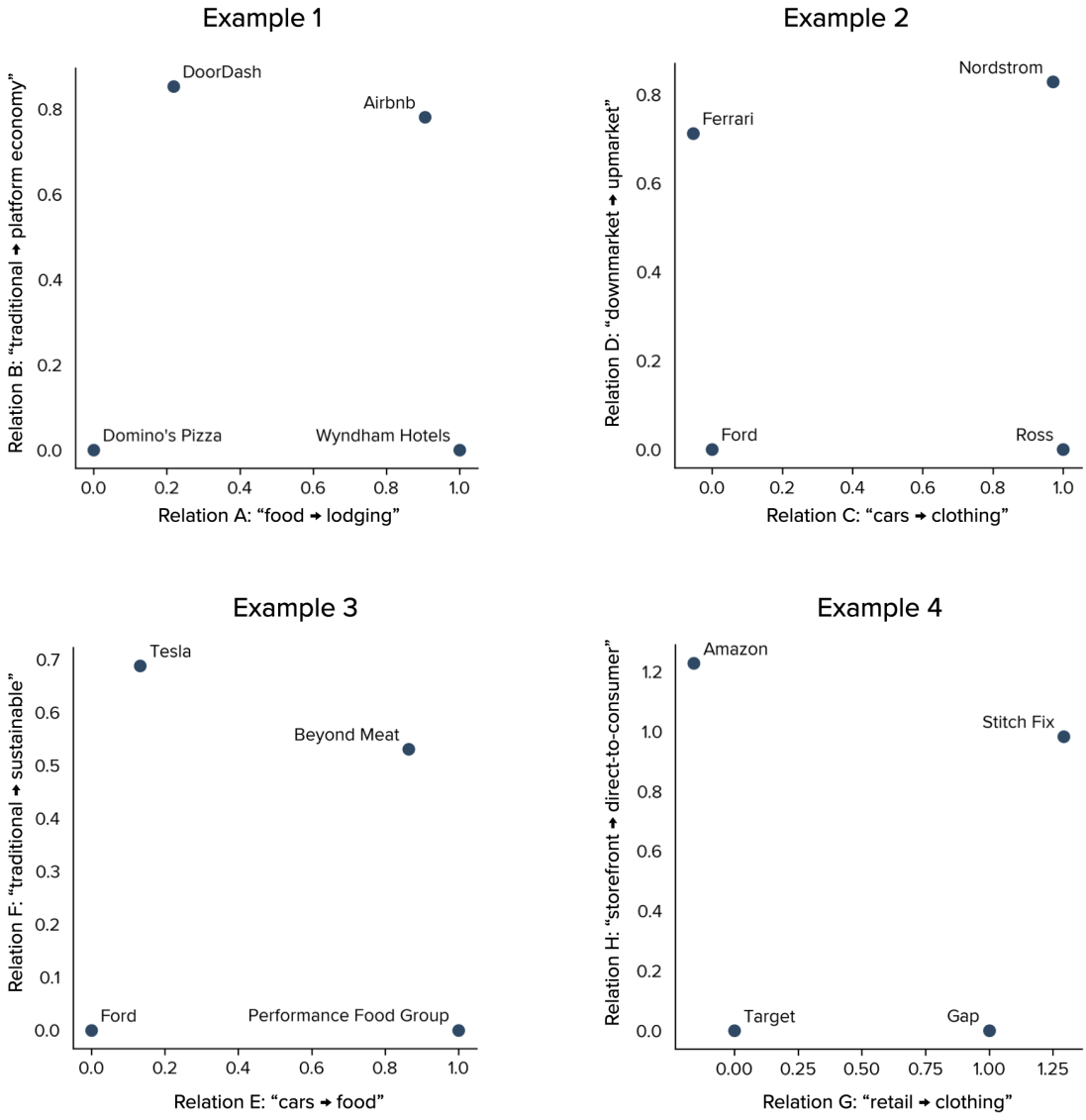
Figure 2 shows examples of economic features of firms that relate to linear projections in representation space. I use the embedding algorithm trained in 2020 to embed several sets of four firms' news coverage in 2021. I project each set of four unit-normed representations onto its first two principal components, and transform these projections so that two target firms span the horizontal axis. Firms that operate in different sectors or industries may still be represented similarly across some features. Some features in this figure, like *"food," "cars," "lodging,"* and *"clothing,"* are often used to structure how the market perceives firms. Other features in this figure, like *"platform economy," "upmarket," "sustainable,"* and *"direct-to-consumer,"* are not traditionally studied drivers of valuation.

Representation similarity relates to market similarity. If two firms are represented similarly, market participants may believe the firms are exposed to similar types of shocks. Investors may therefore be more likely to make similar trading decisions across these firms. Sell-side analysts may also be more likely to cover pairs of firms that are represented similarly. To test these hypotheses, I estimate how representation similarity relates to shared analyst coverage and stock return correlations.

I compute the representation similarity between each pair of firms s and s' in each year using the cosine similarity ψ of their representations.

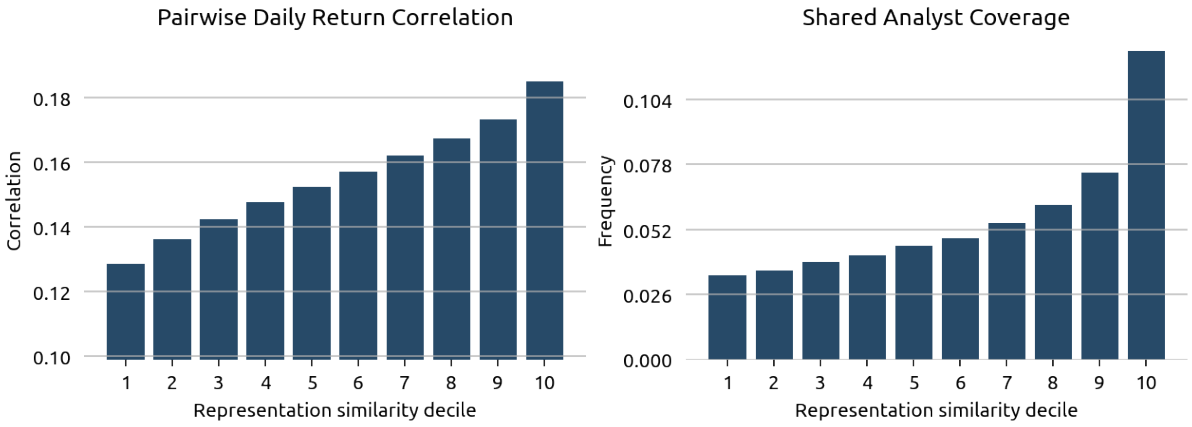
$$\text{Representation Similarity}_{s,s',t} = \psi(\boldsymbol{x}_{s,t}, \boldsymbol{x}_{s',t}) \tag{4}$$

Figure 2: Representations encode relations between firms.



Notes: This figure presents four examples of how representations can encode relations between firms. In each example, I project four representations onto two orthogonal axes. Each axis corresponds to a relation in the language of financial news coverage. For each axis, I use quotation marks to annotate relations between economic features. These annotated features are not part of the representations measure, but describe some of the economic relations between the firms. Some of the features in these annotations, like *"platform economy," "upmarket," "sustainable,"* and *"direct-to-consumer,"* are not traditionally studied drivers of valuations. The market's perceptions of firms along these kinds of features could influence valuations.

Figure 3: Representations encode similarity between firms.



Notes: This figure plots the relationship between representation similarity and standard measures of market similarity. The horizontal axis in each subplot corresponds to the decile of pairwise representation similarity between firms. The vertical axis corresponds to the pairwise daily return correlation or frequency of firms with shared analyst coverage.

To measure pairwise return correlations, I compute the correlation in daily returns of each pair of stocks in year t. To measure shared analyst coverage, I compute an indicator for whether each pair of firms has at least one analyst in common across IBES forecasts in year t.

Figure 3 reports the relationship between representation similarity and the measures of market similarity. Higher representation similarity corresponds to higher pairwise return correlation and higher probability of shared analyst coverage.

3.2. Representations Help to Explain Prices and Cash Flow Forecasts

To further validate the representations measure, I show it helps to explain valuations. I evaluate the split-sample R^2 of estimates of valuation ratios and cash flow forecasts using representations. I find that representations explain a meaningful portion of the cross-sectional variation in these outcome variables, and increase the fit over using only industry information or firm characteristics.

I evaluate how well the estimated outcome $\widehat{Y}_{s,t} = \widehat{\boldsymbol{v}}_t \cdot \boldsymbol{x}_{s,t}$ explains the realized outcome $Y_{s,t}$ by computing the R^2 of

$$Y_{s,t} = \widehat{\boldsymbol{v}}_t \cdot \boldsymbol{x}_{s,t} + \varepsilon_{s,t} \tag{5}$$

Table 1: Representations help to explain prices and forecasts.

R^2: Estimates of price and forecast variables			
Dependent variable $\rightarrow$	$\log(\mathrm{P/B})$	$\log(\mathrm{P/E})$	$\mathbb{F}[\mathrm{LTG}]$
Explanatory variables $\downarrow$	(1)	(2)	(3)
Representation	26.0%	12.1%	19.7%
Industry (FF12)	9.2%	6.8%	12.8%
Sub-Industry (FF48)	11.7%	7.5%	13.6%
Characteristics	12.8%	6.9%	22.2%
Representation+All Others	32.9%	14.5%	30.1%
All Others	19.3%	10.7%	25.7%

Notes: This table reports the R^2 on a series of estimates of stock price and cash-flow forecast variables across sets of explanatory variables. The estimates are from Equation (5): $Y_{s,t} = \widehat{\boldsymbol{v}}_t \cdot \boldsymbol{x}_{s,t} + \varepsilon_{s,t}$, following the split-sample approach of Section 2.3. Each row reports R^2 estimates across a different set of explanatory variables—representations, industry codes, and characteristics, and concatenations of these vectors. Each column reports R^2 estimates across a different price or forecast variable. The set of firms used to fit each estimator does not include the firms used to evaluate the estimates.

I estimate $\boldsymbol{v}_t$ using the split-sample approach from Section 2.3. I compute the average R^2 of each of the five folds across each year. As stock s is not in the data used to learn each valuation function $\widehat{\boldsymbol{v}}_t$, the R^2 in Equation (5) corresponds to goodness-of-fit on unseen data. If a representation $\boldsymbol{x}_{s,t}$ has more parameters than a characteristics vector $\boldsymbol{x}^{\text{char}}_{s,t}$, the representation will not mechanically produce a higher R^2 than the characteristics vector. An increase in R^2 from representations versus characteristics reflects additional information in the representation.

I evaluate the fit of analyses that use the explanatory variables from Section 2.3 to explain the log price-to-book ratio $\log(\mathrm{P/B})_{s,t}$, the log price-to-earnings ratio $\log(\mathrm{P/E})_{s,t}$. and the median long-term growth forecast $\mathbb{F}_t[\mathrm{LTG}_s]$ made at the end of year t. Table 1 reports the R^2 of these estimates. Representations improve the ability to explain prices and forecasts.

4. Representation Changes Help to Explain Stock Price Changes

In my first set of main results, I show that representations help to explain changes in stock prices. I characterize economic mechanisms that could drive these changes. In total, representations help to explain 19% of the cross-sectional variation in annual stock

returns.

I decompose the estimated return into two components. The valuation function change component reflects changes in how business models are valued. The representation change component reflects changes in the perceived business model of the firm. I estimate that two-thirds of the explained variation in cross-sectional returns relates to the valuation function change component, and one-third relates to the representation change component.

I find that the valuation function component increases explanatory power for returns compared to using just stock characteristics and industry information. This result indicates that there is additional information in the representation that helps to explain returns. Adding the representation change component further increases explanatory power. This result indicates that another economic mechanism behind changes in stock prices is the change in the perceived business model.

4.1. Estimating Components of Returns

I estimate the valuation function change and representation change components of the annual return $R_{s,t+1}$. I use a split-sample strategy, so the estimates are not formed using the firm's own return.

Valuation function change component. Recall from Equation (2) that the valuation function change component can be expressed as $[\Delta v]_{s,t+1} \equiv \Delta \boldsymbol{v}_{t+1} \cdot \boldsymbol{x}_{s,t}$. In this expression, $\Delta \boldsymbol{v}_{t+1}$ is the change in valuation function from t to $t+1$.

To estimate $\Delta \boldsymbol{v}_{t+1}$, I use the split-sample procedure from Section 2.3 to estimate the relationship between returns and representations. In each year t, I use the estimated coefficient vector $\widehat{\boldsymbol{\delta}}_{t+1}$ from a ridge regression across stocks $s' \notin \mathcal{K}_s$ not in the fold of stock s

$$R_{s',t+1} = \boldsymbol{\delta}_{t+1} \cdot \boldsymbol{x}_{s',t} + \varepsilon_{s',t+1}$$

to estimate the valuation function change component

$$\widehat{[\Delta v]}_{s,t+1} = \widehat{\boldsymbol{\delta}}_{t+1} \cdot \boldsymbol{x}_{s,t}$$

Representation change component. Recall from Equation (2) that the representation change component can be expressed as $[\Delta x]_{s,t+1} \equiv \boldsymbol{v}_{t+1} \cdot \Delta \boldsymbol{x}_{s,t+1}$. In this expression, $\Delta \boldsymbol{x}_{s,t+1}$ is the change in representation from t to $t+1$.

To estimate $\boldsymbol{v}_{t+1}$, I use the split-sample procedure from Section 2.3 to estimate the relationship between returns and changes in representation. In each year t, I use the estimated coefficient vector $\widehat{\boldsymbol{\nu}}_{t+1}$ from a ridge regression across stocks $s' \notin \mathcal{K}_s$ not in the fold of stock s

$$R_{s',t+1} = \boldsymbol{\nu}_{t+1} \cdot \Delta \boldsymbol{x}_{s',t+1} + \varepsilon_{s',t+1}$$

to estimate the representation change component

$$\widehat{[\Delta x]}_{s,t+1} = \widehat{\boldsymbol{\nu}}_{t+1} \cdot \Delta \boldsymbol{x}_{s,t+1}$$

Total estimated return. To efficiently estimate the total explained return $[\Delta v]_{s,t+1} + [\Delta x]_{s,t+1}$, I apply the estimation procedure to both sets of independent variables jointly. I concatenate the previous two independent variables into $\boldsymbol{c}_{s,t+1} \equiv \begin{bmatrix} \boldsymbol{x}_{s,t} & \Delta \boldsymbol{x}_{s,t+1} \end{bmatrix}$.

In each year t, I use the estimated coefficient vector $\widehat{\boldsymbol{\zeta}}_{t+1}$ from a ridge regression across stocks $s' \notin K_s$

$$R_{s',t+1} = \boldsymbol{\zeta}_{t+1} \cdot \boldsymbol{c}_{s',t+1} + \varepsilon_{s',t+1}$$

to estimate the total estimated return

$$\widehat{[\text{Total}]}_{s,t+1} = \widehat{\boldsymbol{\zeta}}_{t+1} \cdot \boldsymbol{c}_{s,t+1}$$

4.2. Representations Help to Explain Returns

I find that estimated returns explain variation in realized returns. Representations add explanatory power over industry information and characteristics. Accounting for the change in representation adds further explanatory power.

Split-sample cross-sectional test. I compute the R^2 from the specification

$$R_{s,t+1} = \widehat{\mathbb{E}}_t[R_{s,t+1} \mid z] + \varepsilon_{s,t+1} \tag{6}$$

where each $\widehat{\mathbb{E}}_t[\cdot]$ corresponds to a different estimate of the realized return: $\widehat{[\Delta v]}$, $\widehat{[\Delta x]}$, or $\widehat{[\text{Total}]}$. To compute the R^2, I first compute the average R^2 across the folds in each year, and then compute the average of this statistic across all years.

Table 2: Representations help to explain returns.

R^2: Estimates of annual returns			
Estimation strategy $\rightarrow$	$[\Delta v]$	$[\Delta x]$	[Total]
Explanatory variables $\downarrow$	(1)	(2)	(3)
Representation	9.4%	6.8%	13.4%
Industry (FF12)	3.5%	$<$0%	3.5%
Sub-Industry (FF48)	4.1%	$<$0%	4.1%
Characteristics	3.1%	4.4%	8.6%
Representation+All Others	11.5%	10.0%	18.7%
All Others	6.4%	4.4%	11.3%

Notes: This table reports split-sample R^2 statistics of using estimated returns to explain realized cross-sectional returns, following Equation (6): $R_{s,t+1} = \widehat{\mathbb{E}}_t[R_{s,t+1} \mid z] + \varepsilon_{s,t+1}$. Each row reflects a different set of explanatory variables: Representations, industry vectors, characteristics vectors, and combinations of these vectors. Each column reflects a different estimation strategy.

To benchmark these results, I run analogous procedures using industry information and characteristics, as computed in Section 2.3.

Column (1) shows the R^2 of the valuation function change component across these explanatory variables. On their own, representations explain 9.4% of returns. When combined with the traditional characteristics, representations help to explain 11.5% of returns. This is 1.8 times the explanatory power of the traditional characteristics, which explain 6.4% of returns.

Column (2) shows the R^2 of the representation change component across these explanatory variables. Representation changes and characteristics changes explain variation in returns, while industry and sub-industry changes do not. As industry codes are updated only rarely, there is not enough information to explain returns on a separate sample using changes in industry—the split-sample R^2 is negative when using industry information. Representations can measure high-frequency changes in the perceived business model that industry codes do not. The change in representation explains 6.8% of returns on its own. When combined with changes in the other explanatory variables, the change in representation helps to explain 10.0% of returns. This is 2.3 times the explanatory power of changes in the traditional characteristics, which explain 4.4% of returns.

Column (3) shows the R^2 of the estimated total return. Representations explain 13.4% of returns, and 18.7% when combined with industry and characteristics information.

This is 1.7 times the explanatory power of the traditional characteristics, which explain 11.3% of returns.

The valuation function change component $[\Delta v]$ helps to explain returns, and increases explanatory power over traditional characteristics. This result indicates that the representation is more informative of the features that drive stock returns. For example, while changes in auto demand may affect valuations of all automakers, changes in regulatory approval may more strongly affect automakers working on autonomous vehicle, and changes in consumers' discretionary spending may more strongly affect automakers that produce luxury cars. The representation measure contains additional information about the perceived business model.

The representation change component $[\Delta x]$ further explains returns. The additional contribution of the component is $(13.4 - 9.4)/13.4 \approx 30\%$. This result indicates that changes in a firm's perceived business model contribute to changes in valuation. For example, an automaker's valuation may change as its perceived business model changes from car manufacturer to technology developer to battery firm. While it may be challenging to measure these kinds of changes using traditional data, the structure of representations measure helps to estimate changes in perception in more detail.

Alternative explanations for variation explained by the $[\Delta x]$ component. What leads representation changes to explain variation in realized price changes? A natural explanation is that as the market's perception of a firm changes, the firm's valuation changes. An alternative explanation is that information about a firm's valuation that is unrelated to how the market perceives the firm influences the representations measure. In other words, the embedding representation contains information about the price that is not in the market's representation, and the explanatory power of the $[\Delta x]$ component could be driven by this information.

A series of evidence indicates this alternative explanation does not match the data. First, in Section 5.1, I show that time-t information in representations forecasts time-$t+1$ returns. When a firm's representation deviates from the historical mean, the firm's future returns are forecastable. This forecasting power exceeds the forecasting power of the raw deviation in valuation at time t. This result demonstrates that there is information in representations that is not driven by prices. Second, in Appendix D.2, I show that the representation change component does not relate to return shocks that are orthogonal to perceptions. I find that return variation attributable to commodity price shocks does not relate to return variation attributable to representation changes. This result demonstrates

that the representation change component does not correspond to returns driven by shocks that are orthogonal to perceptions.

In addition to these two direct results, the alternative explanation would have to account for the explanatory power of the $[\Delta v]$ component, and the measured relationship between representation similarity and daily return correlations. In summary, these results indicate that the natural explanation for the explanatory power of the $[\Delta x]$ component—when perceptions change, valuations change—better matches the data.

5. Predictable Changes in Representation Indicate Misperception

In my second set of main results, I find that some changes in representation appear non-fundamental, in that they suggest investors have misperceived a firm's business model. I show that when a firm's representation deviates from the historical mean, it predictably reverts. This reversion is associated with predictable returns. I discuss how these results are consistent with misperception of the firm's business model.

To understand what may drive misperception, I measure whether a firm's representation incorporates features that may draw investor attention. I show that after a firm's representation incorporates features associated with trending news coverage or recent extreme performance, it has predictable returns. This evidence suggests that a firm can be misperceived if investors focus on some of its attention-drawing features but neglect its full set of fundamentals.

5.1. Reversion in Representation

I find that when a firm's representation deviates from the historical mean, the representation predictably reverts.

Estimating the deviation in representation. I compute the firm's deviation in representation $\Delta \boldsymbol{x}^h_{s,t} \equiv \boldsymbol{x}_{s,t} - \overline{\boldsymbol{x}}^h_{s,t}$, which is the difference between the firm's representation $\boldsymbol{x}_{s,t}$ and its historical representation $\overline{\boldsymbol{x}}^h_{s,t}$. The historical representation $\overline{\boldsymbol{x}}^h_{s,t} \equiv \frac{1}{h}\sum_{k=1}^{h} \boldsymbol{x}_{s,t-k}$ is the average of the firm's representation over the previous h years.

The deviation in representation $\Delta \boldsymbol{x}^h_{s,t}$ reflects the deviation in features of firm s in year t from their historical level. I use $h = 5$, so that the historical representation does not include representations from several years prior. Equation (12) shows that using too long a history can increase the variance of forecasts, as the firm's fundamentals may have changed over a longer horizon.

Deviations in representation revert. To assess whether these deviations revert, I compute how much the future change in representation $\Delta \boldsymbol{x}_{s,t+1} \equiv \boldsymbol{x}_{s,t+1} - \boldsymbol{x}_{s,t}$ relates to the deviation in representation $\Delta \boldsymbol{x}^h_{s,t}$. Using cosine similarity ψ, I compute

$$\text{Reversion}_{s,t+1} = -\psi(\Delta \boldsymbol{x}_{s,t+1}, \Delta \boldsymbol{x}^h_{s,t})$$

I find that the average reversion is 0.38. A reversion estimate of 1 would mean that representations on average completely revert in the direction of deviation. A reversion estimate of -1 would mean that representations on average continue to deviate in the same direction. The estimate of 0.38 implies that deviations in representation on average revert to the historical mean.

5.2. Reversion in Representation and Returns

I assign a price to the deviation in representation, and find that this priced deviation predicts returns. When investors represent a firm's business model very differently from the past, this deviation appears on average too strong. I find that the unconditional deviation in representation forecasts returns.

To evaluate another mechanism that may influence returns, I assign a price to the deviation in valuation function. When the aggregate valuation function changes, it could both change too much[23] or change too little.[24] I find that the unconditional deviation in valuation function does not forecast returns.

Pricing the deviation in representation. I price the deviation in representation by estimating

$$\text{Representation Deviation (priced)}_{s,t} = \widehat{\boldsymbol{v}}_t \cdot \Delta \boldsymbol{x}^h_{s,t} \tag{7}$$

where $\Delta \boldsymbol{x}^h_{s,t} \equiv \boldsymbol{x}_{s,t} - \overline{\boldsymbol{x}}^h_{s,t}$ is the deviation in representation, and $\widehat{\boldsymbol{v}}_t$ is a function that maps representations to the log price-to-book ratio $\text{pb}_{s,t}$, estimated using the split-sample ridge procedure from Section 2.3.

This measure reflects whether the deviation in representation is toward high-valued

[23]The valuation function could change too much if investors overreact to information about some business models. This mechanism has been studied in conjunction with valuation changes during technology booms (Shiller, 2000).

[24]The valuation function could change too little if investors underreact to information about some business models. This mechanism has been studied in conjunction with the industry momentum effect (Moskowitz and Grinblatt, 1999).

stocks or low-valued stocks. For example, if Tesla's representation deviates to become an "internet-of-cars company," and "internet" has a relatively high valuation, the representation deviation would have a high valuation. If "internet" continues to have a relatively high valuation in the following year, but the representation of Tesla reverts to a "metal bender like everybody else," Tesla would have a lower valuation in the following year.

Pricing the deviation in valuation function. I price the deviation in valuation function by estimating

$$\text{Valuation Function Deviation (priced)}_{s,t} = \widehat{\Delta \boldsymbol{v}_t^h} \cdot \overline{\boldsymbol{x}}_{s,t}^h \tag{8}$$

where $\overline{\boldsymbol{x}}_{s,t}^h$ is the historical representation and $\widehat{\Delta \boldsymbol{v}_t^h}$ is a function that maps the historical representation $\overline{\boldsymbol{x}}_{s,t}^h$ to the deviation in log price-to-book ratio $\text{pb}_{s,t} - \overline{\text{pb}}_{s,t}^h$, estimated using the split-sample ridge procedure from Section 2.3.

Evaluating return predictability. Table 3 reports the results of Fama-MacBeth return forecasting regressions using these two components. The regressions forecast monthly returns, following the Fama-MacBeth specification

$$R_{s,m,t+1} = \alpha_m + \beta_m^T Z_{s,t} + \varepsilon_{s,m,t+1} \tag{9}$$

with respect to sorting variables $Z_{s,t}$. Each sorting variable is re-computed at the end of each year.

Table 3 shows that the priced deviation in representation negatively forecasts returns. When a firm's representation deviates toward high-priced stocks, it has lower returns; when its representation deviates toward low-priced stocks, it has higher returns. This return forecasting ability persists ($t = 3.4$) after conditioning on the valuation function deviation. This result, paired with the result on reversion in representation, suggests the market may misvalue firms if it misrepresents their business models.

In addition, Table 3 shows that the priced deviation in valuation function does not forecast returns. Unconditionally, changes in the market's valuation function do not persist or revert at a one year horizon. However, this is not to say that every movement in valuation functions is one-to-one with fundamentals—it is possible that additional conditioning information could help to measure non-fundamental changes in valuation functions.

Table 3: Deviations in representation forecast returns.

Dependent variable: Monthly return [%]			
	(1)	(2)	(3)
Representation deviation (priced)	-0.45***		-0.47***
	(0.15)		(0.14)
Valuation function deviation (priced)		0.07	0.09
		(0.28)	(0.28)
Forecasting R^2	0.002	0.007	0.009
Months	456	456	456

Notes: This table reports results of Fama–MacBeth forecasting regressions of future returns on priced deviation variables, following Equation (9): $R_{s,m,t+1} = \alpha_m + \beta_m^T Z_{s,t} + \varepsilon_{s,m,t+1}$. Sorting variables are computed at the end of each year, and each sorting variable is cross-sectionally percentile ranked. Fama–MacBeth standard errors are reported in parentheses.

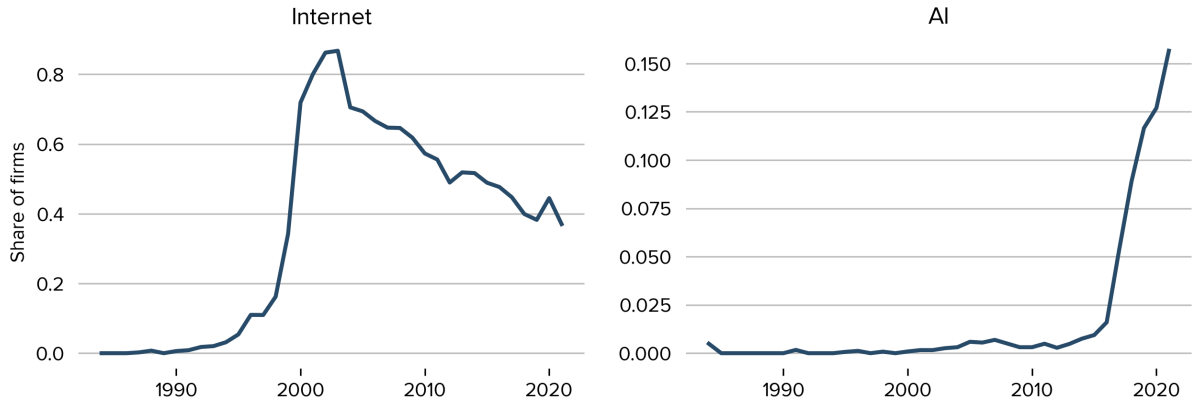
Additional specifications. To benchmark this predictability, I add conditioning variables to the forecasting regression, and conduct portfolio tests.

In a set of robustness regressions, I augment the forecasting regression specification with additional controls. These controls consist of levels and changes in valuation ratios, and past returns. Table E1 reports the results from these regressions. After accounting for these controls, the priced deviation in representation continues to forecast returns. This indicates that the return predictability is distinct from unconditional autocorrelation in prices—the priced deviation in representation has additional forecasting power over returns.

I also construct equal-weighted portfolios that go long the bottom quintile and go short the top quintile of the priced representation deviation. Table E2 reports returns from portfolios across the full sample, and Table E3 reports returns from portfolios that exclude microcap stocks. The long–short portfolio has a monthly return of 0.40% ($t = 3.3$) over the market portfolio, and a monthly return of 0.35% ($t = 3.2$) over the market, SMB, and HML portfolios. The returns on the constructed portfolio correlate with the returns on the HML portfolio. This correlation is natural. The long–short portfolio bets against stocks whose deviation in representation implies a higher price. If misrepresentation leads many stocks to have high prices, the lower returns of these stocks will be correlated with the unconditional lower returns across all high-priced stocks.

I also construct equal-weighted portfolios that go long the bottom quintile and go

Figure 4: Motivating evidence of attention-drawing features: Firm-level word mentions.



Notes: This figure plots the share of firms whose news coverage includes a given word in a given year. The two panels report these statistics for the words "internet" and "AI." Internet began to have a large increase in attention in the late 1990s, and AI began to have a large increase in attention in the late 2010s. During these periods of increases in attention, case study evidence suggests that the market neglected fundamental features of these firms (Cooper et al., 2001; Narayanan and Kapoor, 2024). Figure A2 also includes statistics for "virtual reality," "blockchain," "green," and "wearable."

short the top quintile of the priced valuation function deviation. Returns from these portfolios are also reported in Table E2 and Table E3. As is the case in the forecasting regressions, the valuation function deviation portfolios do not have statistically significant returns. While the unconditional deviation in representation forecasts returns, the unconditional deviation in valuation function does not.

5.3. Motivating Evidence: Attention-Drawing Features, Neglected Fundamentals

What could lead firms to be misrepresented by the market? One reason is that some features may disproportionately draw attention. Distorted attention may lead investors to neglect fundamentals.

As a case study, I first focus on two features—"internet" and "AI"—that have attracted the market's attention. I measure patterns in news coverage associated with these features. I find that many firms' coverage rapidly incorporated these features. Evidence from these episodes suggests that investors' focus on these features led them to neglect some firms' fundamentals.

Rapid increase in coverage. Figure 4 plots the share of firms from 1984–2021 whose coverage included the words "internet" and "AI."[25] Both of these words experienced large coverage increases across a concentrated period. These periods correspond to the dotcom boom in the late 1990s, and the emergence of deep learning in the late 2010s.

Neglect of some firms' fundamentals. Evidence from these episodes suggests that attention to these features led investors to neglect some of these firms' fundamentals.

During the dotcom run-up in the 1990s, Cooper et al. (2001) finds that firms that adopted internet-related names experienced large stock price increases. The authors find that these price increases were unrelated to the firms' involvement with the Internet. During the 2000 crash, these firms' stock prices fell (Glynn and Marquis, 2004). This evidence suggests that as the "internet" feature drew attention, investors neglected other fundamental features of the firms that adopted the "internet" feature.

Over the past few years, many firms have been described as "AI" firms. Not all these descriptions may reflect fundamentals. For example, the venture capital firm MMC Ventures claimed in 2019 that AI was not part of the core business of 40% of European AI startups (Schulze, 2019). Zhang et al. (2024) and Schaeffer et al. (2023) find that many claimed successes from AI may be influenced by misspecified benchmarks. Narayanan and Kapoor (2024) argue that many claims of AI adoption are overstated.

These pieces of evidence suggest that when a firm's representation adopts a feature that draws attention, investors may place too much weight on the feature and neglect the firm's fundamentals.

5.4. Attention-Drawing Features and Returns

I discuss how features associated with trending news coverage or extreme performance may draw attention. I compute features associated with these types of events. I then measure how strongly each firm's representation incorporates these features. I find that these measures of feature incorporation predict returns.

Features associated with trending news coverage. Features that are trending in the news may draw investor attention. Experimental evidence suggests that examples that easily come to mind influence how people make judgments (Tversky and Kahneman, 1973), and empirical evidence suggests that news coverage influences how investors

[25] I use a case-insensitive match for "internet" and a case-sensitive match for "AI."

trade (Barber and Odean, 2008). When seemingly every news article is about AI, we may find it easy to remember stories of how some AI firms have experienced rapid growth.

These stories may not always be the most useful stories for reasoning about the average firm. Representing a customer service firm that adopts a language model like the technology firm that developed the model may be too simple a view of the customer service firm's fundamentals. Consequently, a firm whose representation incorporates a trending feature may be misrepresented and mispriced. If trending features on average command higher prices, a firm whose representation incorporates a trend would on average be overpriced.

I measure trending news coverage using changes in firm-level coverage in the newswires data. I identify firms whose news coverage has increased the most over five years. I take the difference in the log number of news articles between year t coverage and year $t-4$ coverage for each firm s. I compute features associated with trending news coverage, and then measure which firms incorporate these features.

Features associated with extreme performance. Both the features of extreme high performers and the features of extreme low performers may draw the market's attention. Evidence from both the lab and the field suggests that extreme events are more likely to come to investors' minds (Tversky and Kahneman, 1973; Kwon and Tang, 2024). We may find it easy to remember stories like Microsoft's rapid growth from its cloud business, or Lehman Brothers' rapid demise from its high risk exposure.

While we may be more likely to remember these stories, they may not always be the most useful stories for reasoning about the futures of other firms. Representing another firm as the "next Microsoft" or the "next Lehman" may be too simple a view of its fundamentals. Consequently, a firm whose representation incorporates features of extreme high performers may be overpriced, while a firm whose representation incorporates features of extreme low performers may be underpriced.

To to measure past profitability and performance, I use past five-year return on equity and past five-year returns. I compute features associated with high past profitability and performance. I then measure which firms incorporate these features.

Estimating attention-drawing features. I compute the direction in representation space that is associated with a given attention-drawing characteristic—trending news coverage, past profitability, and past stock returns.

Table 4: Incorporation of attention-drawing features forecasts returns.

Dependent variable: Monthly return [%]						
	(1)	(2)	(3)	(4)	(5)	(6)
Incorporate trending features	-0.47***			-0.29**		
	(0.15)			(0.13)		
Incorporate profitable features		-0.50***			-0.33***	
		(0.16)			(0.12)	
Incorporate high-performing features			-0.51***			-0.36**
			(0.18)			(0.15)
Representation deviation (priced)				-0.42***	-0.37***	-0.31***
				(0.13)	(0.12)	(0.11)
Valuation function deviation (priced)				0.09	0.11	0.10
				(0.28)	(0.27)	(0.27)
Forecasting R^2	0.001	0.001	0.002	0.010	0.009	0.010
Months	456	456	456	456	456	456

Notes: This table reports results of Fama–MacBeth forecasting regressions of future returns on feature incorporation variables, following Equation (9): $R_{s,m,t+1} = \alpha_m + \beta_m^T Z_{s,t} + \varepsilon_{s,m,t+1}$. Sorting variables are computed at the end of each year, and each sorting variable is cross-sectionally percentile ranked. Fama–MacBeth standard errors are reported in parentheses.

In each year $t-1$, I estimate

$$Y_{s,t-1} = \boldsymbol{a}_{t-1} \cdot \boldsymbol{x}_{s,t-1} + \varepsilon_{s,t-1}$$

using the penalized estimation procedure from Section 2.3. I estimate the relationship between the representation $\boldsymbol{x}_{s,t-1}$ and the attention-drawing characteristic $Y_{s,t-1}$. The estimated coefficients $\widehat{\boldsymbol{a}}_{t-1}$ correspond to the direction in representation space associated with the characteristic, with shrinkage applied. Higher $\widehat{\boldsymbol{a}}_{t-1}$ implies more exposure to the attention-drawing features.

Estimating feature incorporation. I estimate how much a firm's representation incorporates these attention-drawing features. For each of the three measures of attention, I estimate

$$\text{Incorporation}_{s,t} = \widehat{\boldsymbol{a}}_{t-1} \cdot \Delta \boldsymbol{x}_{s,t}^{h}$$

This measure estimates how much the deviation in representation loads on attention-drawing features from the previous year.

Feature incorporation forecasts returns. I find that incorporation of these features forecasts returns. The more strongly a firm's representation incorporates a trending, profitable, or high-performing feature, the lower its future return.

Table 4 reports the results of Fama-MacBeth return forecasting regressions using the feature incorporation variables. Each of the measures of feature incorporation forecasts returns. Each measure also has additional forecasting power over the priced deviation components from Section 5.2.

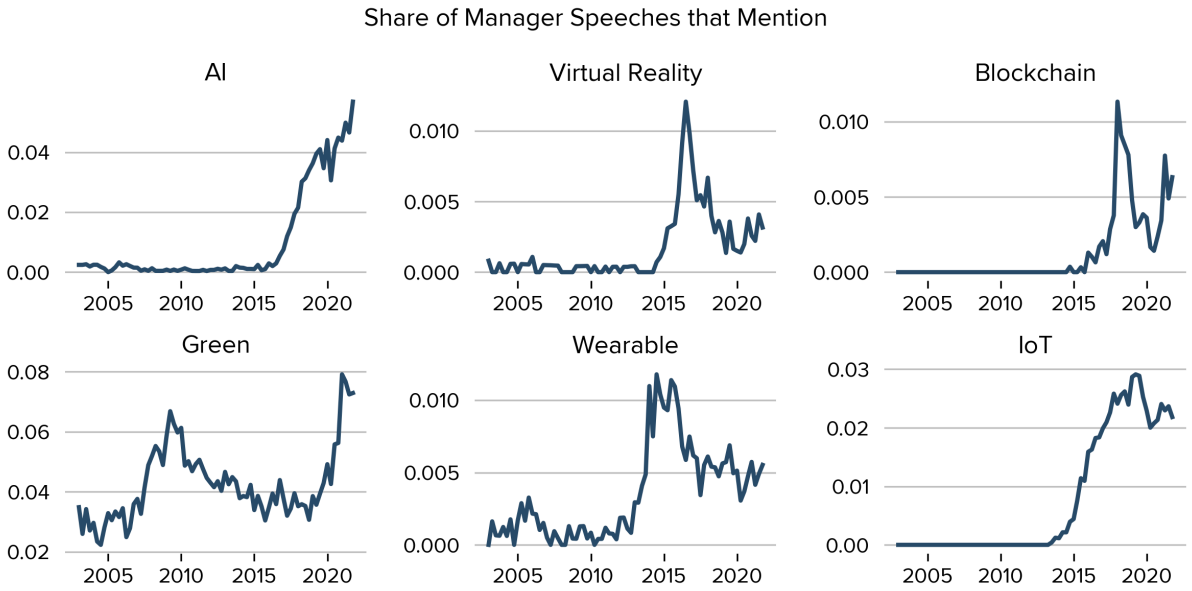
Additional specifications. In a set of robustness regressions, I augment the forecasting regression specification with the additional controls from Section 5.2. These controls relate to valuations and changes in prices. In addition, I run tests that simultaneously condition on all three measures of feature incorporation. Table E4 reports the results from these regressions. After accounting for these controls, the measures of feature incorporation continue to forecast returns.

6. The Role of Communication

What influences how the market perceives firms? To explore one potential mechanism of influence, I analyze communication by firms. I study changes in language used by managers in earnings call speeches. I show that as attention to certain economic features changes, managers' mentions of these features also change. Beyond these individual features, I find that deviations in the overall content of managerial speeches also predictably revert to the historical mean. Furthermore, priced deviations in managerial communication are associated with priced deviations in representations. This evidence suggests that communication by managers may influence how the market perceives firms.

Managerial communication along economic features. I first examine managerial communication across a set of economic features that have attracted attention over the past two decades. I compute the time series of mentions of example economic features in managers' speeches. I use data on earnings calls from StreetEvents from 2003–2021, and use the first call available for each firm in each calendar quarter. I focus on the initial speech segment in each call, which reflects managers' communication to the market before the Q&A period with analysts. I calculate average quarterly mentions of the

Figure 5: Communication across attention-drawing features.



Notes: This figure plots the share of earnings call speeches by managers that include a given phrase in a given quarter. Each of these phrases had a sharp increase in managerial mentions at least once over the sample. For some of these phrases, sharp increases in managerial mentions were followed by sharp decreases in managerial mentions.

phrases "AI," "virtual reality," "blockchain," "green," "wearable," and "IoT."[26]

Figure 5 plots the average mentions of these phrases over the sample. Mentions of "AI" increased sharply in the late 2010s during the emergence of deep learning. This increase coincided with news about AlphaZero and AlphaFold and the release of TensorFlow. It is likely that "AI" mentions have increased even further after 2021.

"Virtual Reality" had a sharp increase and decrease in the late 2010s, around the release of mass-market VR headsets like the Oculus Rift and HTC Vive. "Blockchain" had a sharp increase and decrease in the late 2010s, which coincided with booms in bitcoin and ICOs in this period. "Green" had an increase and decrease in the mid 2000s, around the release of *An Inconvenient Truth* and the clean tech boom. "Green" had another increase during the rise of ESG investing in the late 2010s. "Wearable" had a sharp increase and decrease in the mid-2010s, which coincided with the Fitbit IPO and Apple Watch launch. Finally, "IoT" had a sharp increase and modest decrease in the late 2010s. This period coincided with the spread of commercial internet-of-things devices and consumer devices like the Amazon Echo.

[26] I use case-sensitive matches for "AI" and "IoT," and case-insensitive matches for the other phrases.

For many of these phrases, managerial communication exhibited sharp increases and decreases in phrase frequency. Many of these reversions coincided with technological transformations. These reversions in managerial communication align with the reversions in news coverage in Figure 4 and Figure A2. This evidence suggests that many of these features were not persistent components of these firms' business models.

Deviations in managerial communication predictably revert. Beyond these individual features, I study the overall evolution of managerial communication. Using the RepresentLM model, I compute embeddings of manager speeches, and compute the deviation in these embeddings from the historical mean. Appendix F includes additional details on this procedure. I find that the reversion statistic for manager embeddings is 0.31, which is comparable to the reversion statistic of 0.38 for the representations measure. Like the representations measure, deviations in managerial communication on average revert to the historical mean.

Managerial communication is associated with market representations. I find that deviations in managerial communication are associated with deviations in representations. Using the empirical strategy in Section 5.2, I transform manager embeddings to compute the priced deviation in managerial communication. Appendix F includes additional details on this procedure. Table F1 shows that the priced deviation in managerial communication is associated with the priced deviation in market representations.

What drives the measured relationship between managerial communication and representations? One explanation is that managers may influence the market's perceptions. Managers may supply investors with models for thinking about firms, which influence how investors perceive and value firms. One alternative explanation is that managers may respond to the market's perceptions and change how they perceive or portray their firms. An additional explanation is that managers and the market shift their attention in tandem in response to the economic environment. Across all explanations, the predictable short-run reversion in managerial communication and market representations makes it unlikely that managerial communication is entirely about long-run fundamentals of firms.

What could drive the observed variation in managerial communication? One source could be persuasion and catering by firms. Managers who care about short-run valuations may strategically communicate to the market. If the market does not perfectly understand certain high-priced economic features, and is eager to invest in firms with

those features, managerial communication might emphasize those features. For example, if the market cannot distinguish between firms that develop AI and firms that mention "AI," communication along the "AI" feature could increase a firm's short-run valuation.

Another source of variation could be overoptimism from firms. If managerial communication relates to economic features the market does not perfectly understand, overoptimism could lead to misperception by the market. For example, if firms and the market neglect the effects of competition, they may be too optimistic about the expected firm-level growth associated with some economic features. Competition neglect could affect how the market values firms during technological transformations.

Discussion

People can reason about the same economic object in different ways. As their perceptions change, their valuations may change as well. These perceptions can be challenging to measure using traditional structured data.

Language data could reveal patterns in perception. However, without additional transformation, the structure of raw language data does not lend itself to economic analysis. This paper applies contrastive learning to structure language from financial news into vector representations of firms. These measured representations embed features of firms and similarity between firms, and help to explain stock valuations and cash flow forecasts.

I use representations to characterize mechanisms behind changes in valuations. I show that stock returns relate to both changes in how representations are valued, and changes in representations themselves. Representations help to explain stock returns, and contain additional information beyond traditional characteristics. Changes in representation further increase explanatory power, which suggests that changes in perceived business models contribute to changes in valuations.

Some changes in representation are non-fundamental—investors may come to misperceive some firms' business models. This misperception can lead to misvaluation. Features that draw the market's attention appear to contribute to misperception, and changes in perception are related to changes in managerial communication.

Technological transformations. Technological transformations often coincide with large fluctuations in asset valuations. Influential explanations of these fluctuations often consider a new technology as a whole—and attribute fluctuations in each technology's

overall valuation to variation in sentiment, uncertainty, or other factors (Shiller, 2000; Pástor and Veronesi, 2009; Brunnermeier and Oehmke, 2013).

Another source of valuation fluctuations could be changes in the representations of firms. During a period of technological transformation, the market may reconsider how strongly each firm's business model relates to a new technology. For example, investors may initially believe that a customer service firm that uses a language model is an "AI" firm, in the same sense they believe that the technology firm that developed the language model is an AI firm. Changes in such a firm's characterization could contribute to changes in its valuation. Further analysis of these changes could help to explain patterns in cross-sectional and aggregate valuations during technological transformations.

Valuation by analogy. A person who solves an unfamiliar problem looks for familiar structure (Ross, 1987). Gentner (2003) argues that humans' ability to identify structural similarity is one reason "why we're so smart." Investors commonly use analogical reasoning. For example, in comparables analysis, investors are trained to value a firm based on what they believe to be its set of peers. Reasoning by analogy could help investors learn about the structure of a dynamic environment like the stock market.

While analogical reasoning is a powerful tool, it can distort perceptions when features that generate "surface" similarities are at the center of attention (Holyoak and Koh, 1987). When an analogy is extreme ("Tesla is not a car company") or loads on attention-drawing features ("Tesla is an internet-of-cars company"), investors may not fully appreciate a firm's fundamentals. As new features enter the economy, the focus of investors' reasoning may cycle between surface-level and structural relationships. Empirical studies of analogical reasoning could help to understand how investors perceive and learn about firms.

Representations, decisions, and demand. Many factors that influence decisions are not recorded in traditional structured data. A borrower may consider a credit card's flexible rewards program and colorful card design, while a saver may consider a mutual fund's easy-to-read prospectus and trustworthy brand. A homebuyer may consider a neighborhood's well-maintained streets and friendly residents, while a consumer may consider a cell phone's compatible apps and easy-to-use operating system. A household's macroeconomic representation—how much it considers the high gas prices, rise in automation, closed local shops, and so on—may influence many of its decisions.

The contrastive training procedure in this paper could be applied to learn represen-

tations from language that discusses other kinds of economic objects. These measures could help to further study valuations and decisions.

References

Allen, Carl and Timothy Hospedales, "Analogies Explained: Towards Understanding Word Embeddings," May 2019. arXiv:1901.09813 [cs, stat]. 19

Arora, Sanjeev, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski, "A Latent Variable Model Approach to PMI-based Word Embeddings," *Transactions of the Association for Computational Linguistics*, December 2016, *4*, 385–399. 19

Ashby, F. Gregory and W. Todd Maddox, "Human Category Learning," *Annual Review of Psychology*, February 2005, *56* (1), 149–178. 6

Ashenfelter, Orley and Kathryn Graddy, "Auctions and the Price of Art," *Journal of Economic Literature*, September 2003, *41* (3), 763–786. 1

Ba, Cuimin, J. Aislinn Bohren, and Alex Imas, "Over- and Underreaction to Information," *SSRN Electronic Journal*, 2022. 61

Bajari, Patrick, Zhihao Cen, Victor Chernozhukov, Manoj Manukonda, Suhas Vijaykumar, Jin Wang, Ramon Huerta, Junbo Li, Ling Leng, George Monokroussos, and Shan Wan, "Hedonic Prices and Quality Adjusted Price Indices Powered by AI," April 2023. arXiv:2305.00044 [econ]. 5

Baker, Malcolm and Jeffrey Wurgler, "A Catering Theory of Dividends," *The Journal of Finance*, June 2004, *59* (3), 1125–1165. 7

_ **, Daniel Bergstresser, George Serafeim, and Jeffrey Wurgler**, "The Pricing and Ownership of US Green Bonds," *Annual Review of Financial Economics*, November 2022, *14* (1), 415–437. 7

Barber, Brad M. and Terrance Odean, "All That Glitters: The Effect of Attention and News on the Buying Behavior of Individual and Institutional Investors," *Review of Financial Studies*, April 2008, *21* (2), 785–818. 7, 34

Barberis, Nicholas and Andrei Shleifer, "Style investing," *Journal of Financial Economics*, May 2003, *68* (2), 161–199. 6

_ , _ , **and Jeffrey Wurgler**, "Comovement," *Journal of Financial Economics*, February 2005, *75* (2), 283–317. 7

Bengio, Yoshua, Réjean Ducharme, and Pascal Vincent, "A neural probabilistic language model," *Advances in neural information processing systems*, 2000, *13*. 65

Bergstresser, Daniel and Thomas Philippon, "CEO incentives and earnings management," *Journal of Financial Economics*, June 2006, *80* (3), 511–529. 7

Berry, Steven, James Levinsohn, and Ariel Pakes, "Automobile Prices in Market Equilibrium," *Econometrica*, July 1995, *63* (4), 841. 5

Bhatia, Sudeep, "Associative judgment and vector space semantics.," *Psychological Review*, January 2017, *124* (1), 1–20. 19

Bhojraj, Sanjeev and Charles M. C. Lee, "Who Is My Peer? A Valuation-Based Approach to the Selection of Comparable Firms," *Journal of Accounting Research*, May 2002, *40* (2), 407–439. 5

Black, Fischer, "Noise," *The Journal of Finance*, July 1986, *41* (3), 528–543. 1

Bohren, J. Aislinn, Josh Hascher, Alex Imas, Michael Ungeheuer, and Martin Weber, "A Cognitive Foundation for Perceiving Uncertainty," Technical Report w32149, National Bureau of Economic Research, Cambridge, MA February 2024. 6

Bordalo, Pedro, John Conlon, Nicola Gennaioli, Spencer Yongwook Kwon, and Andrei Shleifer, "How People Use Statistics," Technical Report w31631, National Bureau of Economic Research, Cambridge, MA August 2023. 6, 61

_ **, Nicola Gennaioli, Giacomo Lanzani, and Andrei Shleifer**, "A Cognitive Theory of Reasoning and Choice," Technical Report, Harvard University 2024. 6, 8, 62

_ , _ **, Rafael La Porta, and Andrei Shleifer**, "Finance without exotic risk," Technical Report, National Bureau of Economic Research 2025. 6

Boyer, Brian H., "Style-Related Comovement: Fundamentals or Labels?," *The Journal of Finance*, 2011, *66* (1), 307–332. 7

Bricken, Trenton, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas

Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah, "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning," *Transformer Circuits Thread*, 2023. 19

Brunnermeier, Markus K. and Martin Oehmke, "Bubbles, Financial Crises, and Systemic Risk," in "Handbook of the Economics of Finance," Vol. 2, Elsevier, 2013, pp. 1221–1288. 40

Bryzgalova, Svetlana, Markus Pelger, and Jason Zhu, "Forest Through the Trees: Building Cross-Sections of Stock Returns," *SSRN Electronic Journal*, 2023. 6

Bybee, Leland, Bryan Kelly, and Yinan Su, "Narrative Asset Pricing: Interpretable Systematic Risk Factors from News Text," *The Review of Financial Studies*, November 2023, *36* (12), 4759–4787. 6

_ , _ , Asaf Manela, and Dacheng Xiu, "Business News and Business Cycles," *The Journal of Finance*, October 2024, *79* (5), 3105–3147. 6, 7

Campbell, John, Martin Lettau, Burton Malkiel, and Yexiao Xu, "Idiosyncratic Equity Risk Two Decades Later," Technical Report w29916, National Bureau of Economic Research, Cambridge, MA April 2022. 1, 6

Campbell, John Y, *Financial decisions and markets: a course in asset pricing*, Princeton University Press, 2017. 63

Campbell, John Y. and Robert J. Shiller, "Stock Prices, Earnings, and Expected Dividends," *The Journal of Finance*, July 1988, *43* (3), 661–676. 6

_ , Martin Lettau, Burton G. Malkiel, and Yexiao Xu, "Have Individual Stocks Become More Volatile? An Empirical Exploration of Idiosyncratic Risk," *The Journal of Finance*, February 2001, *56* (1), 1–43. 1, 6

Case, Karl E and Robert J Shiller, "The Efficiency of the Market for Single-Family Homes," *The American Economic Review*, 1989, pp. 125–137. Publisher: JSTOR. 1

Chen, Huaizhi, Lauren Cohen, and Dong Lou, "Industry Window Dressing," *Review of Financial Studies*, December 2016, *29* (12), 3354–3393. 7

Chen, Jiafeng and Suproteem K. Sarkar, "A Semantic Approach to Financial Fundamentals," in "Proceedings of the Second Workshop on Financial Technology and Natural Language Processing" - Kyoto, Japan May 2020, pp. 22–26. 7

Chen, Lingjiao, Matei Zaharia, and James Zou, "How is ChatGPT's behavior changing over time?," October 2023. arXiv:2307.09009 [cs]. 2, 7, 12

Chen, Ting, Calvin Luo, and Lala Li, "Intriguing Properties of Contrastive Losses," October 2021. arXiv:2011.02803 [cs, stat]. 65

Chen, Yifei, Bryan T Kelly, and Dacheng Xiu, "Expected returns and large language models," *Available at SSRN 4416687*, 2024. 7

Cho, Thummim and Christopher Polk, "Putting the Price in Asset Pricing," *The Journal of Finance*, October 2024, p. jofi.13391. 6

Cochrane, John H., "Presidential Address: Discount Rates," *The Journal of Finance*, August 2011, *66* (4), 1047–1108. 6

Cohen, Lauren, Christopher Malloy, and Quoc Nguyen, "Lazy Prices," *The Journal of Finance*, June 2020, *75* (3), 1371–1415. 73

Cohen, Randolph B., Christopher Polk, and Tuomo Vuolteenaho, "The Value Spread," *The Journal of Finance*, April 2003, *58* (2), 609–641. 6

Coleman, Ben, "Why is it okay to average embeddings?," November 2020. Randorithms. 16

Compiani, Giovanni, Ilya Morozov, and Stephan Seiler, "Demand Estimation with Text and Image Data," *SSRN Electronic Journal*, 2024. 5

Cooper, Michael J., Orlin Dimitrov, and P. Raghavendra Rau, "A Rose.com by Any Other Name," *The Journal of Finance*, December 2001, *56* (6), 2371–2388. 4, 7, 32, 33

Da, Zhi, Joseph Engelberg, and Pengjie Gao, "In Search of Attention," *The Journal of Finance*, October 2011, *66* (5), 1461–1499. 7

Damodaran, Aswath, *Investment valuation: tools and techniques for determining the value of any asset* Wiley finance, 3. ed ed., Hoboken, NJ: Wiley, 2012. 5

Daniel, Kent and Sheridan Titman, "Evidence on the Characteristics of Cross Sectional Variation in Stock Returns," *The Journal of Finance*, March 1997, *52* (1), 1–33. 5

Dell, Melissa, "Deep Learning for Economists," July 2024. arXiv:2407.15339 [cs, econ, q-fin]. 2, 12, 14

_ , **Jacob Carlson, Tom Bryan, Emily Silcock, Abhishek Arora, Zejiang Shen, Luca D'Amico-Wong, Quan Le, Pablo Querubin, and Leander Heldring**, "American stories: A large-scale structured text dataset of historical us newspapers," *Advances in Neural Information Processing Systems*, 2024, *36*. 2, 13, 65

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," May 2019. arXiv:1810.04805 [cs]. 13, 16, 65

Didisheim, Antoine, Shikun (Barry) Ke, Bryan Kelly, and Semyon Malamud, "Complexity in Factor Pricing Models," Technical Report w31689, National Bureau of Economic Research, Cambridge, MA September 2023. 6

Dolphin, Rian, Barry Smyth, and Ruihai Dong, "Stock Embeddings: Learning Distributed Representations for Financial Assets," February 2022. arXiv:2202.08968 [q-fin]. 7

Engels, Joshua, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark, "Not All Language Model Features Are Linear," October 2024. arXiv:2405.14860 [cs]. 19

Enke, Benjamin, "The Cognitive Turn in Behavioral Economics," *Technical report*, 2024. 6

_ **and Thomas Graeber**, "Cognitive Uncertainty," *The Quarterly Journal of Economics*, May 2023, p. qjad025. 61

Fama, Eugene F. and James D. MacBeth, "Risk, Return, and Equilibrium: Empirical Tests," *Journal of Political Economy*, May 1973, *81* (3), 607–636. 31, 35, 68, 80, 83, 85

_ **and Kenneth R. French**, "The Cross-Section of Expected Stock Returns," *The Journal of Finance*, June 1992, *47* (2), 427–465. 6

Fedyk, Anastassia and James Hodson, "When can the market identify old news?," *Journal of Financial Economics*, July 2023, *149* (1), 92–113. 7

Flynn, Joel P. and Karthik Sastry, "The Macroeconomics of Narratives," *SSRN Electronic Journal*, 2022. 7

Freyberger, Joachim, Andreas Neuhierl, and Michael Weber, "Dissecting Characteristics Nonparametrically," *The Review of Financial Studies*, May 2020, *33* (5), 2326–2377. 6

Fromkin, VA, Robert Rodman, and V Hyams, *An Introduction to Language 6e*, Orlando, FL: Hartcourt Brace College Publishers, 1998. 19

Fryer, Roland and Matthew O. Jackson, "A Categorical Model of Cognition and Biased Decision Making," *The B.E. Journal of Theoretical Economics*, February 2008, *8* (1). 6

Gabaix, Xavier, "A Sparsity-Based Model of Bounded Rationality," *The Quarterly Journal of Economics*, November 2014, *129* (4), 1661–1710. 6, 61

_ , "Behavioral inattention," in "Handbook of Behavioral Economics: Applications and Foundations 1," Vol. 2, Elsevier, 2019, pp. 261–343. 6

_ **, Ralph S. J. Koijen, and Motohiro Yogo**, "Asset Embeddings," *SSRN Electronic Journal*, 2024. 7

Gao, Tianyu, Xingcheng Yao, and Danqi Chen, "SimCSE: Simple Contrastive Learning of Sentence Embeddings," May 2022. arXiv:2104.08821 [cs]. 65

Gentner, Dedre, "Why we're so smart," in "Language in mind: Advances in the study of language and thought," MIT Press, 2003, pp. 195–235. 6, 40, 62

Gentzkow, Matthew and Jesse M. Shapiro, "What Drives Media Slant? Evidence From U.S. Daily Newspapers," *Econometrica*, 2010, *78* (1), 35–71. 14

Glasserman, Paul and Caden Lin, "Assessing Look-Ahead Bias in Stock Return Predictions Generated by GPT Sentiment Analysis," *The Journal of Financial Data Science*, 2024, *6* (1), 25–42. Publisher: Portfolio Management Research. 2, 7, 12

Glynn, Mary Ann and Christopher Marquis, "When Good Names Go Bad: Symbolic Illegitimacy in Organizations," in "Research in the Sociology of Organizations," Vol. 22, Bingley: Emerald (MCB UP), 2004, pp. 147–170. 33

Golubov, Andrey and Theodosia Konstantinidi, "Where Is the Risk in Value? Evidence from a Market-to-Book Decomposition," *The Journal of Finance*, December 2019, *74* (6), 3135–3186. 6

Gompers, Paul and Josh Lerner, "The Venture Capital Revolution," *Journal of Economic Perspectives*, May 2001, *15* (2), 145–168. 1

Graham, Benjamin and David L. Dodd, *Security analysis*, London: McGraw-Hill, 1934. 5

Green, T. Clifton and Byoung-Hyoun Hwang, "Price-based return comovement," *Journal of Financial Economics*, July 2009, *93* (1), 37–50. 7

Han, Sukjin, Eric H. Schulman, Kristen Grauman, and Santhosh Ramakrishnan, "Shapes as Product Differentiation: Neural Network Embedding in the Analysis of Markets for Fonts," March 2024. arXiv:2107.02739 [econ]. 5

Harris, Lawrence and Eitan Gurel, "Price and Volume Effects Associated with Changes in the S&P 500 List: New Evidence for the Existence of Price Pressures," *The Journal of Finance*, September 1986, *41* (4), 815–829. 7

Hassan, Tarek A, Stephan Hollander, Laurence van Lent, and Ahmed Tahoun, "Firm-Level Political Risk: Measurement and Effects," *The Quarterly Journal of Economics*, November 2019, *134* (4), 2135–2202. 7

Henderson, Matthew, Rami Al-Rfou, Brian Strope, Yun hsuan Sung, Laszlo Lukacs, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil, "Efficient Natural Language Response Suggestion for Smart Reply," May 2017. arXiv:1705.00652 [cs]. 13, 15, 66

Hinton, Geoffrey E, "Distributed representations," *Technical Report*, 1984. Publisher: Carnegie Mellon University. 65

Hirshleifer, David, Sonya S. Lim, and Siew Hong Teoh, "Limited Investor Attention and Stock Market Misreactions to Accounting Information," *Review of Asset Pricing Studies*, December 2011, *1* (1), 35–73. 6

Hoberg, Gerard and Gordon Phillips, "Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis," *Review of Financial Studies*, October 2010, *23* (10), 3773–3811. 7

_ **and** _ , "Text-Based Network Industries and Endogenous Product Differentiation," *Journal of Political Economy*, October 2016, *124* (5), 1423–1465. 7, 73

Holyoak, Keith J. and Kyunghee Koh, "Surface and structural similarity in analogical transfer," *Memory & Cognition*, July 1987, *15* (4), 332–340. 5, 40

Hong, Harrison, Jeremy C. Stein, and Jialin Yu, "Simple Forecasts and Paradigm Shifts," *The Journal of Finance*, June 2007, *62* (3), 1207–1242. 4, 6, 62

Ignatiadis, Nikolaos and Panagiotis Lolas, "σ-Ridge: group regularized ridge regression via empirical Bayes noise level cross-validation," March 2021. 18, 67

Jagannathan, Ravi and Zhenyu Wang, "The Conditional CAPM and the Cross-Section of Expected Returns," *The Journal of Finance*, March 1996, *51* (1), 3. 6

Jegadeesh, Narasimhan and Di Wu, "Word power: A new approach for content analysis," *Journal of Financial Economics*, December 2013, *110* (3), 712–729. 7

Jehiel, Philippe, "Analogy-based expectation equilibrium," *Journal of Economic Theory*, August 2005, *123* (2), 81–104. 6

Jensen, Theis Ingerslev, Bryan Kelly, and Lasse Heje Pedersen, "Is There a Replication Crisis in Finance?," *The Journal of Finance*, October 2023, *78* (5), 2465–2518. 14, 16, 17

Jurafsky, Daniel and James H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd ed. 2024. 2, 19, 64

Kahana, Michael Jacob, *Foundations of human memory*, New York: Oxford university press, 2012. 6

Kapoor, Sayash and Arvind Narayanan, "Leakage and the Reproducibility Crisis in ML-based Science," July 2022. arXiv:2207.07048 [cs, stat]. 7

Ke, Zheng, Bryan T. Kelly, and Dacheng Xiu, "Predicting Returns with Text Data," *SSRN Electronic Journal*, 2020. 7

Kelly, Bryan T., Seth Pruitt, and Yinan Su, "Characteristics are covariances: A unified model of risk and return," *Journal of Financial Economics*, December 2019, *134* (3), 501–524. 6

Koijen, Ralph S. J. and Motohiro Yogo, "A Demand System Approach to Asset Pricing," *Journal of Political Economy*, August 2019, *127* (4), 1475–1515. 5

Kozak, Serhiy and Stefan Nagel, "When Do Cross-sectional Asset Pricing Factors Span the Stochastic Discount Factor?," *SSRN Electronic Journal*, 2023. 6

Kwon, Spencer Yongwook and Johnny Tang, "Extreme Categories and Overreaction to News," *SSRN Electronic Journal*, 2024. 34

Kőszegi, Botond and Adam Szeidl, "A Model of Focusing in Economic Choice," *The Quarterly Journal of Economics*, February 2013, *128* (1), 53–104. 6

Lancaster, Kelvin J, "A new approach to consumer theory," *Journal of political economy*, 1966, *74* (2), 132–157. Publisher: The University of Chicago Press. 5

Levy, Omer and Yoav Goldberg, "Neural word embedding as implicit matrix factorization," *Advances in neural information processing systems*, 2014, *27*. 19

Lewellen, Jonathan, Stefan Nagel, and Jay Shanken, "A skeptical appraisal of asset pricing tests," *Journal of Financial Economics*, May 2010, *96* (2), 175–194. 6

Li, Feng, "Annual report readability, current earnings, and earnings persistence," *Journal of Accounting and Economics*, August 2008, *45* (2-3), 221–247. 7

Liu, Shinhua, "Informational efficiency and GICS classification: Evidence from REITs," *The Quarterly Review of Economics and Finance*, August 2022, *85*, 355–362. 7

Loughran, Tim and Bill McDonald, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *The Journal of Finance*, February 2011, *66* (1), 35–65. 7

_ and _ , "Textual Analysis in Accounting and Finance: A Survey," *Journal of Accounting Research*, 2016, *54* (4), 1187–1230. 73

Magnolfi, Lorenzo, Jonathon McClure, and Alan Sorensen, "Triplet Embeddings for Demand Estimation," *American Economic Journal: Microeconomics*, 2024. 5

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient Estimation of Word Representations in Vector Space," September 2013. arXiv:1301.3781 [cs]. 65

_ , Wen tau Yih, and Geoffrey Zweig, "Linguistic regularities in continuous space word representations," in "Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies" 2013, pp. 746–751. 19

Moskowitz, Tobias J. and Mark Grinblatt, "Do Industries Explain Momentum?," *The Journal of Finance*, August 1999, *54* (4), 1249–1290. 29

Mullainathan, Sendhil, "Thinking Through Categories," *Technical Report*, 2002. 1, 6, 8, 62

_ **and Andrei Shleifer**, "The Market for News," *American Economic Review*, August 2005, *95* (4), 1031–1053. 14

_ **and** _ , "Persuasion in finance," 2005. 7

_ **, Joshua Schwartzstein, and Andrei Shleifer**, "Coarse Thinking and Persuasion," *Quarterly Journal of Economics*, May 2008, *123* (2), 577–619. 6, 62

Narayanan, Arvind and Sayash Kapoor, *AI snake oil: what artificial intelligence can do, what it can't, and how to tell the difference*, Princeton Oxford: Princeton University Press, 2024. 4, 32, 33

Nosofsky, Robert M., "Attention, similarity, and the identification–categorization relationship.," *Journal of Experimental Psychology: General*, 1986, *115* (1), 39–57. 6

O, Ricardo De La, Xiao Han, and Sean Myers, "The Return of Return Dominance: Decomposing the Cross-Section of Prices," *SSRN Electronic Journal*, 2024. 17

Ortoleva, Pietro, "Modeling the Change of Paradigm: Non-Bayesian Reactions to Unexpected News," *American Economic Review*, October 2012, *102* (6), 2410–2436. 6

Park, Kiho, Yo Joong Choe, and Victor Veitch, "The Linear Representation Hypothesis and the Geometry of Large Language Models," November 2023. arXiv:2311.03658 [cs, stat]. 19, 20, 64

_ **,** _ **, Yibo Jiang, and Victor Veitch**, "The Geometry of Categorical and Hierarchical Concepts in Large Language Models," October 2024. arXiv:2406.01506 [cs]. 19

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Andreas Müller, Joel Nothman, Gilles Louppe, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay, "Scikit-learn: Machine Learning in Python," *arXiv:1201.0490 [cs]*, June 2018. arXiv: 1201.0490. 70

Peng, Lin and Wei Xiong, "Investor attention, overconfidence and category learning," *Journal of Financial Economics*, June 2006, *80* (3), 563–602. 6

Pennington, Jeffrey, Richard Socher, and Christopher Manning, "Glove: Global Vectors for Word Representation," in "Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)" Association for Computational Linguistics Doha, Qatar 2014, pp. 1532–1543. 65

Pástor, Lubos and Pietro Veronesi, "Technological Revolutions and Stock Prices," *American Economic Review*, August 2009, *99* (4), 1451–1483. 40

Reimers, Nils and Iryna Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," August 2019. arXiv:1908.10084 [cs]. 13, 15, 65, 66, 71

Rhodes–Kropf, Matthew, David T. Robinson, and S. Viswanathan, "Valuation waves and merger activity: The empirical evidence," *Journal of Financial Economics*, September 2005, *77* (3), 561–603. 6

Roll, Richard, "R2," *The Journal of Finance*, July 1988, *43* (3), 541–566. 1, 6

Rosch, Eleanor H, "Natural categories," *Cognitive psychology*, 1973, *4* (3), 328–350. Publisher: Elsevier. 6

Ross, Brian H., "This is like that: The use of earlier problems and the separation of similarity effects.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, October 1987, *13* (4), 629–639. 6, 40

Sarkar, Suproteem K., "RepresentLM: A Language Model Trained on Historical Semantic Similarity Data," *Language model*, October 2024. 13

_ , "StoriesLM: A Family of Language Models With Time-Indexed Training Data," *SSRN Electronic Journal*, March 2024. 13

_ **and Keyon Vafa**, "Lookahead Bias in Pretrained Language Models," *SSRN Electronic Journal*, 2024. 2, 7, 12

Schaeffer, Rylan, Brando Miranda, and Sanmi Koyejo, "Are Emergent Abilities of Large Language Models a Mirage?," May 2023. arXiv:2304.15004 [cs]. 33

Schulze, Elizabeth, "40% of A.I. start-ups in Europe have almost nothing to do with A.I., research finds," *CNBC*, March 2019. 33

Schwartzstein, Joshua, "Selective Attention and Learning," *Journal of the European Economic Association*, December 2014, *12* (6), 1423–1452. 6

_ **and Adi Sunderam**, "Using Models to Persuade," *American Economic Review*, January 2021, *111* (1), 276–323. 6, 7

Shiller, Robert J., *Irrational exuberance*, Princeton, NJ: Princeton University Press, 2000. 29, 40

_ , *Narrative economics: how stories go viral & drive major economic events*, Princeton: Princeton University Press, 2019. 1, 7

Shleifer, Andrei, "Do Demand Curves for Stocks Slope Down?," *The Journal of Finance*, July 1986, *41* (3), 579–590. 7

Shue, Kelly and Richard R. Townsend, "Can the Market Multiply and Divide? Non-Proportional Thinking in Financial Markets," *The Journal of Finance*, October 2021, *76* (5), 2307–2357. 7

_ **, Richard Townsend, and Chen Wang**, "Categorical Thinking about Interest Rates," *SSRN Electronic Journal*, 2024. 7

Silcock, Emily, Abhishek Arora, and Melissa Dell, "A Massive Scale Semantic Similarity Dataset of Historical English," *Advances in Neural Information Processing Systems*, 2024, *36*. 2, 13, 66

Simon, Herbert A, "Rational choice and the structure of the environment.," *Psychological review*, 1956, *63* (2), 129. Publisher: American Psychological Association. 1

Solomon, David H., "Selective Publicity and Stock Prices," *The Journal of Finance*, April 2012, *67* (2), 599–638. 7

Tetlock, Paul C., "Giving Content to Investor Sentiment: The Role of Media in the Stock Market," *The Journal of Finance*, June 2007, *62* (3), 1139–1168. 7

_ , "Information Transmission in Finance," *Annual Review of Financial Economics*, December 2014, *6* (1), 365–384. 7

_ , "The Role of Media in Finance," in "Handbook of Media Economics," Vol. 1, Elsevier, 2015, pp. 701–721. 14

_ , **Maytal Saar-Tsechansky, and Sofus Macskassy**, "More Than Words: Quantifying Language to Measure Firms' Fundamentals," *The Journal of Finance*, June 2008, *63* (3), 1437–1467. 7

Treisman, Anne M. and Garry Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, January 1980, *12* (1), 97–136. 6

Tversky, Amos and Daniel Kahneman, "Availability: A heuristic for judging frequency and probability," *Cognitive Psychology*, September 1973, *5* (2), 207–232. 33, 34

_ **and** _ , "Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty.," *Science*, September 1974, *185* (4157), 1124–1131. 61

_ **and** _ , "The Framing of Decisions and the Psychology of Choice," *Science*, January 1981, *211* (4481), 453–458. 1, 6

Vafa, Keyon, Emil Palikot, Tianyu Du, Ayush Kanodia, Susan Athey, and David M. Blei, "CAREER: A Foundation Model for Labor Sequence Data," February 2024. arXiv:2202.08370 [cs]. 7

van Binsbergen, Jules H., Martijn Boons, Christian C. Opp, and Andrea Tamoni, "Dynamic asset (mis)pricing: Build-up versus resolution anomalies," *Journal of Financial Economics*, February 2023, *147* (2), 406–431. 6

van Binsbergen, Jules, Svetlana Bryzgalova, Mayukh Mukhopadhyay, and Varun Sharma, "(Almost) 200 Years of News-Based Economic Sentiment," Technical Report w32026, National Bureau of Economic Research, Cambridge, MA January 2024. 7

Wang, Feng and Huaping Liu, "Understanding the behaviour of contrastive loss," in "Proceedings of the IEEE/CVF conference on computer vision and pattern recognition" 2021, pp. 2495–2504. 65

Wang, Tongzhou and Phillip Isola, "Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere," August 2022. arXiv:2005.10242 [cs, stat]. 13, 65

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu,

Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush, "HuggingFace's Transformers: State-of-the-art Natural Language Processing," *arXiv:1910.03771 [cs]*, July 2020. arXiv: 1910.03771. 65

Woodford, Michael, "Modeling Imprecision in Perception, Valuation, and Choice," *Annual Review of Economics*, August 2020, *12* (1), 579–601. 6

Yang, Jeffrey, "A Criterion of Model Decisiveness," *SSRN Electronic Journal*, 2023. 6

Yun, Zeyu, Yubei Chen, Bruno A. Olshausen, and Yann LeCun, "Transformer visualization via dictionary learning: contextualized embedding as a linear superposition of transformer factors," April 2023. arXiv:2103.15949 [cs]. 19

Zhang, Hugh, Jeff Da, Dean Lee, Vaughn Robinson, Catherine Wu, Will Song, Tiffany Zhao, Pranav Raja, Dylan Slack, Qin Lyu, Sean Hendryx, Russell Kaplan, Michele Lunati, and Summer Yue, "A Careful Examination of Large Language Model Performance on Grade School Arithmetic," May 2024. arXiv:2405.00332 [cs]. 7, 33

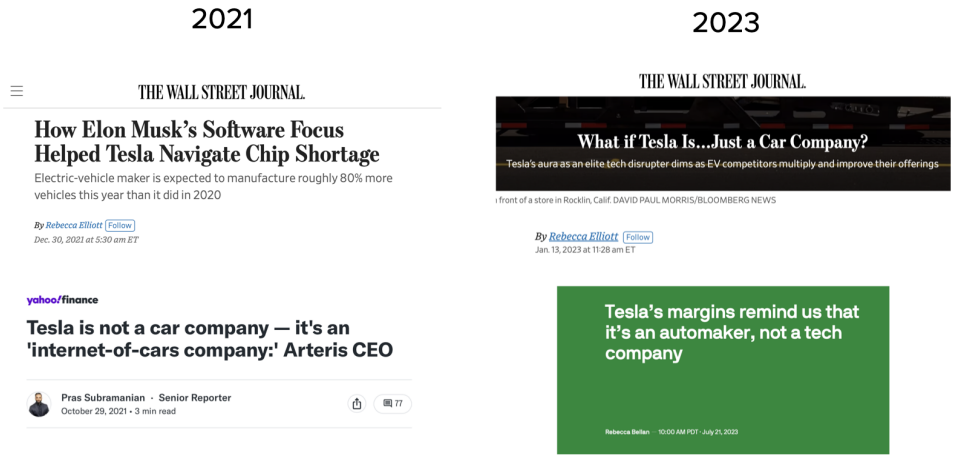
A. News Article Examples and Statistics

I include examples of news coverage and report statistics on word usage in news data.

A.1. Examples

The main text references news coverage of the electric vehicle maker Tesla. Figure A1 presents headlines of news articles about Tesla. I include links to the articles in the figure notes.

Figure A1: Example articles about Tesla.



Notes: This figure shows headlines from four articles about the electric vehicle maker Tesla. The left panel includes an article from the Wall Street Journal from December 30, 2021, and an article from Yahoo Finance from October 29, 2021. The right panel includes an article from the Wall Street Journal from January 13, 2023, and an article from TechCrunch from July 21, 2023.

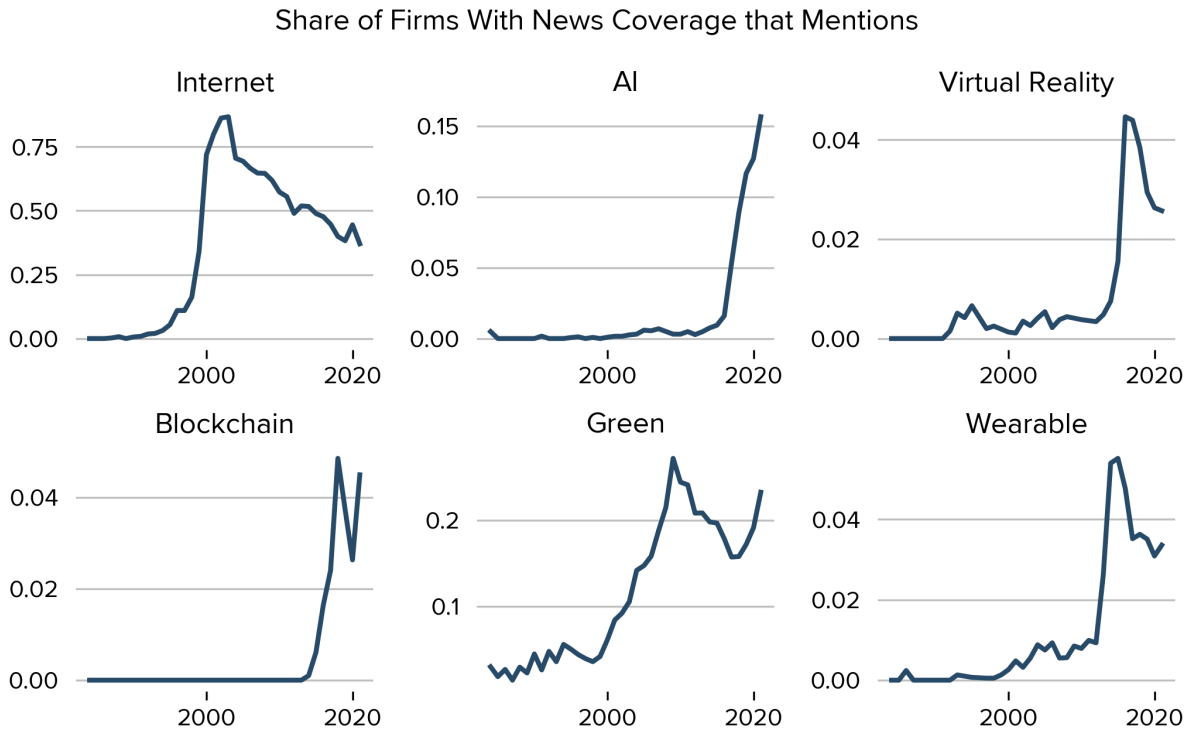
The computation of the decline of "software" mentions in Tesla's news coverage proceeds as follows. I use Factiva to search for Wall Street Journal articles about Tesla in 2021 and 2023. I either run an unconditional search, or a free text search with the term "software." I exclude articles Factiva labels as duplicates.

I count the number of non-duplicate articles in each search. Tesla had 205 unique articles in 2021. 39 of these articles included "software" in a free text search. Tesla had 165 unique articles in 2023. 24 of these articles included "software" in a free text search. The word "software" was included in $39/205 = 19.0\%$ of articles in 2021, and $24/165 = 14.5\%$ of articles in 2023. This reflects a $(19.0 - 14.5)/19.0 \approx 24\%$ decline in relative mentions of "software" over this period.

A.2. Statistics

In each year, I compute the share of firms whose coverage includes a given word or phrase. I use case-sensitive matches for "AI," and case-insensitive matches for the remaining phrases.

Figure A2: Statistics on financial news coverage.



Notes: This figure plots the share of firms whose news coverage includes a given phrase in a given year.

B. Framework Derivations

I derive the equations in Section 1 using simple microfoundations.

B.1. Representations and Prices

I formalize a firm's representation as a vector of intensities across features. Representations and feature-level payoff distributions influence prices.

Setting. A representative investor solves a portfolio choice problem across stocks $s \in [N]$ and a risk-free asset. Each stock s has terminal payoff D_s. The investor makes trading decisions over times $t \in [T]$ before payoffs are realized. The stocks are in fixed supply, denoted by the vector $\boldsymbol{Z}$, and the investor does not learn from prices. The risk-free asset is in zero net supply—its price is normalized to 1 and its gross return is normalized to 0. The investor has CARA utility with risk aversion A and initial wealth W_0. I assume the investor's utility is over terminal wealth, and there are no dynamic trading motives. I denote subjective expectations as $\tilde{\mathbb{E}}$.

The vector $\boldsymbol{x}_{s,t}$ is the investor's representation of each stock s at time t, across K features. The investor believes each stock's payoff is a linear function of its features. At time t, the investor's subjective feature-level payoff distribution is $\tilde{\boldsymbol{d}}_t \sim N(\boldsymbol{\mu}_t, \boldsymbol{\Omega}_t)$. For stock s at time t, the investor's subjective payoff distribution is $\tilde{D}_s = \tilde{\boldsymbol{d}}_t \cdot \boldsymbol{x}_{s,t} + \eta_{s,t}$. This expression combines the representation, feature-level payoff distribution, and an idiosyncratic shifter $\eta_{s,t}$. The idiosyncratic shifter is a known constant to the investor, and is distributed $N(0, \sigma_\eta^2)$ to the analyst. The $N \times K$ matrix $\boldsymbol{X}_t$ stacks representations at time t, the $N \times 1$ vector $\boldsymbol{\eta}_t$ stacks idiosyncratic shifters at time t, and the $N \times 1$ vector $\boldsymbol{D}$ stacks the payoffs across all stocks.

Representations and expectations. The investor's expected payoff for stock s is

$$\tilde{\mathbb{E}}_t[D_s] = \boldsymbol{\mu}_t \cdot \boldsymbol{x}_{s,t} + \eta_{s,t}$$

The investor's expected payoff is a combination of the stock's representation $\boldsymbol{x}_{s,t}$ and expected feature-level payoffs $\boldsymbol{\mu}_t$, as well as the stock-specific shifter $\eta_{s,t}$.

Representations and prices. In each period, the investor solves the portfolio choice problem

$$\max_{\boldsymbol{Q}_t} \ \tilde{\mathbb{E}}_t[-\exp\{-A(W+\boldsymbol{Q}_t\cdot[\boldsymbol{D}-\boldsymbol{P}_t])\}]$$

where $\boldsymbol{Q}_t$ is a vector of portfolio positions and $\boldsymbol{P}_t$ is a vector of prices. The first-order condition implies the investor's vector of positions for the stocks is

$$\boldsymbol{Q}_t = \frac{1}{A}\Sigma_t^{-1}\big(\tilde{\mathbb{E}}_t[\boldsymbol{D}]-\boldsymbol{P}_t\big)$$

where $\Sigma_t = \boldsymbol{X}_t\Omega_t\boldsymbol{X}_t^T$ is the investor's subjective covariance matrix between all stocks' payoffs. Imposing market clearing $\boldsymbol{Q}_t = \boldsymbol{Z}$ and rearranging leads to $\boldsymbol{P}_t = \boldsymbol{X}_t^T\boldsymbol{\mu}_t - A\Sigma_t\boldsymbol{Z}+\boldsymbol{\eta}_t$.

The price of stock s is therefore

$$P_{s,t} = \boldsymbol{v}_t\cdot\boldsymbol{x}_{s,t}+\eta_{s,t} = (\boldsymbol{\mu}_t-\boldsymbol{\lambda}_t)\cdot\boldsymbol{x}_{s,t}+\eta_{s,t} \tag{10}$$

where $\boldsymbol{v}_t$ is a valuation function that maps representations to prices. The valuation function incorporates both the investor's expected feature-level payoffs $\boldsymbol{\mu}_t = \tilde{\mathbb{E}}_t[\tilde{\boldsymbol{d}}_t]$ and the risk adjustment $\boldsymbol{\lambda}_t = A\Omega_t\boldsymbol{X}_t^T\boldsymbol{Z}$.

B.2. Misrepresentation and Reversion in Representation

In a stylized dynamic framework, I discuss how changes in representation can revert if investors misrepresent firms. Deviations between a stock's representation and its historical representation can predict future changes in its representation.

Dynamics of misrepresentation. The representation $\boldsymbol{x}_{s,t}$ of stock s at time t is

$$\boldsymbol{x}_{s,t} = \overbrace{\boldsymbol{x}^*_{s,t}}^{\text{fundamental features}} + \overbrace{\boldsymbol{m}_{s,t}}^{\text{misrepresentation}}$$

In this framework, fundamentals follow a random walk $\boldsymbol{x}^*_{s,t} = \boldsymbol{x}^*_{s,t-1} + \boldsymbol{u}_{s,t}$ with $\boldsymbol{u}_{s,t} \sim N(0, \text{diag}(\boldsymbol{\sigma}^2_u))$.[27] In each period, misrepresentation follows $\boldsymbol{m}_{s,t} \sim N(0, \text{diag}(\boldsymbol{\sigma}^2_m))$.[28]

Predictable reversals in representation. Define the historical representation $\overline{\boldsymbol{x}}^h_{s,t}$ as the mean representation over the past h periods

$$\overline{\boldsymbol{x}}^h_{s,t} \equiv \frac{1}{h} \sum_{k=1}^{h} \boldsymbol{x}_{s,t-k}$$

Claim 1 (Reversion in representation)**.** *Along feature i, the expected change in representation is*

$$\mathbb{E}\Big[[\Delta x_{s,t+1}]_i \mid [\Delta x^h_{s,t}]_i \Big] = -\beta_i \times [\Delta x^h_{s,t}]_i \tag{11}$$

where $\beta_i = \frac{[\sigma^2_m]_i}{\left(1+\frac{1}{h}\right)[\sigma^2_m]_i + \left(\frac{h}{3}+\frac{1}{2}+\frac{1}{6h}\right)[\sigma^2_u]_i} > 0$.

The more the stock's representation changes $[\Delta x^h_{s,t}]_i$ along feature i, the more its representation will revert $[\Delta x_{s,t+1}]_i$ along feature i in the next period. The strength of reversion β_i depends on how much $[\sigma^2_m]_i$ misrepresentation affects feature i, and how much $[\sigma^2_u]_i$ fundamentals tend to change along feature i. The higher the relative variation in misrepresentation $\frac{[\sigma^2_m]_i}{[\sigma^2_u]_i}$ along feature i, the larger the reversion.

Claim 2 (Variance-reducing horizon)**.** *The history horizon h^* that minimizes the conditional variance $\mathbb{V}[[\Delta x_{s,t+1}]_i \mid [\Delta x^d_{s,t}]_i]$ of the representation change forecast across feature i is*

$$h^* = \sqrt{\frac{6[\sigma^2_m]_i + [\sigma^2_u]_i}{2[\sigma^2_u]_i}} \tag{12}$$

When the variance $[\sigma^2_m]_i$ of misrepresentation is high, increasing the horizon reduces the variance of the forecast, as the historical representation is a better estimate of fundamentals. When the variance $[\sigma^2_u]_i$ of fundamentals evolution is high, reducing the

[27] Fundamentals may not follow a random walk if a firm's operations relate to the business cycle. However, business cycle-related reversion in fundamentals may occur at a longer horizon than the horizons studied in this paper.

[28] This formulation assumes that misrepresentation is corrected after one period. Some misrepresentation could be persistent—and may not be completely corrected until after many periods. If misrepresentation is persistent, using a longer history length could reduce the conditional variance of mean reversion forecasts.

horizon reduces the variance of the forecast, as the historical representation of the firm may differ more from its current fundamentals. The minimum-variance horizon trades off these two forces.

Derivation of Claim 1. Note that

$$\Delta \boldsymbol{x}_{s,t+1} = \boldsymbol{u}_{s,t+1} + \boldsymbol{m}_{s,t+1} - \boldsymbol{m}_{s,t}$$

and

$$\begin{aligned}\Delta \boldsymbol{x}^h_{s,t} &= \boldsymbol{m}_{s,t} - \frac{1}{h}\sum_{k=1}^{h} \boldsymbol{m}_{s,t-k} + \boldsymbol{u}_{s,t} + \sum_{k=1}^{h} \boldsymbol{u}_{s,t-k} - \frac{1}{h}\sum_{k=1}^{h}\sum_{i=k}^{h} \boldsymbol{u}_{s,t-k} \\ &= \boldsymbol{m}_{s,t} - \frac{1}{h}\sum_{k=1}^{h} \boldsymbol{m}_{s,t-k} + \boldsymbol{u}_{s,t} + \sum_{k=1}^{h} \frac{h-k}{h} \times \boldsymbol{u}_{s,t-k}\end{aligned}$$

Along feature i, the covariance between the future change and the deviation is

$$\mathrm{Cov}([\Delta x_{s,t+1}]_i, [\Delta x^h_{s,t}]_i) = \sigma^2_{m,i}$$

and the variance of the deviation[29] is

$$\mathrm{Var}([\Delta x^h_{s,t}]_i) = \left(1 + \frac{1}{h}\right)\sigma^2_{m,i} + \left(\frac{h}{3} + \frac{1}{2} + \frac{1}{6h}\right)\sigma^2_{u,i}$$

By normality, the expected change across dimension i is

$$\mathbb{E}[[\Delta x_{s,t+1}]_i \mid [\Delta x^h_{s,t}]_i] = -\frac{\sigma^2_{m,i}}{\left(1 + \frac{1}{h}\right)\sigma^2_{m,i} + \left(\frac{h}{3} + \frac{1}{2} + \frac{1}{6h}\right)\sigma^2_{u,i}} \times [\Delta x^h_{s,t}]_i$$

The more the representation deviates across the feature $[\Delta x^h_{s,t}]_i$, the more its representation will revert across that feature $[\Delta x_{s,t+1}]_i$ in the next period. If some features that are more likely to be affected by misrepresentation (high $\sigma^2_{m,i}$ relative to $\sigma^2_{u,i}$) the representation will revert even more along those features.

[29] The second term in the variance expression follows from

$$1 + \sum_{k=1}^{h} \frac{(h-k)^2}{h^2} = 1 + \sum_{j=0}^{h-1} \frac{j^2}{h^2} = 1 + \frac{(h-1)h(2h-1)}{6h^2} = \frac{h}{3} + \frac{1}{2} + \frac{1}{6h}$$

Derivation of Claim 2. Note that the conditional variance is

$$\begin{aligned}\mathbb{V}[[\Delta x_{s,t+1}]_i \mid [\Delta x^h_{s,t}]_i] &= \mathbb{V}[[\Delta x_{s,t+1}]_i] - \sigma_{12}\sigma_{22}^{-1}\sigma_{21} \\ &= \sigma^2_{u,i} + 2\sigma^2_{m,i} - \frac{(\sigma^2_{m,i})^2}{\left(1+\frac{1}{h}\right)\sigma^2_{m,i} + \left(\frac{h}{3}+\frac{1}{2}+\frac{1}{6h}\right)\sigma^2_{u,i}}\end{aligned}$$

The horizon h^* that minimizes this conditional variance is

$$h^* = \sqrt{\frac{6\sigma^2_{m,i} + \sigma^2_{u,i}}{2\sigma^2_{u,i}}}$$

Note that

$$\begin{aligned}\frac{\partial h^*}{\partial \sigma^2_{m,i}} &= \frac{3}{\sqrt{2\sigma^2_{u,i}(6\sigma^2_{m,i} + \sigma^2_u)}} \\ \frac{\partial h^*}{\partial \sigma^2_{u,i}} &= -\frac{3}{\sigma^2_{u,i}\sqrt{2\sigma^2_{u,i}(6\sigma^2_{m,i} + \sigma^2_{u,i})}}\end{aligned}$$

B.3. Connection to Representation-Based Theories

Many theories of economic behavior make predictions about cognitive representations. Some of the predictions from these theories could be applied to the types of vector representations used in this paper's framework.

Transformation theories (e.g. attention, anchoring, attenuation, and analogy): In some theories, an investor may neglect or otherwise transform some features of a firm. Feature transformation could be modeled as

$$\boldsymbol{x}_s = \boldsymbol{a} \odot \boldsymbol{x}^*_s + (\mathbf{1} - \boldsymbol{a}) \odot \boldsymbol{x}'_s$$

The vector $\boldsymbol{a}$ encodes attention to features and $\boldsymbol{x}'_s$ is a "default." For example, $\boldsymbol{a}$ could result from attention optimization (e.g. Gabaix, 2014) or bottom-up attention (e.g. Bordalo et al., 2023). The default representation $\boldsymbol{x}'_s$ could be an anchor (Tversky and Kahneman, 1974) or an average (Ba et al., 2022; Enke and Graeber, 2023).

Clustering theories (e.g. categorization, prototyping, paradigm shifts): In some theories, an investor may coarsen the representation to a set of discrete categories. This

coarsening could be modeled as

$$\boldsymbol{x}_s = C(\boldsymbol{x}_s^*, w_s)$$

where C is a coarsening function that takes into account additional signals w_s about stock s. These additional signals could be news about an asset that is imperfectly incorporated into the coarsened representations (e.g. Mullainathan, 2002; Hong et al., 2007), or persuasion by a firm (e.g. Mullainathan et al., 2008).

Attention and categorization may be jointly driven by similarity (e.g. Bordalo et al., 2024), which can then influence how the fundamental features $\boldsymbol{x}_s^*$ map into the representation $\boldsymbol{x}_s$.

B.4. Dynamics From Misrepresentation Versus Bayesian Learning

Assume that an investor learns about fundamentals from imperfect signals $\boldsymbol{\xi}_{s,t} = \boldsymbol{x}_{s,t}^* + \boldsymbol{\varepsilon}_{s,t}$ of the fundamental features process $\boldsymbol{x}_{s.t}^* = \boldsymbol{x}_{s,t-1}^* + \boldsymbol{u}_{s,t}$, with $\boldsymbol{u}_t \sim N(0, \text{diag}(\sigma_u^2))$. The representation $\boldsymbol{x}_{s,t}$ reflects the investor's expectation of the fundamental features. At time $t = 0$, the investor's prior belief is $\boldsymbol{x}_{s,0}^* \sim N(\boldsymbol{x}_{s,0}, \text{diag}(\sigma_0^2))$. Future beliefs about fundamentals respond to signals $\boldsymbol{\xi}_{s,t} = \boldsymbol{x}_{s,t}^* + \boldsymbol{e}_{s,t}$ where $\boldsymbol{e}_t \sim N(0, \text{diag}(\sigma_e^2))$. Given this setting, updating follows a Kalman filter. For small t, if the prior is biased ($\boldsymbol{x}_{s,0} \neq \boldsymbol{x}_{s,0}^*$), representations are on average persistent. In this learning setup, if fundamental features follow a random walk, persistence in representation is more likely than reversion in representation.

B.5. Analogical Learning

In her work on analogy, Gentner (2003) argues that a core part of learning is the transition from surface-level to structural similarity assessments.[30] When some features draw attention, investors may focus on surface-level similarities between firms across the attention-drawing features, and neglect fundamental differences between firms across other features. As time progresses, investors may come to better understand structural relationships between existing features, and form more accurate representations of

[30] Gentner (2003) writes: *"Although comparison is an inborn process, its manifestation—whether a sense of sameness is perceived for a given pair of potential analogues—depends on how the situations are represented, and this in turn depends on experience. . . . Early in learning, comparisons are made only between situations that match overwhelmingly. . . . As similarity comparisons evolve from being perceptual and context bound to becoming increasingly sensitive to common relational structure, children show an increasing capacity to reason at the level of abstract commonalities and rules."*

firms. However, as new features enter the economy, investors' attention may again be disproportionately drawn to those new features.

Along these new features, some analogies may be less appropriate than others—for example, a customer service firm that uses a language model may be fundamentally less of an AI firm than the technology firm that builds the language model. If investors focus on the fact that both firms' operations relate to language models, they may consider both to be "AI" firms. In other words, investors may initially draw too strong an analogy between the customer service firm and the broader set of AI firms. Transitions from surface-level to structural assessments could lead to reversions in representation.

B.6. Representations and Present Value Logic

In an alternative derivation, I discuss how the modeled relationship between representations and prices can be expressed in terms of present value logic.

Investors form expectations $\tilde{\mathbb{E}}_t$ over cash flows and returns. Following Campbell (2017), the log price $p_{s,t}$ of stock s at time t is

$$p_{s,t} = \overbrace{\tilde{\mathbb{E}}_t \textstyle\sum_{j=0}^{\infty} \rho^j (1-\rho) d_{s,t+1+j}}^{p^{\text{CF}}_{s,t}} - \overbrace{\tilde{\mathbb{E}}_t \textstyle\sum_{j=0}^{\infty} \rho^j r_{s,t+1+j}}^{-p^{\text{DR}}_{s,t}} + c$$

Assume that these expectations linearly relate to a representation vector $\boldsymbol{x}_{s,t}$

$$p^{\text{CF}}_{s,t} = \boldsymbol{v}^{\text{CF}}_t \cdot \boldsymbol{x}_{s,t} + \omega_{s,t} \quad \text{and} \quad -p^{\text{DR}}_{s,t} = \boldsymbol{v}^{\text{DR}}_t \cdot \boldsymbol{x}_{s,t} + \kappa_{s,t}$$

Then the price can be expressed as

$$\begin{aligned} p_{s,t} &= (\boldsymbol{v}^{\text{CF}}_t - \boldsymbol{v}^{\text{DR}}_t) \cdot \boldsymbol{x}_{s,t} + \eta_{s,t} \\ p_{s,t} &= \boldsymbol{v}_t \cdot \boldsymbol{x}_{s,t} + \eta_{s,t} \end{aligned}$$

The valuation function encodes feature-level expectations of cash flows and returns.

Rearranging like in Section 1.1, the change in price can be expressed as

$$\Delta p_{s,t+1} = \overbrace{\Delta \boldsymbol{v}_{t+1} \cdot \boldsymbol{x}_{s,t}}^{\text{valuation function change}} + \overbrace{\boldsymbol{v}_{t+1} \cdot \Delta \boldsymbol{x}_{s,t+1}}^{\text{representation change}} + \varepsilon_{s,t+1}$$

C. Model Training and Outcome Estimation

I provide background on language embeddings, further details on model training, and further details on outcome estimation.

C.1. Background on Language Embeddings

Computer science and linguistics research has developed and evaluated several algorithms for embedding language, and have demonstrated that language embeddings are a class of *semantic representation*.[31] Two commonly-explored properties of semantic representations are *features* and *similarity*. Both these properties could help representations to explain economic decisions. This subsection briefly describes the structure of language embeddings, the design of embedding algorithms, and the geometric properties of representation spaces. For the interested reader, I have included references with more detail on these topics.

Computational and structural advantages of language embeddings. Language embeddings are vectors that represent word sequences. For a language with vocabulary $\mathcal{W}$, an embedding algorithm $g : \mathcal{W}^* \to \mathbb{R}^K$ maps a sequence of words from the vocabulary to a K-dimensional vector. Compared to an encoding of the full discrete sequence of words, an embedding reduces dimensionality and allows for structured similarity measurement between sequences.For example, the sequences $z_1 =$ *"the firm launched another app"* and $z_2 =$ *"the firm released another service"* are more semantically similar to one another than they are to the sequence $z_3 =$ *"the firm fired another CEO."* Under some discrete encoding schemes, these three language sequences may be represented as equally similar. Using the edit distance $\mathrm{H}(\cdot,\cdot)$ over words—the number of word substitutions required to transform one sequence into another—these similarities are the same, $\mathrm{H}(z_1, z_2) = \mathrm{H}(z_1, z_3) = 2$. In a embedding space that measures similarity through an inner product, the embedding similarity between z_1 and z_2 could be higher than the embedding similarity between z_1 and z_3.

The set of potential word sequences in a language is very large, which means inference on full word sequences requires a very large number of parameters. In a language

[31] Jurafsky and Martin (2024) and Park et al. (2023) discuss the structure of semantic representations in more detail. It is worth noting that language embedding algorithms produce vector representations of word sequences that—in addition to semantic information—incorporate information about other linguistic features like syntax and morphology. While I refer to these vector representations as semantic representations, semantic information is a subset of the information they contain.

with vocabulary $\mathcal{W}$, there are $|\mathcal{W}|^L$ possible sequences of length L. In a language with vocabulary size 500,000, for example, there are $500{,}000^{10} \approx 8.9 \times 10^{84}$ possible sequences of 10 words.[32] Embeddings allow for inference on word sequences with a smaller number of free parameters. Neural network approaches to learning embeddings build on the idea of a distributed representation (Hinton, 1984)—a representation of an entity through a collection of computing elements. Bengio et al. (2000) introduced a neural probabilistic language model that represented words using vectors. Further developments in word-level embeddings include Mikolov et al. (2013a, word2vec), who developed models to predict words from context, and Pennington et al. (2014, GloVe), who developed models that consider word co-occurence patterns. Reimers and Gurevych (2019, SentenceTransformers) introduced an architecture that efficiently learns language sequence embeddings for semantic similarity tasks. Training these models using contrastive loss functions can produce embeddings which group language sequences with similar features more closely together (Wang and Liu, 2021; Wang and Isola, 2022; Chen et al., 2021; Gao et al., 2022).

C.2. Details on Training Procedures

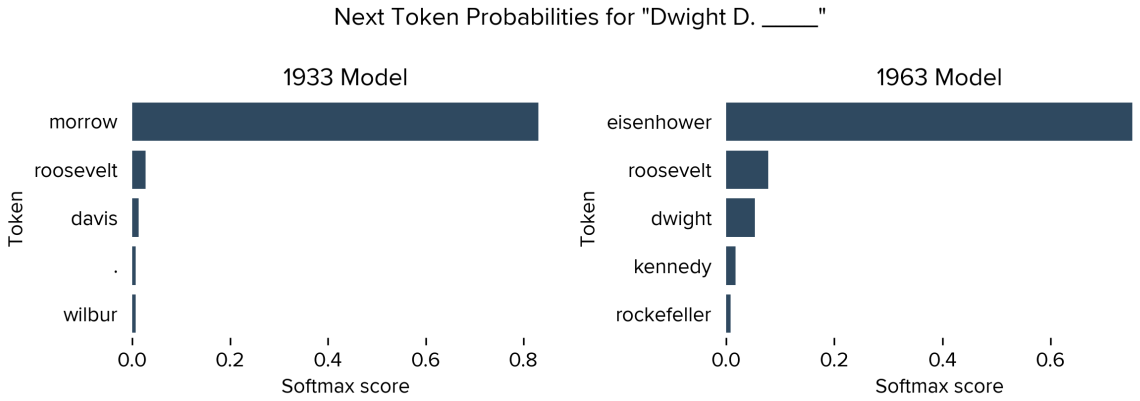
I include further detail on how the StoriesLM and RepresentLM models were trained.

StoriesLM training. StoriesLM is a family of language models with sequentially-expanding pretraining windows. The pretraining data for the model family comes from the American Stories dataset (Dell et al., 2024). This dataset is a collection of language from historical American news articles. The first language model in the StoriesLM family is pretrained on language data from 1900. In this paper, I use the StoriesLM-v1 family, which uses a BERT architecture (Devlin et al., 2019) initialized from scratch, using the default architecture configuration. Each subsequent language model is initialized with the previous year's model checkpoint and trained on the following year's language data. The final model in the family uses language data ending in 1963.

Each model in the StoriesLM family was trained using the HuggingFace Transformers library (Wolf et al., 2020). All models were trained for one epoch using masked language modeling with masking probability 0.15. Each training run was conducted on an Nvidia Tesla V100 GPU. Each document in the pretraining data was tokenized using the "bert-base-uncased" tokenizer, with truncation and padding to a uniform sequence length of 512 tokens.

[32] The Oxford English Dictionary contains more than 500,000 words. Link.

Figure C1: Time-stamped language models avoid information leakage.



Notes: This figure shows how time-stamped language models can avoid information leakage from pretraining. Using the StoriesLM-v1-1933 and StoriesLM-v1-1963 models, I generate fill-mask softmax scores for the sequence "Dwight D. [MASK]." The 1933 model does not generate the sequence "eisenhower"—this should be expected since "Dwight D. Eisenhower" was not a common language sequence in training data that ends in 1933. The 1933 model instead finishes the sequence with "morrow"—Dwight Morrow was an American politician and diplomat who had held public office in the early 1900s. By 1963, Dwight D. Eisenhower had spent time as a general and president, and was well known. The 1963 model finishes the sequence with "eisenhower." Time-stamped language models can prevent information from the future from leaking into language analysis that should only use data from the past.

RepresentLM training. RepresentLM is a language model that produces embeddings. The model's pretraining data comes from the Headlines dataset (Silcock et al., 2024). This dataset is a collection of matched headlines that refer to the same article. As a pair of matched headlines is semantically related, each matched pair can be used as a positive example for a semantic similarity algorithm. I initialized training using the StoriesLM-v1-1963 model.

The RepresentLM model was trained using the SentenceTransformers library (Reimers and Gurevych, 2019). I filtered the Headlines data to observations from 1920–1979. I restricted to articles with at least five matched headlines. To avoid dominating training with many examples from widely-printed articles, I restricted to at most eight randomly selected pairs from the same article. The model was trained for one epoch on an Nvidia Tesla V100 GPU using a multiple negatives ranking loss function (Henderson et al., 2017) using the default SentenceTransformers parameterization.

C.3. Hyperparameter Tuning for Outcome Estimation

For each fold in each year, I fit parameters and tune hyperparameters on the other four folds in the year. In each iteration of fitting, I standardize the industry and characteristics vectors, and their changes, using the mean and standard deviation on the four training folds in each year. I do not standardize the embeddings or embedding changes to preserve the geometry of the embedding space. To handle these different scalings across variable types, I use the group ridge procedure from Ignatiadis and Lolas (2021).

This procedure sets a groupwise ridge objective

$$\widehat{\boldsymbol{v}} = \widehat{\boldsymbol{v}}(\boldsymbol{\lambda}) \in \arg\min_{\boldsymbol{v}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (Y_i - \boldsymbol{x}_i \cdot \boldsymbol{v})^2 + \sum_{g=1}^{K} \frac{\lambda_g}{2} \|\boldsymbol{v}_{G_g}\|_2^2 \right\}$$

where the regularization hyperparameter λ_g can vary for each group g of coefficients. The optimization procedure proceeds across a common hyperparameter σ that generates the vector of regularization parameters across groups. I use a separate group for each explanatory variable vector type {embedding, characteristics, industry} and each explanatory variable vector timing {level, change}. As in Ignatiadis and Lolas (2021), I use the accelerated leave-one-out objective to speed up computation of the hyperparameter search.

D. Additional Results

I include additional results on validating the representations measure, and on explaining returns using the measure.

D.1. Additional Validation Results

Similarity. I evaluate the relationship between representation similarity and established measures of similarity using Fama–MacBeth regressions.

Table D1: Representation similarity relates to established measures of similarity.

	Return Correlation	Shared Analyst Coverage
	(1)	(2)
Representation Similarity	0.057***	0.079***
	(0.006)	(0.004)
Intercept	0.127***	0.016**
	(0.013)	(0.008)
Years	38	38
R^2	0.020	0.008

Notes: This table reports the results of Fama–MacBeth regressions of pairwise daily return correlation and shared analyst coverage on representation similarity. The dependent variable is percentile ranked in each year. Fama–MacBeth standard errors are reported in parentheses.

Valuation ratios and cash flow forecasts. The main analysis dataset matches language to non-missing market equity and positive book equity, as these are the key variables throughout the analysis. I do not match to positive earnings and non-missing long term growth forecasts in the main analysis dataset because these variables are not used outside of Section 3.2. The statistics in columns (2) and (3) of Table 1 restrict to observations with the defined and non-missing observations of the relevant outcome variables, which means that these columns correspond to statistics on fewer observations than the full dataframe used for column (1). In this table, I report statistics for the 37,267-observation subsample that is fully matched across the three outcome variables.

Table D2: Representations help to explain prices and forecasts.

R^2: Estimates of price and forecast variables			
Dependent variable $\rightarrow$	$\log(\text{P/B})$	$\log(\text{P/E})$	$\mathbb{F}[\text{LTG}]$
Explanatory variables $\downarrow$	(1)	(2)	(3)
Representation	29.6%	12.7%	19.8%
Industry (FF12)	9.9%	7.0%	14.7%
Sub-Industry (FF48)	11.1%	7.4%	15.0%
Characteristics	19.4%	8.6%	24.2%
Representation+All Others	35.9%	15.2%	31.6%
All Others	25.6%	11.5%	26.8%

Notes: This table reports the R^2 on a series of estimates of stock price and cash-flow forecast variables across sets of explanatory variables. The estimates are from Equation (5): $Y_{s,t} = \widehat{\boldsymbol{v}}_t \cdot \boldsymbol{x}_{s,t} + \varepsilon_{s,t}$, following the split-sample approach of Section 2.3. Each row reports R^2 estimates across a different set of explanatory variables—representations, industry codes, and characteristics, and concatenations of these vectors. Each column reports R^2 estimates across a different price or forecast variable. The set of firms used to fit each estimator does not include the firms used to evaluate the estimates.

D.2. Additional Return Explanation Results

First, I re-estimate results from Section 4.2 with alternative linguistic transformations, a random vector placebo, and alternative language sources. Second, I assess whether perception-orthogonal shocks relate to the $[\Delta x]$ component. Third, I re-estimate results from Section 4.2 with log returns.

Alternative linguistic transformations. I construct alternative linguistic measures $\boldsymbol{x}_{s,t}^{\text{alt}}$. I construct both word count-based measures and alternative transformer embeddings. I use these measures to evaluate split-sample explanatory power over returns, following Equation (6).

I compute word counts and transformer embeddings across the coverage of firm s in year t. For word count-based measures, I concatenate all articles for each firm in each year. In each year, I use a token count vectorizer to compute token occurrences in the concatenated string of news coverage. To match the dimensionality of the representations measure, I use 768 features. For vectorization, I use either hashed count vectorizers or TF-IDF vectorizers. I use the default scikit-learn implementation (Pedregosa et al., 2018), and I exclude the default list of English stopwords.[33] For transformer embeddings, I use both the StoriesLM-v1-1963 model and the RepresentLM-v1 model. The StoriesLM-v1 model was trained using masked language modeling (MLM), and I use the model's [CLS] embedding for each input chunk. The RepresentLM-v1 model was trained using semantic textual similarity (STS), and I use the model's embedding of each input chunk. For both embedding types, I pool across 512-token chunks in each article, and compute the average embedding across each firm in each year.

I replicate the analysis in Table 2 using these alternative linguistic measures. Table D3 reports the results of this additional analysis.

[33]The English stopword dictionary includes common function words like "the," "this," or "that."

Table D3: Alternative linguistic transformations and returns.

R^2: Estimates of annual returns			
Estimation strategy $\rightarrow$	$[\Delta v]$	$[\Delta x]$	[Total]
Explanatory variables $\downarrow$	(1)	(2)	(3)
Word Counts	4.6%	2.4%	5.7%
TF-IDF	5.7%	3.1%	7.3%
Untargeted Embedding (MLM)	5.1%	3.6%	7.2%
Baseline Embedding (STS)	7.9%	5.6%	11.0%
Main Representation (STS+FT)	9.4%	6.8%	13.4%

Notes: This table reports split-sample R^2 statistics of using estimated returns to explain realized cross-sectional returns, following Equation (6): $R_{s,t+1} = \widehat{\mathbb{E}}_t[R_{s,t+1} \mid z] + \varepsilon_{s,t+1}$. Each row reflects a different set of explanatory variables computed using an alternative linguistic transformation. Each column reflects a different estimation strategy.

This table reports goodness-of-fit across the alternative linguistic measures. The STS embedding outperforms both word count-based measures. The MLM embedding outperforms raw word counts, but does not outperform TF-IDF in the $[\Delta v]$ component. As Reimers and Gurevych (2019) discuss, models trained using MLM are not targeted to semantic tasks, and may not produce optimal embeddings of language sequences. The additional semantic training in the STS step generates a meaningful increase in explanatory power over the raw MLM embedding.

The main embedding measure in this paper, which is generated through additional contrastive training on financial language, has even further explanatory power over returns. The representation measure increases total explanatory power from 11.0% to 13.4% over the STS embedding. This result indicates that targeting embeddings to the relevant economic context helps to better explain economic outcomes.

Random vector placebo. The analysis in Section 4.2 uses a split-sample approach, so that more parameters do not mechanically lead to better goodness-of-fit. To be clear, this analysis is not a holdout sample test, as returns are correlated within $t+1$. The goal of the estimation procedure is to explain common variation in returns $R_{s,t+1}$ using information in representations $\boldsymbol{x}_{s,t}$, and to avoid mechanical increases in goodness-of-fit from more parameters.

I evaluate the fit of the split-sample strategy on uninformative vectors with equal dimension as the main representations measure. I generate random 768-dimensional vectors $\boldsymbol{x}_{s,t}^{\text{random}}$ from independent standard multivariate normal draws. I then evaluate the relationship between these vectors and the realized return using the split-sample approach in Section 4.2. I find that the R^2 from this approach is -0.003. With the split sample approach, this uninformative vector measure—with equal number of parameters as the representations measure—does not mechanically generate explanatory power over returns.

Regulatory filings. Regulatory filings are another rich source of language about firms. Hoberg and Phillips (2016) use language in 10-K filings to study product similarity between firms.

10-K filings are written for regulatory compliance. Therefore, they may contain information about firms that is different from the information in news language. As the focus of this paper is not on how firms report business models, but on how the market perceives business models, it is feasible that the detailed and dynamic information in news language is a better proxy for perception and could better explain returns. In addition, 10-Ks are filed annually, which may lead filings data to have less total information than news data over the period. I benchmark explained variation from the representations measure against explained variation from 10-K filings.

I use 10-K filings from 1994–2021, sourced from the data repository of Loughran and McDonald (2016). I use regular expressions to filter to the business description section in the first 10-K filing in each calendar year. I embed each firm's business description section in each year using the RepresentLM-v1 model, pooling over 512 token chunks. I merge this embedding with the main analysis dataframe, which results in 49,304 observations. I then replicate the analysis in Section 4.2 using this measure.

Table D4: 10-K filing language and returns (matched sample).

R^2: Estimates of annual returns		
Estimation strategy $\rightarrow$	$[\Delta v]$	$[\Delta x]$
Explanatory variables $\downarrow$	(1)	(2)
Representation	9.3%	7.1%
10-K Filing	3.3%	0.4%

Notes: This table reports split-sample R^2 statistics of using estimated returns to explain realized cross-sectional returns, following Equation (6): $R_{s,t+1} = \widehat{\mathbb{E}}_t[R_{s,t+1} \mid z] + \varepsilon_{s,t+1}$. Each row reflects a different set of explanatory variables: The representation and the filings embedding. Each column reflects a different estimation strategy.

Representations explain more variation in returns than the filing embedding. Explanatory power from the 10-K filing embedding is particularly low for the $[\Delta x]$ component. This is intuitive: Language in 10-K filings changes less frequently than language in the news, and market participants may not always pay attention to changes in 10-K filings (Cohen et al., 2020). This result indicates that news language has additional information that could be useful for measuring the market's perceived business model.

Earnings calls speeches. Earnings call speeches are another rich source of language about firms. However, as earnings calls are held quarterly, this data source may have less total information than the total content of news data. In addition, as with regulatory filings, the market's perceived business model may differ from an executive's characterization of the firm in a speech.

I use earnings call speeches from 2003–2021, using the same 31,698 firm-year sample as in Section 6. As in that section, I use the average speech embedding for each firm in each calendar year from the RepresentLM-v1 model, pooling over 512 token chunks in each speech. I then replicate the analysis in Section 4.2 using this measure.

Table D5: Earnings speech language and returns (matched sample).

R^2: Estimates of annual returns		
Estimation strategy $\rightarrow$	$[\Delta v]$	$[\Delta x]$
Explanatory variables $\downarrow$	(1)	(2)
Representation	7.1%	5.8%
Earnings Call	2.1%	0.8%

Notes: This table reports split-sample R^2 statistics of using estimated returns to explain realized cross-sectional returns, following Equation (6): $R_{s,t+1} = \widehat{\mathbb{E}}_t[R_{s,t+1} \mid z] + \varepsilon_{s,t+1}$. Each row reflects a different set of explanatory variables: The representation and the earnings call embedding. Each column reflects a different estimation strategy.

Representations explain more variation in returns than the earnings call embedding. This result indicates that news language has additional information that could be useful for measuring the market's perceived business model.

Interpretation of the $[\Delta x]$ component. I assess whether the $[\Delta x]$ component relates to perception-orthogonal shocks to returns. I show that commodity price shocks in year $t+1$ more strongly affect year $t+1$ returns of firms in industries more related to the commodities. However, such commodity price shock-driven return variation does not relate to variation in the $[\Delta x]$ component.

I obtain monthly data on commodity prices from the World Bank. To construct the commodity price shock in year $t+1$, I divide the difference between the December commodity prices at $t+1$ and t by the December commodity price at t. I correlate the average crude oil price with the return on energy stocks, the iron ore CFR spot price with the return on manufacturing stocks, and the natural gas index price with the return on utility stocks. I identify each industry using the Fama–French 12 industry classification.

I use the two-stage least squares specification

$$\begin{aligned}[\Delta x]_{s,t+1} &= \beta \times \text{Return}_{s,t+1} + \alpha_{t+1} + \alpha_I + u_{s,t+1} \\ \text{Return}_{s,t+1} &= \gamma \times 1(\text{Exposed Industry})_{s,t} \times \text{Shock}_{t+1} + \alpha_{t+1} + \alpha_I + v_{s,t+1}\end{aligned}$$

where $1(\text{Exposed Industry})_{s,t}$ indicates whether stock s is a member of the target industry in year t, Shock_{t+1} is the change in price of the target commodity in year $t+1$, α_I is an industry fixed effect, and α_{t+1} is a time fixed effect. The identification condition is that membership in the exposed industry of firm s in year t is orthogonal to the commodity price shock in year $t+1$.

Table D6 and Table D7 report the results of these regressions. In each table, Column (1) shows that commodity price shocks differentially explain the returns of exposed industries. Column (2) shows that commodity price shocks do not explain the $[\Delta x]$ component. Finally, Column (3) shows that returns instrumented using commodity price shocks do not explain the $[\Delta x]$ component. These results demonstrate that perception-orthogonal shocks to returns do not relate to returns explained by changes in representation.

Table D6: Returns attributed to oil price shocks do not explain returns attributed to changes in representation.

	Return	$[\Delta x]$	$[\Delta x]$
	(1)	(2)	(3)
Oil price shock $\times$ Energy firm	0.36^{***}	0.01	
	(0.05)	(0.01)	
Return			0.02
			(0.02)
F-test (first stage, projected)			104.2
Projected R^2	0.001	0.000	
Observations	81,708	81,708	81,708
Year fixed effects	✓	✓	✓
Industry fixed effects	✓	✓	✓
Regression type	First Stage	Reduced Form	IV

Notes: This table reports the results of regressions that relate returns attributed to the $[\Delta x]$ component to oil price shocks. Standard errors clustered by year and industry are reported in parentheses.

Table D7: Returns attributed to iron price shocks do not explain returns attributed to changes in representation.

	Return	$[\Delta x]$	$[\Delta x]$
	(1)	(2)	(3)
Iron price shock $\times$ Manufacturing firm	0.11^{***}	0.01	
	(0.02)	(0.01)	
Return			0.06
			(0.06)
F-test (first stage, projected)			31.4
Projected R^2	0.000	0.000	
Observations	81,708	81,708	81,708
Year fixed effects	✓	✓	✓
Industry fixed effects	✓	✓	✓
Regression type	First Stage	Reduced Form	IV

Notes: This table reports the results of regressions that relate returns attributed to the $[\Delta x]$ component to iron price shocks. Standard errors clustered by year and industry are reported in parentheses.

Table D8: Returns attributed to natural gas price shocks do not explain returns attributed to changes in representation.

	Return	$[\Delta x]$	$[\Delta x]$
	(1)	(2)	(3)
Natural gas price shock $\times$ Utility firm	0.10^{***}	0.00	
	(0.02)	(0.00)	
Return			0.02
			(0.04)
F-test (first stage, projected)			17.8
Projected R^2	0.000	0.000	
Observations	81,708	81,708	81,708
Year fixed effects	✓	✓	✓
Industry fixed effects	✓	✓	✓
Regression type	First Stage	Reduced Form	IV

Notes: This table reports the results of regressions that relate returns attributed to the $[\Delta x]$ component to natural gas price shocks. Standard errors clustered by year and industry are reported in parentheses.

Log returns. I re-estimate the results from Section 4.2 using log annual returns as the outcome variable.

Table D9: Representations help to explain returns.

R^2: Estimates of log annual returns			
Estimation strategy $\rightarrow$	$[\Delta v]$	$[\Delta x]$	[Total]
Explanatory variables $\downarrow$	(1)	(2)	(3)
Representation	11.7%	9.0%	16.2%
Industry (FF12)	4.6%	<0%	4.6%
Sub-Industry (FF48)	5.4%	<0%	5.4%
Characteristics	5.6%	5.1%	12.1%
Representation+All Others	14.4%	11.6%	20.9%
All Others	9.2%	5.1%	14.8%

Notes: This table reports split-sample R^2 statistics of using estimated returns to explain realized cross-sectional log returns, following Equation (6): $r_{s,t+1} = \widehat{\mathbb{E}}_t[r_{s,t+1} \mid z] + \varepsilon_{s,t+1}$. Each row reflects a different set of explanatory variables: Representations, industry vectors, characteristics vectors, and combinations of these vectors. Each column reflects a different estimation strategy.

E. Additional Return Forecasting Results

I include additional results on return forecasting. I report additional controls for the baseline specification, portfolio results, and additional controls for the attention-drawing features specification.

Additional controls. I augment the baseline return forecasting specification with a series of control variables. I control for the deviation in the log price-to-book ratio over the same horizon as I compute the priced deviation, i.e. $\log(P/B)_{s,t} - \frac{1}{5}\sum_{k=1}^{5}\log(P/B)_{s,t-k}$. I also control for past 12-month and 36-month returns, excluding the most recent month. Finally, I control for the log price-to-book ratio.

Table E1: Additional controls: Deviations in representation forecast returns.

Dependent variable: Monthly return [%]					
	(1)	(2)	(3)	(4)	(5)
Representation Deviation (priced)	-0.45***	-0.46***	-0.28***	-0.30**	-0.27***
	(0.16)	(0.14)	(0.10)	(0.13)	(0.10)
Valuation Function Deviation (priced)	0.06	0.18	0.24	0.18	0.26
	(0.24)	(0.19)	(0.19)	(0.21)	(0.16)
Deviation in $\log(P/B)$	0.08				-0.14
	(0.24)				(0.19)
$\text{Return}_{12,1}$		0.04			0.37
		(0.31)			(0.24)
$\text{Return}_{36,1}$			-0.52		-0.46
			(0.40)		(0.36)
$\log(P/B)$				-0.52	-0.35
				(0.32)	(0.27)
Forecasting R^2	0.013	0.016	0.017	0.015	0.028
Months	456	456	456	456	456

Notes: This table reports results of Fama–MacBeth forecasting regressions of future returns on priced deviation variables and controls, following Equation (9): $R_{s,m,t+1} = \alpha_m + \beta_m^T Z_{s,t} + \varepsilon_{s,m,t+1}$. Sorting variables are computed at end of each year, and each sorting variable is cross-sectionally percentile ranked. Fama–MacBeth standard errors are reported in parentheses.

Portfolio sorts. I report results from portfolio sorts of the priced deviation variables. These results correspond to returns from equal-weighted portfolios that go long the bottom quintile of priced deviation and go short the top quintile of priced deviation.

Table E2: Portfolio sorts.

	Representation Deviation			Valuation Function Deviation		
	(1)	(2)	(3)	(4)	(5)	(6)
Alpha [%]	0.41***	0.40***	0.35***	-0.06	-0.18	-0.24
	(0.12)	(0.12)	(0.11)	(0.22)	(0.23)	(0.23)
MKT		0.02	0.01		0.16**	0.21***
		(0.04)	(0.03)		(0.08)	(0.08)
SMB			0.23***			-0.12
			(0.07)			(0.23)
HML			0.28***			0.22*
			(0.06)			(0.13)
Months	456	456	456	456	456	456

Notes: This table reports the monthly returns of long–short portfolios that bet against the priced representation deviation and priced valuation function deviation, following $R_m^{\text{long}-\text{short}} = \alpha + \beta^T f_m + \varepsilon_m$. Robust standard errors are reported in parentheses.

Portfolio sorts, excluding microcaps. I report results from portfolio sorts of the priced deviation variables, excluding stocks smaller than the 20th percentile of NYSE stocks.

Table E3: Portfolio sorts (excluding microcaps).

	Representation Deviation			Valuation Function Deviation		
	(1)	(2)	(3)	(4)	(5)	(6)
Alpha [%]	0.43***	0.42***	0.36***	-0.07	-0.19	-0.26
	(0.13)	(0.13)	(0.12)	(0.24)	(0.25)	(0.25)
MKT		0.02	0.02		0.16**	0.22***
		(0.04)	(0.03)		(0.08)	(0.08)
SMB			0.20***			-0.19
			(0.06)			(0.23)
HML			0.28***			0.21
			(0.06)			(0.13)
Months	456	456	456	456	456	456

Notes: This table reports the monthly returns of long–short portfolios that bet against the priced representation deviation and priced valuation function deviation, following $R_m^{\text{long}-\text{short}} = \alpha + \beta^T f_m + \varepsilon_m$. Robust standard errors are reported in parentheses.

Additional controls. I augment the attention-drawing features return forecasting specification with a series of control variables. I control for the deviation in the log price-to-book ratio over the same horizon as I compute the priced deviation variables, i.e. $\log(P/B)_{s,t} - \frac{1}{5}\sum_{k=1}^{5}\log(P/B)_{s,t-k}$. I also control for past 12-month and 36-month returns, excluding the most recent month. I also control for the log price-to-book ratio. Finally, I control for the priced deviation variables.

Table E4: Additional controls: Incorporation of attention-drawing features.

Dependent variable: Monthly return [%]						
	(1)	(2)	(3)	(4)	(5)	(6)
Incorporate trending features	-0.24**			-0.27*	-0.15	-0.15
	(0.11)			(0.14)	(0.12)	(0.11)
Incorporate profitable features		-0.34***		-0.23*	-0.29**	-0.29**
		(0.10)		(0.14)	(0.12)	(0.11)
Incorporate high-performing features			-0.27***	-0.32*	-0.12	-0.07
			(0.10)	(0.19)	(0.13)	(0.12)
Deviation in $\log(P/B)$	-0.06	-0.07	-0.05		-0.07	-0.16
	(0.19)	(0.19)	(0.19)		(0.19)	(0.17)
$\text{Return}_{12,1}$	0.37	0.37	0.34		0.37	0.38
	(0.27)	(0.28)	(0.27)		(0.27)	(0.26)
$\text{Return}_{36,1}$	-0.46	-0.42	-0.42		-0.40	-0.40
	(0.39)	(0.39)	(0.38)		(0.38)	(0.37)
$\log(P/B)$	-0.30	-0.34	-0.31		-0.31	-0.35
	(0.32)	(0.32)	(0.32)		(0.32)	(0.28)
Representation deviation (priced)						-0.17**
						(0.09)
Valuation function deviation (priced)						0.28
						(0.19)
Forecasting R^2	0.024	0.024	0.024	0.004	0.025	0.029
Months	456	456	456	456	456	456

Notes: This table reports results of Fama–MacBeth forecasting regressions of future returns on feature incorporation variables, following Equation (9): $R_{s,m,t+1} = \alpha_m + \beta_m^T Z_{s,t} + \varepsilon_{s,m,t+1}$. Sorting variables are computed at end of each year, and each sorting variable is cross-sectionally percentile ranked. Fama–MacBeth standard errors are reported in parentheses.

F. Additional Communication Results

I show that deviations in manager communication correspond to deviations in market representations. I use the RepresentLM-v1 model to embed each firm's earnings call speech in each quarter. If a speech exceeds 512 tokens, I take the mean embedding across 512-token chunks of the speech. I compute the average of the speech embeddings for each firm in each year to construct a firm-by-year manager embedding $\boldsymbol{e}_{s,t}$ across 2003–2021.

I then compute the deviation in manager embedding $\Delta\boldsymbol{e}^{h}_{s,t} \equiv \boldsymbol{e}_{s,t} - \frac{1}{h}\sum_{k=1}^{h}\boldsymbol{e}_{s,t-k}$. As with the representations measure learned from news language, I use $h = 5$. In addition, I compute the deviation in manager embedding using any available previous years of manager embeddings—for example, if a firm only has one previous year of earnings call data, the deviation in manager embedding is $\boldsymbol{e}_{s,t} - \boldsymbol{e}_{s,t-1}$. I merge the manager embedding dataframe with the main analysis dataset, which results in 31,698 firm-year observations.

Using cosine similarity ψ, I compute

$$\text{Managerial Communication Reversion}_{s,t+1} = -\psi(\Delta\boldsymbol{e}_{s,t+1}, \Delta\boldsymbol{e}^{h}_{s,t})$$

I find that the average reversion is 0.31.

I price the deviation in managerial communication by estimating

$$\text{Managerial Communication Deviation (priced)}_{s,t} = \widehat{\boldsymbol{v}}_t \cdot \Delta\boldsymbol{e}^{h}_{s,t} \tag{13}$$

where $\widehat{\boldsymbol{v}}_t$ is a function that maps manager embeddings to the log price-to-book ratio $\text{pb}_{s,t}$, estimated using the split-sample ridge procedure from Section 2.3.

Table F1: Deviations in managerial communication relate to deviations in representations.

	Representation Deviation (priced)
	(1)
Managerial Communication Deviation (priced)	0.146***
	(0.048)
Intercept	0.427***
	(0.024)
Years	18
R^2	0.061

Notes: This table reports the results of a Fama–MacBeth regression of the priced deviation in representation on the priced deviation in managerial communication. Both the variables are percentile ranked in each year. Fama–MacBeth standard errors are reported in parentheses.

Table F1 reports the results of a Fama–MacBeth regression of the priced deviation in market representation on the priced deviation in managerial communication. If the deviation in managerial communication implies a higher price, the deviation in market representation also implies a higher price. This result shows that communication by managers is associated with the market's perceptions.

Acknowledgment of Data and Funding

I use data from the Dow Jones Newswires historical news archive. The Dow Jones received this paper prior to its circulation "for the sole purpose of verifying that Dow Jones and its services are correctly attributed."

I am supported by a National Science Foundation Graduate Research Fellowship and a Two Sigma PhD Fellowship. This project is supported, in part, by funding from Two Sigma Investments, LP. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of Two Sigma Investments, LP. Neither the National Science Foundation nor Two Sigma had a right to review this paper prior to its circulation.