Specific Signals in a Noisy World: Idiosyncratic Forward-Looking Disclosures and Predictable Returns

Haowei Yuan*

July 26, 2025

Abstract

This paper examines how idiosyncratic (firm-specific) versus systematic (non-specific) forward-looking statements in corporate disclosures affect markets differently. Using natural language processing, I classify forward-looking statements from 10-K filings and develop neural network-based growth probability measures for each type. Analysis of U.S. public firms (1998-2022) reveals idiosyncratic forward-looking statements significantly outperform systematic ones in predicting future growth. A one standard deviation increase in idiosyncratic growth measure yields a 1.25% excess stock return at 180 days, while systematic information loses predictive power when orthogonalized against idiosyncratic content. Additionally, firms with higher idiosyncratic forward-looking growth experience reduced stock price volatility. Analysts respond positively to idiosyncratic—but not systematic—forward-looking information specifically during downward forecast revisions, consistent with managers selectively disclosing negative information. These findings demonstrate the superior information content of firm-specific forward-looking statements and document investor underreaction to idiosyncratic information, consistent with limited attention theories.

JEL Code: G14, G12, D83, M41
Keywords: Forward-looking statements; Idiosyncratic information; Market efficiency; Corporate disclosure; Growth predictability; Machine Learning

*Haowei Yuan is from Penn State University, email: hky5193@psu.edu. I am grateful to my committee members Alexey Zhdanov and Mihail Velikov for their invaluable guidance and support. I also thank Matthew Gustafson, Anh Le, Lawrence Jin, and all the other participants at Smeal Ph.D. Colloquium for their helpful comments and suggestions. All remaining errors are my own.

1 Introduction

Forward-looking statements in corporate disclosures represent a critical channel through which firms communicate expectations about future performance to capital markets. These statements, typically found in the Management Discussion and Analysis (MD&A) section of firms' 10-K filings, provide investors with insights into management's perspectives on future growth opportunities, strategic initiatives, and potential challenges (Muslu et al., 2015; Li, 2010). Despite the regulatory environment encouraging forward-looking disclosures through safe harbor provisions established by the Private Securities Litigation Reform Act of 1995, the information content and market implications of such statements remain incompletely understood.

This study investigates a fundamental but underexplored distinction within forward-looking disclosures: the differential information content and market reaction to idiosyncratic (firm-specific) versus systematic (non-specific) forward-looking statements. I define idiosyncratic forward-looking statements as those that specifically address firm-level operations, strategies, and performance expectations, while systematic forward-looking statements refer to broader industry trends, macroeconomic conditions, and market-wide factors that may affect future performance. This distinction is theoretically important because it aligns with the fundamental categorization of risk and information in finance theory, which separates firm-specific factors from systematic ones (Merton, 1989; Roll, 1988).

Understanding firm growth is fundamental to both corporate finance and asset pricing. Firm growth drives value creation, affects capital allocation decisions, and ultimately determines long-term shareholder returns(Berk et al., 1999). For investors, the ability to identify firms with superior growth prospects is central to investment strategies and portfolio formation (La Porta et al., 1997; Chan et al., 2003). However, forecasting firm growth remains challenging due to information asymmetry between managers and external stakeholders (Myers and Majluf, 1984).

Measuring growth probability—not just understanding realized growth—represents a crucial advancement in this domain. Growth probability quantifies the likelihood of future performance improvements based on current information, creating a forward-looking metric that

extends beyond traditional backward-looking measures (Campbell and Shiller, 1988). The distinction between idiosyncratic and systematic forward-looking statements provides a particularly relevant framework for measuring growth probability because these two information types theoretically contain different signals. Idiosyncratic forward-looking statements reflect firm-specific plans, initiatives, and management perspectives that should, in principle, provide more precise signals about firm-specific growth opportunities. Conversely, systematic forward-looking statements address broader economic conditions that may affect all firms in an industry or market, potentially offering less discriminatory power for predicting individual firm outcomes. By separately measuring growth probabilities derived from these distinct information sources, I provide a more nuanced understanding of how different types of forward-looking information contribute to growth expectations and how markets process these differential signals.

The efficient markets hypothesis suggests that all publicly available information should be rapidly incorporated into asset prices (Fama, 1970). However, a growing body of literature documents various market anomalies and inefficiencies, particularly related to the processing of complex textual information (Tetlock, 2007; Loughran and McDonald, 2011). Behavioral finance theories suggest that investors may face cognitive limitations when processing detailed, firm-specific information that requires greater attention and analytical resources (Hong and Stein, 1999; Hirshleifer et al., 2003). This creates an intriguing tension: do capital markets process idiosyncratic and systematic forward-looking information differently, and if so, what are the implications for market efficiency and asset pricing?

To address these questions, I develop a novel methodological approach leveraging recent advancements in natural language processing. Specifically, I employ a pretrained BERT (Bidirectional Encoder Representations from Transformers) model[1] to distinguish between firm-specific and non-specific forward-looking statements from firms' 10-K filing MD&A sections. I then utilize OpenAI's text-embedding-3-large model to convert these statements into 128-dimensional vectors and train a feed-forward neural network with one hidden layer to predict future sales growth. This approach allows me to construct two distinct measures

[1]I thank Yi Yang for making Finbert-FLS publicly available. Detailed description of this pretrained model is in Section 2.2.

of growth probability based on idiosyncratic and systematic forward-looking statements, respectively.

Using a comprehensive sample of U.S. public firms from 1998 to 2022 as my testing period, I document several important findings that highlight the differential impact of idiosyncratic versus systematic forward-looking information. First, idiosyncratic forward-looking statements significantly outperform systematic forward-looking statements in predicting future firm growth, suggesting that firm-specific disclosures contain more precise and value-relevant information than general market or industry statements. Second, I find a strong positive relationship between forward-looking growth measures and future stock returns, with the effect strengthening over longer horizons. A one standard deviation increase in my idiosyncratic information measure translates into a 1.25% excess stock return at the 180-day horizon. To disentangle the effects of idiosyncratic and systematic information, I perform an orthogonalization analysis which reveals that when systematic forward-looking growth information is purged of its idiosyncratic component, it loses its predictive power for future returns. In contrast, idiosyncratic forward-looking growth information maintains robust predictive power even after removing any systematic influences, further supporting the unique value of firm-specific disclosures. Third, firms with higher idiosyncratic forward-looking growth measure experience lower stock price response following disclosures, indicating reduced information asymmetry. Systematic forward-looking growth measure, however, shows no significant relationship with post-disclosure response. Fourth, examining analyst behavior reveals that while analysts generally do not react to either forward-looking measure when controlling for revenue surprises, they do respond positively to idiosyncratic forward-looking information—but not systematic information—specifically in cases of downward revisions. This asymmetric reaction is consistent with the selective disclosure hypothesis, whereby managers typically release positive news early to analysts but withhold negative information due to litigation concerns, making idiosyncratic forward-looking disclosures particularly informative during periods of negative news revelation.

The predictive power of idiosyncratic forward-looking information is vividly illustrated by the case of Agios Pharmaceuticals. Despite reporting positive sales growth from 2013 to 2014, the firm's 10-K filing on February 24, 2015, contained numerous firm-specific forward-looking

statements that received notably low growth probability scores from my model. Examining these statements reveals subtle but important linguistic signals of future underperformance. For instance, the statement with the lowest score (0.09) candidly acknowledged: "*We are also unable to predict when, if ever, material net cash inflows will commence from AG-221, AG-120, AG-348, or any of our other product candidates.*" Other low-scoring statements emphasized continuing losses ("*We expect to continue to incur significant expenses and operating losses over the next several years,*" scoring 0.14) and distant revenue prospects ("*Our commercial revenues, if any, will be derived from sales of medicines that we do not expect to be commercially available for many years, if at all*", scoring 0.17). Despite ongoing clinical development activities and partnership with Celgene, the linguistic patterns in these firm-specific disclosures signaled caution about near-term growth prospects[2].

These signals proved prescient—Agios experienced negative sales growth in fiscal year 2015, and its stock price declined by 39% from February 2015 to January 2016, as shown in Figure 1. This case demonstrates how my model effectively captures substantive information in idiosyncratic forward-looking statements that may contradict recent financial metrics. While conventional analysis might have focused on Agios's positive historical growth trend, analysis of firm-specific forward-looking language provided early warning signals about future performance challenges. This example highlights how systematic parsing of idiosyncratic versus systematic components of forward-looking statements can reveal insights that are not immediately reflected in backward-looking financial metrics or market prices, potentially providing investors with actionable information about future growth prospects.

A potential threat to the validity of my findings is the presence of look-ahead bias in the construction of my forward-looking information measures. As the training sample for OpenAI's text-embedding-3-large model is based on historical data up to 2021, there is a risk that the model may have learned from future information that was not available at the time of the training. To address this concern, before calling the OpenAI's embedding model, I substitute the year number in all the sentences with "t" if the year number is the filing's corresponding fiscal year, and $t+i$ ($t-i$) if the year is the next i (previous i) fiscal year. This adjustment ensures that the model does not learn about the exact year of the filing,

[2]See Appendix A. for the whole Specific Forward-looking Sentences parsed from the MD&A section.

mitigating the risk of look-ahead bias.

As a robustness check, I conduct portfolio sorts and additional Fama-MacBeth regressions to further validate the predictive power of my idiosyncratic forward-looking growth measure. I find that the idiosyncratic forward-looking growth measure consistently predicts future stock returns across portfolio sort and Fama-MacBeth regression analyses, even after controlling for various firm characteristics and other known return predictors.

These findings contribute to several streams of literature. First, I extend the disclosure literature by providing a more nuanced understanding of how the specificity of forward-looking statements affects their information content and market implications (Bozanic et al., 2018; Hope et al., 2016). Li (2010) and Muslu et al. (2015) find that forward-looking statements contain valuable information about future performance, but do not distinguish between idiosyncratic and systematic components. Frankel et al. (2016) and Bochkay and Levine (2019) use machine learning technique to extract information that explains future earnings. More recently, Kim and Nikolaev (2024b) show that contextualization of accounting numbers substantially improves the informativeness of financial reports. My research builds upon these findings by demonstrating that the idiosyncratic component of forward-looking statements contains significantly more informative content about future growth prospects than the systematic component.

Second, I contribute to the market efficiency literature by documenting differential investor reactions to idiosyncratic versus systematic information, supporting theories of limited attention and information processing costs (Hirshleifer et al., 2011; Cohen and Lou, 2012; Cohen et al., 2020). My findings reveal a significant post-filing drift following disclosures with rich idiosyncratic forward-looking content, indicating investors initially underreact to specific growth signals, with the effect persisting until actual growth materializes. This pattern aligns with Cohen et al. (2020), who demonstrate investor underreaction to complex information requiring greater cognitive effort, and with theoretical frameworks suggesting cognitive constraints limit investors' ability to fully process detailed firm-specific disclosures (Hirshleifer et al., 2003; Hong and Stein, 1999).

Third, my findings contribute to the broader literature on textual analysis in finance and

economics [3]. A recent survey by Hoberg and Manela (2025) classify text-based research into three categories: Research Objective Categories (ROCs)—Targeted, Holistic, and Comparative. My study falls into the ROC-Holistic category, extracting comprehensive document information to predict economic variables. While many studies in this domain rely on dictionaries (Tetlock, 2007; Tetlock et al., 2008; Loughran and McDonald, 2011; Garcia et al., 2023) or keyword lists (Li et al., 2013; Hassan et al., 2019; Bourveau et al., 2020; Jiang et al., 2024), my approach leverages advanced embedding models. I combine FinBERT (Yang et al., 2020) with OpenAI's embedding model and follow machine learning best practices (Gu et al., 2020; Chen et al., 2022; Kim and Nikolaev, 2024a) to train a neural network specifically optimized for growth prediction.

My research also has important implications for corporate disclosure policies and investor behavior. The documented underreaction to idiosyncratic forward-looking information suggests that firms may benefit from increasing the specificity of their forward-looking disclosures to improve information dissemination. For investors, my findings highlight potential opportunities from systematic analysis of firm-specific forward-looking statements, which appear to contain valuable information not immediately reflected in stock prices.

The remainder of this paper is organized as follows. Section 2 describes my data and methodology. Section 3 presents my empirical results on the predictive power of idiosyncratic versus systematic forward-looking statements for future growth and stock returns. Section 4 presents additional portfolio sorts and test as a support. Section 5 concludes with implications and directions for future research.

2 Data and Methodology

This Section provides a detailed description of the data and methodology used in this study. Subsection 2.1 describes the data and sample selection process. Subsection 2.2 explains the classification of forward-looking statements into specific and non-specific categories. Subsection 2.3 discusses the contextualization process for embedding. Subsection 2.4 outlines

[3]See literature reviews by Gentzkow et al. (2019); Loughran and McDonald (2020); Ash and Hansen (2023); Cong et al. (2021).

the model design and training process. Subsection 2.5 presents the performance metrics of the models trained on specific and non-specific forward-looking statements. Finally, Subsection 2.6 describes how I construct the idiosyncratic and systematic forward-looking growth measures.

2.1 Data and Sample Selection

My primary data source is the Management Discussion and Analysis (MD&A) section of corporate 10-K filings obtained from the U.S. Securities and Exchange Commission's EDGAR database. Following the filtering procedure established by Loughran and McDonald (2011), I download all cleaned 10-K and 10-K405 filings between 1995 and 2023 from Loughran& McDonald's website[4] and match each filing with the summary data file from their website.

I exclude duplicates and retain only the first filing per CIK-year. I require at least 180 days between a given firm's 10-K filings and match with CRSP permno. The sample is restricted to common shares listed on NYSE, AMEX, or NASDAQ exchanges with a price greater than $3 on the day before filing. I require at least 60 days of returns and volume data in the year prior to and following the file date. Each filing is matched with the COMPUSTAT annual data on same filing period and I further require available book-to-market data with book value greater than zero.

From each filing, I extract the MD&A section and require the number of sentences to be between 10 and 1,000. The final sample includes 69,677 firm-year observations from 9,937 unique firms spanning 1995 to 2023. The detailed sample filtering process with observations loss for each step is presented in Table 1.

For return calculations, I use the CRSP value-weighted index downloaded from Kenneth French's website as the benchmark. All accounting variables, returns, and estimated measures are winsorized at the 1st and 99th percentiles to mitigate the influence of outliers.

[4] https://sraf.nd.edu/sec-edgar-data/cleaned-10x-files/

2.2 Specific and Non-specific Forward-looking Statements

FinBERT by Yang et al. (2020) is a domain-specific language model based on BERT (Bidirectional Encoder Representations from Transformers), tailored for financial language processing. It is designed to improve the performance of natural language processing (NLP) tasks in the financial domain, such as sentiment analysis and text classification.

I deploy the FinBERT-FLS model[5] that is publicly available on huggingface to classify sentences in the MD&A section of 10-K filings into three categories: Specific Forward-looking Statements (SPFLS), Non-specific Forward-looking Statements (NSPFLS), and Not Forward-looking Statements (NFLS).

Panel A of Table 2 presents the distribution of sentences across these two categories. On average, an MD&A section contains 329 sentences, with 15(5%) of them classified as SPFLS and 42(12.5%) as NSPFLS. The % of Negative words are calculated by dividing the number of negative words in the whole 10-K documents by the total number of words in the documents, using the 10-K summary data from Loughran&McDonald's website. The average % of negative words is 1.7% and the average % of positive words is 0.5%.

2.3 Contextualization and Embedding

To contextualize the forward-looking statements, I subsequently create context windows for each identified SPFLS and NSPFLS sentence. The context window is a fixed-size segment of text surrounding the identified sentence, which provides additional context for the embedding model to analyze. I set the context window size to 2 sentences before and 2 sentences after the identified sentence. This means that for each SPFLS or NSPFLS sentence, I extract a total of 5 sentences (the identified sentence plus 2 preceding and 2 following sentences) to form the context window[6]. This approach allows the embedding model to capture the surrounding context and improve the quality of the generated embeddings.

[5]The model is pretrained on a large corpus of financial text and fine-tuned on a manually annotated dataset of 3,500 sentences from the MD&A section of annual reports of Russell 3000 firms. It takes input text and output the label from the following three categories, "Specific-FLS , Non-specific FLS, or Not-FLS".

[6]Kim and Nikolaev (2024b) show a peak contextual embedding predictive performance using 1 sentence before and after, I use 2 sentences to preserve more contextual meaning as my model is trained in a different way from theirs.

I then use OpenAI's text-embedding-3-large model to convert each of the context windows into vectors. OpenAI embeddings include 3,072 dimensions ordered by importance. I follow a similar practice as Kim et al. (2024) and only use the first 128 dimensions of the embedding vectors[7]. Thus, the input to the embedding model is a 5-sentence context window, and the output is a 128-dimensional vector representing the contextualized information of the SPFLS or NSPFLS sentence.

2.4 Model Design and Training

I model the interpretation process as a process of learning important prospects related with future growth in the multi-dimensional space of the embedding vectors,

$$\mathrm{E}_t[T_{t+1,i}|C_{t,i,j}] = \Omega_t(C_{t,i,j}) + \epsilon_{t,i,j} \tag{1}$$

where $T_{t+1,i}$ is the materialized future sales growth of firm i, $C_{t,i,j}$ is the context information vector of the j-th Forward-looking statements in the MD&A section of the 10-K filing of firm i at time t, $\Omega_t(.)$ is the non-linear function that maps each of the context information vector to the future sales growth, and $\epsilon_{t,i,j}$ is the error term.

I use a feed-forward neural network with one hidden layer of 32 neurons and a one-dimensional output layer to model the non-linear function $\Omega_t(.)$. The network takes sentence-level 128-dimensional embedding vector representation $C_{t,i,j}$ as input. The output layer is a single neuron representing the predicted future sales growth. I use the rectified linear unit (ReLU) activation function for the hidden layer and a linear activation function for the output layer.

In this neural network architecture, each neuron in the hidden layer maintains weighted connections to all input layer neurons. These connection weights can be conceptualized as vectors indicating directions of growth-relevant information in the embedding space. When an input context vector aligns with these weight vectors, it signifies that the contextual information contains significant growth-predictive content. This structure effectively imple-

[7]OpenAI's text-embedding-3-large model is trained using a technique called "Matryoshka Representation learning", which allows the condensing of the dimensionality without losing the embedding's contextual properties (Kusupati et al., 2022).

ments a learning mechanism that identifies and extracts salient growth indicators from the multidimensional embedding representation of forward-looking statements.

The model is trained with Adam optimizer and with a Binary Cross-Entropy logits loss function. The training, validation, and test samples are constructed on a rolling basis over time. The training sample consists of the three years prior to the test year, while the validation sample includes the year before the test year. As our initial data is from 1995, the first test sample year is 1998 (i.e. the train set is from 1995 to 1996, the validation set is 1997). I then roll the training and validation samples forward each year and, thus, have 25 test samples from 1998 to 2022.

Following the standard practice in machine learning, my training consists of two phrases. First, I use a grid search to find an optimal set of hyperparameters for the model[8]. The optimal hyperparameters and corresponding training epochs are then used to train the final model on the most recent two years prior to the test year. (i.e. 1996-1997 for the test year 1998). This two-stage approach allows the model to learn from the most relevant and recent data, improving its predictive performance. For each test year, I train model seperately for SPFLS and NSPFLS sentences, with the first model aims to capture the idiosyncratic forward-looking signals, and the second model aims to capture the systematic forward-looking signals.

2.5 Model Performance

Table 3 presents the performance metrics of the models trained on the SPFLS and NSPFLS sentences. The main evaluation metrics are accuracy, F1 score, and AUC (Area Under the Receiver Operating Characteristic Curve). The accuracy measures the proportion of correctly classified sentences, while the F1 score combines precision and recall to provide a balanced measure of model performance. The AUC considers both the true positive rate and false positive rate, providing a comprehensive evaluation of the model's ability to predict future sales growth. Both models use each sentence as a training sample or testing sample,

[8]I set four learning rates: (1e-5, 1e-4, 1e-3, 1e-2), three batch sizes: (128, 256, 512) for grid searching. I also adopt an early stopping criteria based on the validation loss, which stops the training process if the validation loss does not improve for 10 consecutive epochs.

which results in a large number of testing sample size per year. As there are more sentences in MD&A section that are classified as NSPFLS, both the training and testing samples of NSPFLS sentences are around 3 times larger than those of SPFLS sentences. On average, models trained on SPFLS sentences achieve an accuracy of 0.6068, an F1 score of 0.7278, and an AUC of 0.5360, whereas the models trained on NSPFLS sentences yield an accuracy of 0.6069, an F1 score of 0.7205, and an AUC of 0.5188. However, despite the large difference in sample size, the performance of the two models is comparable using a pairwise t-test method following Chen et al. (2022). The AUC of the SPFLS model is slightly higher than that of the NSPFLS model (0.5360 vs. 0.5188, with a p-value of 0.0199), and the accuracy and F1 score of the SPFLS model are not significantly different from those of the NSPFLS model (0.6068 vs. 0.6069, with a p-value of 0.9943 and 0.7278 vs. 0.7205, with a p-value of 0.7643).

As Sarkar and Vafa (2024) point out, the look-ahead bias of pre-trained language models for prediction tasks is a potential concern. Models trained using embeddings from the pre-trained embedding models may unavoidably learn from future information that was not available at the time of training. To mitigate this risk, I implement a contextualization process to ensure that the model does not learn about the exact year of the filing. Specifically, I replace the year number in all sentences with "t" if the year number corresponds to the filing's fiscal year, and $t + i$ ($t - i$) if the year is the next i (previous i) fiscal year. This adjustment ensures that the model does not know the exact fiscal year of the filing, thereby reducing the risk of look-ahead bias.

Nevertheless, as a further precautionary check, I focus on the last two rows (2021-2022) of Table 3 to examine the model's performance in the pure out-of-sample tests (Kim and Nikolaev, 2024a,b). All of the three performance metrics of both models are still comparable to the overall sample average, with almost no deterioration in performance. As a result, look-ahead bias is unlikely to be a significant contributing factor to the model's performance.

2.6 Measure of Growth Probability

For each test year from 1998 to 2022, I then use the trained models for SPFLS and NSPFLS sentences to construct the idiosyncratic and systematic forward-looking growth measures. For each firm-year observation, I take the average of the predicted probabilities of future

sales growth across all SPFLS(NSPFLS) sentences in the MD&A section of the 10-K filing of that firm-year. I further standardize the predicted probabilities by calculating the z-score of all firms in the same Fama-French industry-year group. The purpose of this standardizing process is to control for the industry-year fixed effects, which helps to isolate the idiosyncratic and systematic components of the forward-looking growth probability measures. The summary statistics of the idiosyncratic and systematic forward-looking growth measures are presented in Panel B of Table 2. $Salegr^{adj}_{SP}$ *(*$Salegr^{adj}_{NSP}$*)* has a mean of 0.002 (0.003) and a standard deviation of 0.963 (0.953).

Next, I calculate the correlation matrix between the idiosyncratic and systematic forward-looking growth measures and other variables. The correlation matrix is presented in Table 4. The idiosyncratic forward-looking growth measure ($Salegr^{adj}_{SP}$) is positively correlated with the systematic forward-looking growth measure ($Salegr^{adj}_{NSP}$) (0.385), indicating that firms with higher idiosyncratic forward-looking growth measures also tend to have higher systematic forward-looking growth measures. Both of these measures also show positive correlations with current log sales growth (0.103 and 0.076, respectively), log size (0.115 and 0.068, respectively). They are negatively correlated with the log of book-to-market ratio (-.065 and -0.047, respectively), and percentage of either positive or negative words, and percentage of SPFLS or NSPFLS sentences. Overall, the correlation matrix suggests that the idiosyncratic and systematic forward-looking growth measures are not capturing the normal unidimensional signals that are typically used in the literature, such as the sentiment or the length.

3 Empirical Findings

The previous section describes the training details and construction of the idiosyncratic and systematic forward-looking growth measures. I find no significant difference in the performance of the two models trained using SPFLS and NSPFLS sentences, respectively. Nor do I find look-ahead bias significantly drives the model performance. The two growth measures are positively correlated with each other and do not show significant correlation with any of the derived measures in the literature. In this section, I present the main

empirical findings of the study. Subsection 3.1 summarizes the characteristics of quintile portfolios sorted based on the idiosyncratic and systematic forward-looking growth measures. Subsection 3.2 presents the predictive power of the idiosyncratic and systematic forward-looking growth measures for future growth. Subsection 3.3 investigates the stock returns and forward-looking growth measures. Subsection 3.4 examines the stock market response to forward-looking growth measures.

3.1 Characteristics of Quintile Portfolios

Table 5 presents the characteristics of quintile portfolios sorted based on the idiosyncratic and systematic forward-looking growth measures. I sort firms into quintiles based on their idiosyncratic forward-looking growth measure ($Salegr^{adj}_{SP}$) and systematic forward-looking growth measure ($Salegr^{adj}_{NSP}$) at the end of June each year. Panel A shows that high $Salegr^{adj}_{SP}$ firms are larger, low book-to-market, high growth, high profitability, low market leverage, but high book leverage, and low cash holdings. In panel B, high $Salegr^{adj}_{NSP}$ firms exhibit same characteristics as high $Salegr^{adj}_{SP}$ firms, but with a lower magnitude.

The findings in Table 5 reveal important patterns regarding the characteristics of firms with high growth probabilities as indicated by idiosyncratic and systematic forward-looking statements. These patterns provide valuable insights into the information content of these two distinct types of forward-looking disclosures.

First, the directional consistency between the characteristics associated with high $Salegr^{adj}_{SP}$ and high $Salegr^{adj}_{NSP}$ firms suggests that both idiosyncratic and systematic forward-looking statements contain similar signals about future growth prospects. This alignment indicates that managers tend to include optimistic forward-looking content—whether firm-specific or broad market-related—when the firm exhibits characteristics traditionally associated with growth opportunities: larger size, lower book-to-market ratios, higher historical growth, and stronger profitability. This consistency provides validation that both measures capture meaningful information about growth potential, despite their different focus.

Second, the capital structure patterns—low market leverage but high book leverage for firms with high growth probability measures—merit particular attention. This seemingly contradictory finding can be reconciled by considering that high-growth firms typically have

higher market valuations (explaining the lower market leverage) while still utilizing debt financing to fund their growth initiatives (explaining the higher book leverage). The stronger magnitude of this pattern for idiosyncratic forward-looking statements suggests that managers more confidently provide firm-specific forward-looking information when they have successfully leveraged their balance sheet to support growth initiatives without weakening their market position.

Third, the lower cash holdings observed for firms with high growth probability measures may initially seem counterintuitive, as growth firms often maintain cash reserves for investment flexibility. However, this finding can be explained by considering that firms actively pursuing growth opportunities may be deploying their cash reserves toward investments, resulting in lower cash balances at the time of disclosure.

These portfolio characteristics collectively strengthen the case that idiosyncratic forward-looking statements contain more precise signals about future performance than systematic forward-looking statements. The consistency in direction coupled with differences in magnitude between the two measures provides further evidence that markets should, in principle, extract more valuable information from firm-specific forward-looking disclosures and the underreaction to idiosyncratic forward-looking information represents a form of market inefficiency that cannot be explained by differences in firm characteristics alone.

3.2 Predictive Power of Forward-looking Growth Measures

In order to examine the predictive power of the idiosyncratic and systematic forward-looking growth measures for future growth, I conduct a panel regression analysis with high-dimensional fixed effects included:

$$Y_{t+1,i} = \alpha + \beta_{SP} Salegr^{SP}_{t,i} + \beta_{NSP} Salegr^{NSP}_{t,i} + \gamma X_{t,i} + \iota_{k,t} + \epsilon_{t,i} \tag{2}$$

where $Y_{t+1,i}$ is one of the future growth measures, $Salegr_{t,i}$ is the idiosyncratic or systematic forward-looking growth measure, $X_{t,i}$ is a vector of control variables including $leverage$, $logsize$, $logbm$, $LMneg$, $LMpos$, and $Similarity$. $\iota_{k,t}$ is the Fama-French 48 industry-year fixed effect, and $\epsilon_{t,i}$ is the error term. The dependent variable $Y_{t+1,i}$ is one of the following

measures: 1) future sales growth rate, 2) future sales growth dummy, 3) future asset growth rate, and 4) future asset growth dummy. The standard errors are double clustered at the firm and fiscal year levels to account for potential correlation in the error terms across firms and over time and all variables are winsorized at the 1st and 99th percentiles.

Table 6 presents the results of the panel regression analysis. The first four columns compare the relative predictive power of the idiosyncratic and systematic forward-looking growth measures for future sales growth. The results show that the idiosyncratic forward-looking growth measure ($Salegr_{SP}^{adj}$), when included in the regression only, has a positive and significant coefficient (0.010) with a t-stat of 3.87 for future sales growth rate, while the systematic forward-looking growth measure ($Salegr_{NSP}^{adj}$) has insiginicant coefficient (-0.001), when included in the regression together with idiosyncratic forward-looking growth measure. The idiosyncratic forward-looking growth measure also has a positive and significant coefficient (0.037) with a t-stat of 5.51 for future sales growth dummy, while the systematic forward-looking growth measure has a less significant coefficient (0.014) with a t-stat of 2.43, when included in the regression together with idiosyncratic forward-looking growth measure. The differences of β_{SP} and β_{NSP} are significant in both specifications as in column (2) and (4), with t-stat of 3.16 and 2.35, respectively. The last four columns further compare the predictive power of the idiosyncratic and systematic forward-looking growth measures for future asset growth. The results show that the idiosyncratic forward-looking growth measure ($Salegr_{SP}^{adj}$), when included in the regression only, has a positive and significant coefficient (0.013) with a t-stat of 5.61 for future asset growth rate, while the systematic forward-looking growth measure ($Salegr_{NSP}^{adj}$) has insiginicant coefficient (0.002), when included in the regression together with idiosyncratic forward-looking growth measure. The idiosyncratic forward-looking growth measure also has a positive and significant coefficient (0.034) with a t-stat of 5.55 for future asset growth dummy, while the systematic forward-looking growth measure has a less significant coefficient (0.009) with a t-stat of 2.58, when included in the regression together with idiosyncratic forward-looking growth measure. The differences of β_{SP} and β_{NSP} are also significant in both specifications as in column (6) and (8), with t-stat of 3.50 and 3.73, respectively. The results overall suggest that the idiosyncratic forward-looking growth measure is a better predictor of future growth

than the systematic forward-looking growth measure.

The results in Table 6 provide strong evidence that the idiosyncratic forward-looking growth measure ($Salegr^{adj}_{SP}$) significantly predicts future growth, while the systematic forward-looking growth measure ($Salegr^{adj}_{NSP}$) does not. This finding is consistent with the notion that firm-specific forward-looking statements contain more precise and value-relevant information than general market or industry statements.

3.3 Stock Returns and Forward-looking Growth Measures

To investigate the market reaction to forward-looking growth measures, I conduct event studies regression as in Garcia et al. (2023) and Loughran and McDonald (2011). Similarly, I estimate the following regression model:

$$R_{t,i} = \alpha + \beta_1 Salegr_{t,i} + \gamma X_{t,i} + \eta_t + \iota_k + \epsilon_{t,i} \tag{3}$$

where t is the 10-K filing date; $R_{t,i}$ is the firm's buy-and-hold stock return minus the CRSP value-weighted market index return over the different length (4-day, 10-day, 30-day, 60-day, 120-day, and 180-day) windows (from close at $t-1$ to $t+2$, $t+8$, and so on); $Salegr_{t,i}$ is the idiosyncratic or systematic forward-looking growth measure; $X_{t,i}$ is a vector of control variables including Loughran-McDonald sentiment measures ($LMneg$ and $LMpos$), $logbm$, $logsize$, $turnover$, and Nasdaq dummy. I winsorize all continuous variables at 1/99% percentiles to mitigate the influence of outliers. As a standard practice for a event study with unbalanced panel, I include both time (fiscal year-quarter) and industry fixed effects (Fama-French 48 industry) in the regression. I report standard errors clustered on FF49 industries and fiscal year-quarters in all specifications.

The result of the event study regression is presented in Table 7. I find that the idiosyncratic forward-looking growth measure ($Salegr^{adj}_{SP}$) positively and significantly predicts future stock returns for all event windows, with the effect strengthening over longer horizons. For example, a one standard deviation increase in the idiosyncratic forward-looking growth measure translates into a 1.21% (1.252*0.963=1.21%) excess stock return at the 180-day horizon. In contrast, the systematic forward-looking growth measure ($Salegr^{adj}_{NSP}$) does not

show significant predictive power for stock returns, especially when I control for the idiosyncratic forward-looking growth measure and the results are presented in the Appendix Table B1.

The results in Table 7 provide strong evidence that the idiosyncratic forward-looking growth measure ($Salegr_{SP}^{adj}$) significantly predicts future stock returns and creates a positive post-filing drift, with the effect strengthening up to 180 days after the filing date. This finding is consistent with the notion that firm-specific forward-looking statements contain more precise and value-relevant information than general market or industry statements, and that investors underreact to this information at the time of disclosure. As the realized growth materializes, the market corrects this underreaction, leading to a positive post-filing drift in stock returns.

3.4 Stock Market Response and Forward-looking Growth Measures

To further investigate the market response to forward-looking growth measures, I conduct a similar event study regression as in Section 3.3 but with the dependent variable being the absolute stock return over the different length (4-day, 10-day, 30-day, 60-day) windows. The regression model is as follows:

$$|R_{t,i}| = \alpha + \beta_1 Salegr_{t,i} + \gamma X_{t,i} + \eta_t + \iota_k + \epsilon_{t,i} \tag{4}$$

where $|R_{t,i}|$ is the absolute stock return over the different length (4-day, 10-day, 30-day, 60-day) windows. Except from the controls that are used in the event study regression, I also include the percentage of FLS sentences ($FLS\%$) calculated as (% of SPFLS + % of NSPFLS) in the MD&A section of the 10-K filing as a control variable. As is shown in Bozanic et al. (2018), percentage of forward-looking statements in the MD&A section is positively correlated with the stock market response.

Table 8 presents the results of the event study regression above. I find that the idiosyncratic forward-looking growth measure ($Salegr_{SP}^{adj}$) negatively and significantly predicts event window stock market response, with the results keeping stable across different event win-

dows. For example, a one standard deviation increase in the idiosyncratic forward-looking growth measure translates into a 11.94 basis point (0.124*0.963=0.1194) decrease in the absolute stock return at the 10-day horizon. However, the systematic forward-looking growth measure ($Salegr^{adj}_{NSP}$) does not show significant predictive power for stock market response, after controlling for the idiosyncratic forward-looking growth measure, and the results are presented in the Appendix Table B2.

The findings in Table 8 provide strong evidence that the idiosyncratic forward-looking growth measure ($Salegr^{adj}_{SP}$) significantly predicts the stock market response to forward-looking growth measures, with the effect being negative and significant across all event windows. Such findings are also consistent with the notion that firm-specific forward-looking statements contain more precise and value-relevant information than general market or industry statements, and that investors resolve this uncertainty more effectively when they are provided with firm-specific forward-looking information that are more associated with the future growth.

3.5 Analyst Revision and Forward-looking Growth Measures

The previous sections have demonstrated that idiosyncratic forward-looking statements contain more precise and value-relevant information than systematic statements, as evidenced by their superior ability to predict future growth and stock returns. If these idiosyncratic forward-looking disclosures truly contain superior information content, we might expect financial analysts—sophisticated information intermediaries in capital markets—to incorporate this information into their forecasts. To test this hypothesis, I examine analyst revenue forecast revisions around 10-K filing and earnings announcement windows. The regression model is as follows:

$$Revision_{t,i,k} = \alpha + \beta_1 Salegr_{t,i} + \gamma X_{t,i} + \eta_t + \iota_i + \epsilon_{t,i,k} \tag{5}$$

where $Revision_{t,i,k}$ is the percentage change in the analyst k's yearly revenue forecast revision for firm i around fiscal year earnings announcement date t. $X_{t,i}$ is a vector of control variables including Loughran-McDonald sentiment measures ($LMneg$ and $LMpos$),$Fls\%$,

Similarity, *logbm*, *logsize* and revenue surprise compared with analyst consensus in the pre-announcement period. Firm and industry by year fixed effects are included in the regression. Standard errors are double clustered at the firm and industry by year levels.

Table 9 presents the results of the regression analysis. In the full sample (Column (1) and (2)), neither the idiosyncratic nor systematic forward-looking growth measures show a significant relationship with analyst revisions after controlling for revenue surprises. This initial result might suggest that analysts efficiently incorporate the information contained in both types of forward-looking statements, or that they rely primarily on the realized revenue surprise rather than forward-looking disclosures.

However, when I partition the sample based on the direction of analyst revisions, a more nuanced pattern emerges. Column (3) to (5) presents results for the subsample of downward revisions (where analysts reduce their revenue forecasts following the announcement). In this subsample, the idiosyncratic forward-looking growth measure ($Salegr_{SP}^{adj}$) exhibits a positive and significant coefficient, indicating that higher idiosyncratic forward-looking growth measures attenuate the magnitude of downward revisions. In contrast, the systematic forward-looking growth measure ($Salegr_{NSP}^{adj}$) remains insignificant even in this subsample. In column (5), I include both idiosyncratic and systematic forward-looking growth measures in the regression. The coefficient of the idiosyncratic forward-looking growth measure ($Salegr_{SP}^{adj}$) remains positive and significant, while the coefficient of the systematic forward-looking growth measure ($Salegr_{NSP}^{adj}$) remains insignificant. The difference in the coefficients of the two measures is significant with a t-stat of 2.31.

This asymmetric reaction pattern aligns with the selective disclosure hypothesis in the literature on corporate disclosure and analyst behavior. Managers typically have incentives to release positive news early to analysts through various communication channels (conference calls, private meetings, etc.) due to career concerns and compensation incentives tied to stock performance (Kothari et al., 2009; Houston et al., 2010). However, they tend to withhold negative information until required disclosure dates due to litigation concerns (Skinner, 1994; Kasznik and Lev, 1995). Consequently, forward-looking statements in mandatory filings become particularly informative when they contain negative news that managers were previously reluctant to disclose.

The finding that only idiosyncratic forward-looking statements-rather than systematic ones—influence analyst behavior during downward revisions further emphasizes the superior information content of firm-specific disclosures. When analysts are processing negative news and updating their forecasts downward, they appear to discriminate between different types of forward-looking information, placing greater weight on firm-specific statements that contain more precise signals about future performance.

This result complements my earlier findings on stock market underreaction to idiosyncratic forward-looking information. While the market generally underreacts to firm-specific forward-looking statements (as evidenced by the post-filing drift), analysts—as sophisticated information intermediaries—show a more nuanced response, incorporating this information specifically when revising forecasts downward. This differential reaction between market prices and analyst forecasts suggests that the post-filing drift documented earlier may be driven primarily by less sophisticated investors who fail to fully process the information content of idiosyncratic forward-looking statements, particularly when these statements contradict recent performance metrics.

4 Robustness Tests

As a robustness check, I conduct several additional tests to ensure the robustness of my main findings. First, I take a more robust approach to construct the a trading strategy based on the idiosyncratic forward-looking growth measure ($Salegr_{SP}^{adj}$). The Long-Short portfolio earns 34 basis points (0.34%) per month on average. Second, I run the Fama-MacBeth (1973) regression to examine the predictive power of the idiosyncratic forward-looking growth measure ($Salegr_{SP}^{adj}$) for future stock returns.

4.1 Trading Strategy

I construct a trading strategy based on the idiosyncratic forward-looking growth measure ($Salegr_{SP}^{adj}$) to further validate its predictive power for future stock returns. I only consider 10-K filings with a filing date between Feburary and April each year, and I sort firms into quintiles based on their idiosyncratic forward-looking growth measure ($Salegr_{SP}^{adj}$) at the end

of April each year from 1999 to 2023. Then, I form a long-short portfolio by going long on the top quintile (Q5) and shorting the bottom quintile (Q1) of firms with high idiosyncratic forward-looking growth measure ($Salegr_{SP}^{adj}$). Figure 2 presents the cumulative excess returns of the long-short portfolio over the holding period from May to December each year. The long-short portfolio earns 34 basis points (0.34%) per month on average. A further look at the upper panel of Figure 2 shows that the short portfolio (Q1) contributes most because of the negative returns following the filing date. Such a result also corroborates the story that the idiosyncratic forward-looking growth measure ($Salegr_{SP}^{adj}$) is a good predictor of future stock returns, and the market underreacts to the firm-specific forward-looking information.

4.2 Fama-MacBeth Regression

Additionally, I run the Fama-MacBeth (1973) regression to examine the predictive power of the idiosyncratic forward-looking growth measure ($Salegr_{SP}^{adj}$) for future stock returns. Specifically, I estimate the following regression model:

$$r_{tj} = \beta' X_{t-1,j} + \epsilon_{tj} \tag{6}$$

where r_{tj} is the excess return of stock j at month t, $X_{t-1,j}$ is a vector of independent variables including the idiosyncratic forward-looking growth measure ($Salegr_{SP}^{adj}$) at month $t-1$. I run the regression for each month from May to December from 1999 to 2023, and then take the average of the estimated coefficients across all months. The results are presented in Table 10. The idiosyncratic forward-looking growth measure ($Salegr_{SP}^{adj}$) has a positive and significant coefficient (0.16) with a t-stat of 3.87, and it remains significant after controlling for different sets of control variables. This finding further supports the finding that the idiosyncratic forward-looking growth measure ($Salegr_{SP}^{adj}$) significantly predicts future stock returns, and the market underreacts to the firm-specific forward-looking information in the following months.

5 Conclusion

This paper examines how capital markets process idiosyncratic versus systematic forward-looking statements in corporate disclosures. Using advanced natural language processing techniques, I document that idiosyncratic (firm-specific) forward-looking statements significantly outperform systematic (non-specific) statements in predicting future firm growth and stock returns. This finding reveals an important asymmetry in market efficiency: investors appear to process broader, systematic information appropriately while underreacting to detailed, firm-specific disclosures. Additionally, analysts incorporate idiosyncratic—but not systematic—forward-looking information specifically during downward forecast revisions, consistent with the selective disclosure hypothesis that managers withhold negative information until mandatory disclosure dates.

These results have important implications for corporate disclosure policies and investment strategies. Firms seeking to effectively communicate growth prospects should prioritize specific, detailed forward-looking information. For investors, the predictable pattern of returns following disclosures with rich idiosyncratic content suggests potential opportunities from systematic analysis of firm-specific statements. By distinguishing between idiosyncratic and systematic components of forward-looking disclosures, this study provides a more nuanced understanding of information processing in capital markets, contributing to the literature on corporate disclosure, market efficiency, and investor behavior.

References

E. Ash and S. Hansen. Text algorithms in economics. *Annual Review of Economics*, 15(1): 659–688, 2023.

J. B. Berk, R. C. Green, and V. Naik. Optimal investment, growth options, and security returns. *The Journal of finance*, 54(5):1553–1607, 1999.

K. Bochkay and C. B. Levine. Using md&a to improve earnings forecasts. *Journal of Accounting, Auditing & Finance*, 34(3):458–482, 2019.

T. Bourveau, G. She, and A. Žaldokas. Corporate disclosure as a tacit coordination mechanism: Evidence from cartel enforcement regulations. *Journal of Accounting Research*, 58 (2):295–332, 2020.

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

J. Y. Campbell and R. J. Shiller. Stock prices, earnings, and expected dividends. *the Journal of Finance*, 43(3):661–676, 1988.

L. K. Chan, N. Jegadeesh, and J. Lakonishok. Momentum strategies. *Handbook of the Economics of Finance*, 1:427–509, 2003.

Y. Chen, B. T. Kelly, and D. Xiu. Expected returns and large language models. *Available at SSRN 4416687*, 2022.

L. Cohen and D. Lou. Complicated firms. *Journal of financial economics*, 104(2):383–400, 2012.

L. Cohen, C. Malloy, and Q. Nguyen. Lazy prices. *The Journal of Finance*, 75(3):1371–1415, 2020.

L. W. Cong, T. Liang, B. Yang, and X. Zhang. Analyzing textual information at scale. In *Information for efficient decision making: Big data, blockchain and relevance*, pages 239–271. World Scientific, 2021.

E. F. Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417, 1970.

R. Frankel, J. Jennings, and J. Lee. Using unstructured and qualitative disclosures to explain accruals. *Journal of Accounting and Economics*, 62(2-3):209–227, 2016.

D. Garcia, X. Hu, and M. Rohrer. The colour of finance words. *Journal of Financial Economics*, 147(3):525–549, 2023.

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

S. Gu, B. Kelly, and D. Xiu. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273, 2020.

T. A. Hassan, S. Hollander, L. Van Lent, and A. Tahoun. Firm-level political risk: Measurement and effects. *The Quarterly Journal of Economics*, 134(4):2135–2202, 2019.

D. Hirshleifer, S. S. Lim, and S. H. Teoh. Limited attention, information disclosure, and financial reporting. *Journal of Accounting and Economics*, 36(1-3):337–386, 2003.

D. Hirshleifer, S. S. Lim, and S. H. Teoh. Limited investor attention and stock market misreactions to accounting information. *The Review of Asset Pricing Studies*, 1(1):35–73, 2011.

G. Hoberg and A. Manela. The natural language of finance. *USC Marshall School of Business Research Paper Sponsored by iORB*, January 2025. Available at SSRN: https://ssrn.com/abstract=5119322 or http://dx.doi.org/10.2139/ssrn.5119322.

H. Hong and J. C. Stein. A unified theory of underreaction, momentum trading, and overreaction in asset markets. *The Journal of Finance*, 54(6):2143–2184, 1999.

O.-K. Hope, D. Hu, and H. Lu. The benefits of specific risk-factor disclosures. *Review of Accounting Studies*, 21:1005–1045, 2016.

J. F. Houston, B. Lev, and J. W. Tucker. To guide or not to guide? causes and consequences of stopping quarterly earnings guidance. *Contemporary accounting research*, 27(1):143–185, 2010.

H. Jiang, N. Khanna, Q. Yang, and J. Zhou. The cyber risk premium. *Management Science*, 70(12):8791–8817, 2024.

R. Kasznik and B. Lev. To warn or not to warn: Management disclosures in the face of an earnings surprise. *Accounting review*, pages 113–134, 1995.

A. Kim, M. Muhn, V. V. Nikolaev, and Y. Zhang. Learning fundamentals from text. *Chicago Booth Accounting Research Center Research Paper, Fama-Miller Working Paper*, 2024.

A. G. Kim and V. V. Nikolaev. Contextualizing profitability. (Chicago Booth Research Paper No. 23-11, Becker Friedman Institute Working Paper No. 2023-76), June 2024a. doi: 10.2139/ssrn.4459383. URL https://ssrn.com/abstract=4459383.

A. G. Kim and V. V. Nikolaev. Context-based interpretation of financial information. *Journal of Accounting Research*, 2024b.

S. P. Kothari, S. Shu, and P. D. Wysocki. Do managers withhold bad news? *Journal of Accounting research*, 47(1):241–276, 2009.

A. Kusupati, G. Bhatt, A. Rege, M. Wallingford, A. Sinha, V. Ramanujan, W. Howard-Snyder, K. Chen, S. Kakade, P. Jain, et al. Matryoshka representation learning. *Advances in Neural Information Processing Systems*, 35:30233–30249, 2022.

R. La Porta, J. Lakonishok, and R. W. Vishny. Good news for value stocks: Further evidence on market efficiency. *Journal of finance*, 52(2):859–874, 1997.

F. Li. The information content of forward-looking statements in corporate filings—a naïve bayesian machine learning approach. *Journal of accounting research*, 48(5):1049–1102, 2010.

F. Li, R. Lundholm, and M. Minnis. A measure of competition based on 10-k filings. *Journal of Accounting Research*, 51(2):399–436, 2013.

T. Loughran and B. McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *Journal of Finance*, 66(1):35–65, 2011.

T. Loughran and B. McDonald. Textual analysis in finance. *Annual Review of Financial Economics*, 12(1):357–375, 2020.

R. C. Merton. Risk, hedging, and the creation of financial instruments. *Journal of Applied Corporate Finance*, 1(4):3–13, 1989.

V. Muslu, S. Radhakrishnan, K. Subramanyam, and D. Lim. Forward-looking md&a disclosures and the information environment. *Management Science*, 61(5):931–948, 2015.

S. C. Myers and N. S. Majluf. Corporate financing and investment decisions when firms have information that investors do not have. *Journal of financial economics*, 13(2):187–221, 1984.

R. Roll. r^2. *Journal of Finance*, 43(3):541–566, 1988.

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

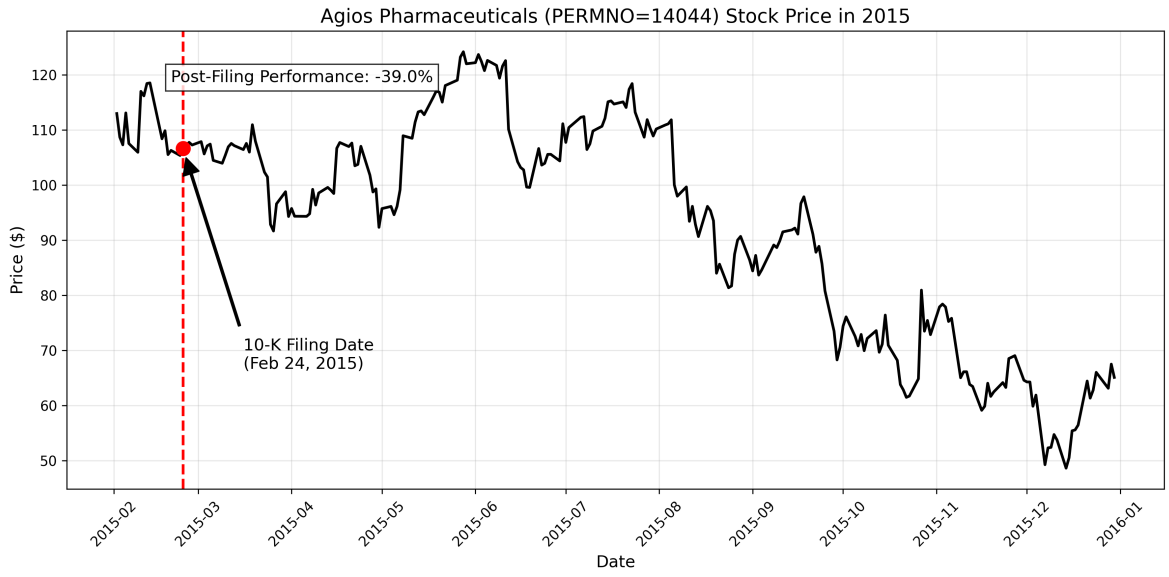
D. J. Skinner. Why firms voluntarily disclose bad news. *Journal of accounting research*, 32 (1):38–60, 1994.

P. C. Tetlock. Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, 62(3):1139–1168, 2007.

P. C. Tetlock, M. Saar-Tsechansky, and S. Macskassy. More than words: Quantifying language to measure firms' fundamentals. *The journal of finance*, 63(3):1437–1467, 2008.

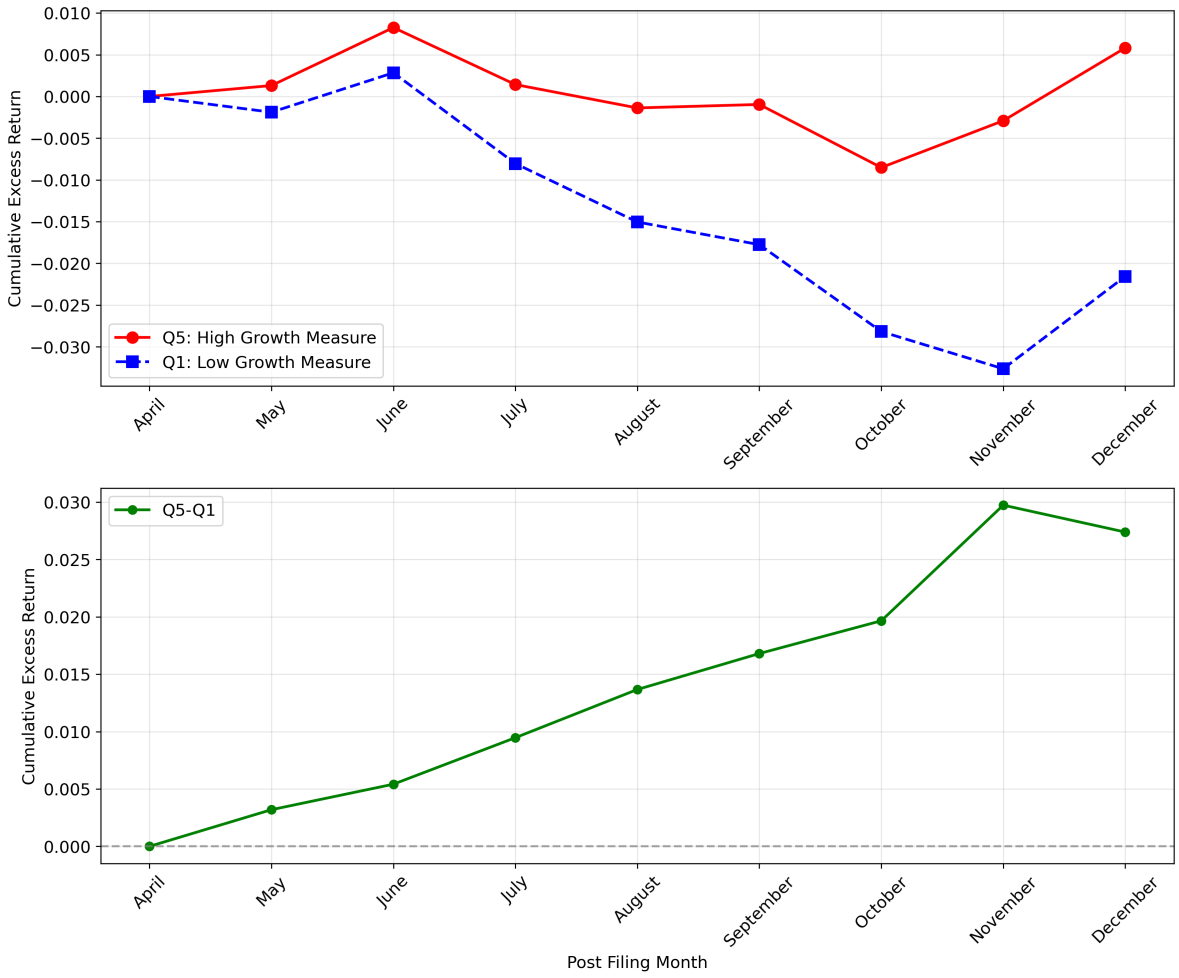
Y. Yang, M. C. S. Uy, and A. Huang. Finbert: A pretrained language model for financial communications. *arXiv preprint arXiv:2006.08097*, 2020.

Figure 1: Agios Pharmaceuticals's stock price in 2015



This figure presents Agios Pharmaceuticals (PERMNO=14044)'s stock price in 2015.

Figure 2: Cumulative Porfolio Excess Returns



This figure presents the cumulative portfolio excess returns sorted on $Salegr_{SP}^{adj}$ at the end of each April from 1999 to 2023. $Salegr_{SP}^{adj}$ is the estimated growth probability using Specific forward-looking sentences and further demeaned by fiscal year and industry. In each April, all firms that file 10-K filings from Feb. to Apr. are sorted into quintiles based on $Salegr_{SP}^{adj}$. Upper figure plots the cumulative equal-weighted monthly excess return against CRSP market index of top and bottom portfolio. Bottom figure plots the cumulative equal-weighted monthly excess return against CRSP market index of Long-Short portfolio.

Table 1: Impact of data filter on sample size

This table reports the impact of data filter on the sample size. I download all 10-K and 10-K405 filings from EDGAR between 1995 and 2023. I exclude duplicates, keep the first filing per cik-year, require at least 180 days between a given firm's 10-K filings, match with CRSP permno, require common shares and NYSE, AMEX or NASDAQ exchange listing, require price on filing date day minus one to be greater than 3, require at least 60 days of returns and volume in year prior to and following file date, require book-to-market COMPUSTAT data available and book value greater than 0, extract MD&A section, and require MD&A sentences to be greater than 10 and less than 1000. The table reports the number of firms and the average number of years in the sample.

Full 10-K Document	Observations
EDGAR 10-K / 10-K405 1995-2023 complete sample (excluding duplicates)	240,365
Only keep first filing per cik-year	235,965
At least 180 days between a given firm's 10-K filings	195,882
CRSP permno match	115,467
Common shares and NYSE, AMEX or NASDAQ exchange listing	104,147
Price on filing date day minus one > 3	87,541
At least 60 days of returns and volume in year prior to and following file date	87,377
Book-to-market COMPUSTAT data available and book value > 0	84,649
MD&A section extracted	82,663
MD&A sentences > 10 and < 1000	69,677
Firm-year sample:	
Uniqne firms	9,937
Average years	7

Table 2: Summary Statistics

This table reports summary statistics for the variables used in the analysis. The sample consists of 69,677 firm-year Item 7 filings from 1995 to 2023. *# Item 7 sentences* is the number of sentences in the Item 7 section of the 10-K filing. *% of SPFLS* (*% of NSPFLS*) is the percentage of Specific-FLS (Non-specific FLS) in the Item 7 section. *# of SPFLS* (*# of NSPFLS*) is the number of Specific-FLS (Non-specific FLS) in the Item 7 section. *% of Positive Words* (*% of Negative Words*) is the percentage of positive (negative) words in the whole 10-K filing calculated by Loughran and McDonald (2011) dictionary. $Salegr_{SP}^{adj}$ *(*$Salegr_{NSP}^{adj}$*)* is the estimated sales growth probability using firms' Specific-FLS (Non-specific FLS) in the Item 7 section further adjusted by year and industry. $\Delta logsale_{t+1}$ is the change in the natural logarithm of sales from year t to year $t+1$. $1(salegr)_{t+1}$ is a dummy variable that equals 1 if the sales growth rate is positive in year $t+1$. $\Delta logat_{t+1}$ is the change in the natural logarithm of total assets from year t to year $t+1$. $1(atgr)_{t+1}$ is a dummy variable that equals 1 if the total assets growth rate is positive in year $t+1$. r_{4d} is the 4-day cumulative buy-and-hold return minus CRSP value-weighted market return. $|r_{4d}|$ is the absolute value of r_{4d}. *Turnover* is the logarithm of past year's trading volume divided by the number of shares outstanding, and at least 60 observations are required. *Leverage* is book leverage. *Logbm* is the natural logarithm of book-to-market ratio. *Logsize* is the natural logarithm of market capitalization. $\Delta logsale_t$ is the change in the natural logarithm of sales from year $t-1$ to year t.

	Count	Mean	Std	P5	P50	P95
# Item 7 sentences	69,677	329.488	171.051	11.000	305.000	999.000
% of SPFLS	69,677	0.050	0.034	0.000	0.042	0.438
# of SPFLS	69,677	15.202	12.117	0.000	12.000	124.000
% of NSPFLS	69,677	0.125	0.062	0.000	0.116	0.647
# of NSPFLS	69,677	42.064	31.445	0.000	37.000	353.000
% of Negative words	69,676	0.017	0.004	0.006	0.017	0.028
% of Positive words	69,676	0.005	0.002	0.002	0.005	0.010
$Salegr_{SP}^{adj}$	60,929	0.002	0.963	-2.565	0.046	2.171
$Salegr_{NSP}^{adj}$	61,299	0.003	0.954	-2.637	0.040	2.234
$\Delta logsale_{t+1}$	64,004	0.085	0.283	-0.999	0.072	1.227
$1(salegr)_{t+1}$	65,242	0.692	0.462	0.000	1.000	1.000
$\Delta logat_{t+1}$	65,159	0.082	0.255	-0.694	0.055	1.115
$1(atgr)_{t+1}$	65,242	0.682	0.466	0.000	1.000	1.000
r_{4d}	69,660	-0.192	7.232	-24.878	-0.134	23.587
$\|r_{4d}\|$	69,660	5.051	5.670	0.056	3.154	31.092
$Turnover$	69,677	7.179	0.991	4.490	7.287	9.394
$Leverage$	69,373	0.203	0.190	0.000	0.162	0.740
$logbm$	69,648	6.078	0.839	3.373	6.183	7.836
$logsize$	69,655	6.445	1.835	2.754	6.324	11.182
$\Delta logsale_t$	67,182	0.110	0.294	-0.876	0.082	1.396

Table 3: Baseline Model Performance Comparison

This table presents the performance metrics of the models trained on NSPFLS(Non-specific Forward-looking Statements) and SPFLS(Specific Forward-looking Statements) samples. I first transform each statement with a context window of 2 sentences into OpenAI embeddings of 128 dimensions. The target variable is a binary indicator that equals one if the sales growth rate is positive in the next fiscal year and zero otherwise. For each test year, I then train a feedforward neural network with 1 hidden layer of 32 neurons and ReLU activation function. The output layer has one neuron. Section 2.4 describes the model design and training process in detail. The table includes the number of samples, accuracy, F1 score, and AUC (area under the curve) for each year from 1998 to 2022 seperately for SPFLS and NSPFLS samples. The difference between the sample sizes of the two models is due to the fact that MD&A sections of the 10-K filings contain more NSPFLS sentences than SPFLS sentences as shown in Table 2 and I use each sentence as a data point during both training and testing. I perform paired t-tests to compare the average performance of two models. */**/*** indicate significance at the 10/5/1% level.

	NSPFLS				**SPFLS**			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Year	**Count**	**Accuracy**	**F1**	**AUC**	**Count**	**Accuracy**	**F1**	**AUC**
1998	69,103	0.6431	0.7828	0.5117	39,146	0.5981	0.7208	0.5356
1999	66,637	0.6589	0.7884	0.5396	31,313	0.6442	0.7680	0.5604
2000	62,834	0.4946	0.6529	0.5157	27,264	0.5384	0.6861	0.5510
2001	103,133	0.4776	0.6465	0.4991	33,397	0.5180	0.6825	0.5119
2002	112,319	0.3811	0.2984	0.4981	32,310	0.7101	0.8305	0.5278
2003	155,604	0.7567	0.8615	0.5070	47,901	0.6193	0.7380	0.5424
2004	170,290	0.7203	0.8355	0.5129	50,163	0.7365	0.8460	0.5524
2005	123,655	0.7309	0.8436	0.5268	49,567	0.7373	0.8468	0.5723
2006	116,470	0.7003	0.8232	0.5241	47,844	0.7007	0.8207	0.5618
2007	112,996	0.5782	0.7327	0.4920	43,620	0.6024	0.7439	0.5370
2008	99,344	0.3664	0.5363	0.5005	36,509	0.3572	0.5264	0.4600
2009	113,373	0.3637	0.1990	0.5332	41,594	0.3909	0.3411	0.5006
2010	113,947	0.6861	0.8131	0.5434	41,700	0.7103	0.8306	0.4600
2011	111,688	0.6114	0.7429	0.5368	40,105	0.5938	0.7230	0.5297
2012	111,967	0.6422	0.7813	0.5522	41,015	0.6526	0.7898	0.5265
2013	115,267	0.6474	0.7694	0.5456	41,403	0.6440	0.7590	0.5824
2014	119,913	0.5593	0.7016	0.5023	42,144	0.5429	0.6683	0.5419
2015	119,809	0.6045	0.7535	0.5388	40,097	0.5598	0.7178	0.5196
2016	118,580	0.6147	0.7371	0.5258	38,235	0.5290	0.6072	0.5444
2017	117,600	0.7472	0.8553	0.5509	38,491	0.7109	0.8197	0.6036
2018	107,898	0.6494	0.7831	0.5285	35,913	0.6238	0.7608	0.5437
2019	98,952	0.4650	0.6348	0.5006	33,458	0.4414	0.6070	0.5177
2020	114,064	0.7442	0.8533	0.4713	41,546	0.7270	0.8419	0.4948
2021	105,621	0.7156	0.8342	0.5167	40,004	0.6851	0.8131	0.5560
2022	99,283	0.6137	0.7519	0.4957	36,592	0.5956	0.7064	0.5655
Average	**110,413.9**	**0.6069**	**0.7205**	**0.5188**	**39,653.2**	**0.6068**	**0.7278**	**0.5360**

Comparison:	**Diff.**	**p-value**
Difference in Accuracy: (6) vs. (2)	-0.0001	0.9943
Difference in F1: (7) vs. (3)	0.0073	0.7643
Difference in AUC: (8) vs. (4)	0.0172**	0.0199

Table 4: Future Sales Growth Probability Correlation

This table reports correlation matrix. $Salegr^{adj}_{SP}$ ($Salegr^{adj}_{NSP}$) is the estimated growth probability using Specific (Non-Specific) forward-looking sentences and further demeaned by fiscal year and industry. $item7\#$ is the number of sentences in MD&A section. $spfls\%$ ($nspfls\%$) is the percentage of Specific (Non-Specific) forward-looking sentences in MD&A section. $LMneg\%$ ($LMpos\%$) is the percentage of negative (positive) words in the whole 10-K filing. $\Delta logsale_t$ is the change in log sales from year $t-1$ to year t. $logbm$ is the log of book-to-market ratio. logsize is the log of total assets.

Variables	$Salegr^{adj}_{SP}$	$Salegr^{adj}_{NSP}$	$item7\#$	$spfls\%$	$nspfls\%$	$LMneg$	$LMpos$	$\Delta logsale_t$	$logbm$	$logsize$
$Salegr^{adj}_{SP}$	1.000									
$Salegr^{adj}_{NSP}$	0.385	1.000								
$item7\#$	-0.017	-0.059	1.000							
$spfls\%$	-0.076	-0.086	-0.196	1.000						
$nspfls\%$	-0.027	-0.088	0.095	0.066	1.000					
$LMneg$	-0.086	-0.078	0.232	-0.071	0.241	1.000				
$LMpos$	-0.041	-0.050	-0.022	0.203	0.272	0.256	1.000			
$\Delta logsale_t$	0.103	0.076	-0.070	0.096	0.051	-0.060	0.070	1.000		
$logbm$	-0.065	-0.047	0.080	-0.178	-0.142	-0.012	-0.254	-0.179	1.000	
$logsize$	0.115	0.068	0.364	-0.011	0.011	0.095	0.051	0.039	-0.377	1.000

Table 5: Characteristic of Firms in Sorted Portfolios

This table reports quintile portfolio characteristics of the $Salegr^{adj}_{SP}$ and $Salegr^{adj}_{NSP}$ sorted at the end of June of year t from 1999 to 2022. $Salegr^{adj}_{SP}$ $(Salegr^{adj}_{NSP})$ is the estimated sales growth probability using firms' Specific-FLS (Non-specific FLS) further adjusted by year and industry. $Size$ is the natural logarithm of the market value of equity. BM is the book-to-market ratio. $SaleY$ is the sales-to-price ratio. BEg is the book equity growth rate. ATg is the asset growth rate. $SALEg$ is the sales growth rate. ROA is the return on assets. $GProf$ is the gross profitability. $Mlev$ is the market leverage. $Blev$ is the book leverage. $Cash$ is the cash-to-assets ratio. Quintiles are formed at the end of June of year t and all accounting variables are from the previous fiscal yearend reportings. I first take the median value for each year and then report the time-series mean for each quintile. Panel A and B reports average characteristics for the firms in each portfolio sorted by $Salegr^{adj}_{SP}$ and $Salegr^{adj}_{NSP}$. H-L is the difference between the average of the top quintile (quintile 5) and the average of the bottom quintile (quintile 1). The t-statistics and p-values are reported in the last two rows.

Panel A: Quintile portfolios sorted on $Salegr^{adj}_{SP}$

Quintiles	$Salegr^{adj}_{SP}$	$Size$	β	BM	$SaleY$	BEg	ATg	$SALEg$	ROA	$GProf$	$Mlev$	$Blev$	$Cash$
1	-1.259	6.125	1.005	0.566	-0.399	0.044	0.039	0.040	0.003	0.274	0.356	0.144	0.131
2	-0.469	6.387	1.012	0.537	-0.447	0.059	0.051	0.059	0.004	0.312	0.343	0.171	0.112
3	0.047	6.521	1.010	0.523	-0.425	0.071	0.064	0.078	0.006	0.332	0.338	0.190	0.099
4	0.529	6.704	1.017	0.498	-0.491	0.081	0.073	0.089	0.007	0.348	0.324	0.199	0.095
5	1.210	6.804	1.014	0.473	-0.526	0.097	0.086	0.102	0.008	0.371	0.313	0.188	0.099
H-L	2.469	0.680	0.009	-0.093	-0.128	0.052	0.048	0.062	0.005	0.097	-0.043	0.044	-0.032
t-stat	197.590	3.486	0.294	-3.191	-1.616	5.880	5.329	4.446	5.577	6.358	-3.527	2.687	-3.471
p-val	0.000	0.001	0.770	0.003	0.113	0.000	0.000	0.000	0.000	0.000	0.001	0.010	0.001

Panel B: Quintile portfolios sorted on $Salegr^{adj}_{NSP}$

Quintiles	$Salegr^{adj}_{NSP}$	$Size$	β	BM	$SaleY$	BEg	ATg	$SALEg$	ROA	$GProf$	$Mlev$	$Blev$	$Cash$
1	-1.227	6.246	1.026	0.543	-0.466	0.051	0.043	0.047	0.003	0.269	0.346	0.142	0.139
2	-0.463	6.387	1.009	0.551	-0.455	0.063	0.058	0.067	0.004	0.310	0.348	0.179	0.106
3	0.041	6.526	1.005	0.533	-0.450	0.071	0.064	0.078	0.006	0.336	0.340	0.187	0.096
4	0.515	6.678	1.006	0.513	-0.459	0.078	0.071	0.085	0.006	0.347	0.331	0.195	0.093
5	1.192	6.681	1.010	0.475	-0.503	0.090	0.080	0.097	0.008	0.370	0.314	0.182	0.110
H-L	2.419	0.436	-0.016	-0.068	-0.037	0.039	0.036	0.049	0.005	0.101	-0.031	0.041	-0.029
t-stat	187.196	2.241	-0.530	-2.336	-0.487	4.165	4.026	3.349	6.296	6.230	-2.482	2.124	-2.243
p-val	0.000	0.030	0.599	0.024	0.629	0.000	0.000	0.002	0.000	0.000	0.017	0.039	0.030

Table 6: Relative Predictive Power of two measures of forward-looking sentences

This table reports predictive regression results of future sales growth and asset growth on two model estimated growth probabilities. $Salegr^{adj}_{SP}$ ($Salegr^{adj}_{NSP}$) is the estimated growth probability using Specific (Non-Specific) forward-looking sentences and further demeaned by fiscal year and industry. *leverage* is the book leverage from the most recent filing. *logbm* is the logrithm of book-to-market ratio. *logsize* is the logrithm of size. Industry-by-year and firm fixed effects are included. T-statistics in parentheses are based on standard errors double clustered by firm and year. */**/*** indicates significance at the 10/5/1% level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	$\Delta logsale_{t+1}$		$1(salegr)_{t+1}$		$\Delta logat_{t+1}$		$1(atgr)_{t+1}$	
$Salegr^{adj}_{SP}$	0.010***	0.011***	0.037***	0.032***	0.013***	0.013***	0.034***	0.031***
	(3.87)	(3.99)	(5.51)	(5.38)	(5.61)	(5.37)	(5.55)	(5.53)
$Salegr^{adj}_{NSP}$		-0.001		0.014**		0.002		0.009**
		(-0.59)		(2.43)		(0.91)		(2.58)
leverage	-0.030***	-0.030***	-0.024	-0.025	-0.131***	-0.131***	-0.310***	-0.311***
	(-3.27)	(-3.18)	(-1.07)	(-1.16)	(-7.23)	(-7.27)	(-12.62)	(-12.70)
logbm	-0.054***	-0.054***	-0.078***	-0.077***	-0.073***	-0.073***	-0.097***	-0.097***
	(-8.67)	(-8.63)	(-11.39)	(-11.11)	(-10.32)	(-10.25)	(-14.99)	(-14.73)
logsize	0.000	0.000	0.020***	0.020***	0.001	0.001	0.021***	0.021***
	(0.08)	(0.09)	(4.26)	(4.27)	(0.63)	(0.63)	(6.05)	(5.97)
LMneg	-2.129***	-2.132***	-8.754***	-8.546***	-3.666***	-3.649***	-14.239***	-14.127***
	(-3.96)	(-3.96)	(-7.90)	(-7.72)	(-7.19)	(-7.23)	(-11.22)	(-11.20)
LMpos	-0.042	-0.072	-3.194	-2.956	-2.991	-2.964	-11.467***	-11.305***
	(-0.02)	(-0.04)	(-0.74)	(-0.71)	(-1.64)	(-1.63)	(-2.92)	(-2.92)
Similarity	-0.122***	-0.121***	-0.006	-0.007	-0.052***	-0.053***	0.078***	0.077***
	(-7.26)	(-7.21)	(-0.33)	(-0.36)	(-2.87)	(-2.91)	(3.61)	(3.60)
%FLS	0.189***	0.190***	-0.034	-0.010	0.074*	0.075*	-0.033	-0.020
	(4.47)	(4.60)	(-0.63)	(-0.19)	(1.88)	(1.95)	(-0.64)	(-0.39)
β_{SP} - β_{NSP}		0.012***		0.019**		0.011***		0.022***
t-statistic		3.16		2.35		3.50		3.73
p-value ($\beta_{SP} > \beta_{NSP}$)		0.002		0.014		0.001		0.001
Industry × Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	51,026	50,971	51,759	51,704	51,729	51,674	51,759	51,704
Adj. R^2	0.140	0.140	0.183	0.183	0.124	0.124	0.142	0.142

Table 7: Future Sales Growth Probability and Stock Return

This table reports standard event study regressions. r_{4d} is the event window buy-and-hold return from $t-1$ to $t+2$ minus CRSP value-weighted market return. $Salegr^{adj}_{SP}$ ($Salegr^{adj}_{NSP}$) is the estimated growth probability using Specific (Non-Specific) forward-looking sentences and further demeaned by fiscal year and industry. $LMneg$ ($LMpos$) is the percentage of negative (positive) words in the whole 10-K filing, $logbm$ is the log of book-to-market ratio, $logsize$ is the log of market capitalization, $turnover$ is the turnover ratio, and 1_{nasdaq} is an indicator variable that equals 1 if the firm is listed on NASDAQ. The sample period is from 1998 to 2022. Industry and year-quarter fixed effects are included. T-statistics in parentheses are based on standard errors double clustered by industry and fiscal year-quarter. */**/*** indicate significance at the 10/5/1% level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	r_{4d}		r_{10d}		r_{30d}		r_{60d}		r_{120d}		r_{180d}	
$Salegr^{adj}_{SP}$	0.116**		0.205**		0.340**		0.463***		0.764***		1.252***	
	(2.34)		(2.61)		(2.44)		(2.74)		(4.07)		(5.69)	
$Salegr^{adj}_{NSP}$		0.070		0.093		0.196		0.203		0.379*		0.547**
		(1.37)		(1.33)		(1.44)		(1.10)		(1.80)		(2.22)
$LMneg$	-11.749	-14.973	-16.701	-22.814	-6.741	-13.966	10.919	0.556	-24.067	-41.934	-39.980	-67.222
	(-1.01)	(-1.32)	(-1.00)	(-1.38)	(-0.21)	(-0.44)	(0.18)	(0.01)	(-0.30)	(-0.53)	(-0.53)	(-0.93)
$LMpos$	-49.242	-49.984	-62.530	-63.957	-112.008	-113.328	-116.442	-122.118	-167.867	-176.342	-279.545	-293.549
	(-1.47)	(-1.48)	(-1.13)	(-1.18)	(-1.44)	(-1.46)	(-0.83)	(-0.86)	(-0.95)	(-1.02)	(-1.39)	(-1.48)
$logbm$	0.215*	0.210*	0.419*	0.410*	0.670*	0.663*	0.808*	0.807*	1.450**	1.431**	2.212***	2.193**
	(1.84)	(1.82)	(1.76)	(1.74)	(1.90)	(1.88)	(1.90)	(1.88)	(2.48)	(2.43)	(2.71)	(2.66)
$logsize$	0.205***	0.208***	0.285***	0.293***	0.336*	0.349*	0.310	0.333	0.614*	0.650**	1.151**	1.220**
	(3.88)	(3.96)	(2.95)	(3.08)	(1.73)	(1.81)	(1.35)	(1.47)	(1.92)	(2.06)	(2.42)	(2.58)
$turnover$	-0.403***	-0.402***	-0.630**	-0.617**	-1.003**	-0.982**	-1.316**	-1.294**	-2.429***	-2.402***	-3.279***	-3.247***
	(-2.95)	(-2.92)	(-2.63)	(-2.58)	(-2.32)	(-2.27)	(-2.48)	(-2.43)	(-3.26)	(-3.24)	(-3.39)	(-3.40)
1_{nasdaq}	-0.005	-0.002	-0.045	-0.044	0.056	0.074	-0.022	0.008	0.189	0.246	0.760	0.852
	(-0.04)	(-0.02)	(-0.21)	(-0.20)	(0.17)	(0.23)	(-0.06)	(0.02)	(0.32)	(0.42)	(1.14)	(1.28)
Year×Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	60,893	61,260	60,893	61,260	60,893	61,260	60,893	61,260	60,893	61,260	60,893	61,260
Adj. R^2	0.014	0.014	0.026	0.026	0.044	0.044	0.041	0.041	0.049	0.048	0.054	0.053

Table 8: Future Sales Growth Probability and Stock Market Response

This table reports standard event study regressions. $|r_{4d}|$ is the absolute value of event window buy-and-hold return from $t-1$ to $t+2$ minus CRSP value-weighted market return. $Salegr^{adj}_{SP}$ ($Salegr^{adj}_{NSP}$) is the estimated growth probability using Specific (Non-Specific) forward-looking sentences and further demeaned by fiscal year and industry. $LMneg$ ($LMpos$) is the percentage of negative (positive) words in the whole 10-K filing, *logbm* is the log of book-to-market ratio, *logsize* is the log of market capitalization, *turnover* is the turnover ratio, and 1_{nasdaq} is an indicator variable that equals 1 if the firm is listed on NASDAQ. The sample period is from 1998 to 2022. Industry and fiscal year-quarter fixed effects are included. T-statistics in parentheses are based on standard errors double clustered by industry and year-quarter. */**/*** indicate significance at the 10/5/1% level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	$\|r_{4d}\|$		$\|r_{10d}\|$		$\|r_{30d}\|$		$\|r_{60d}\|$	
$Salegr^{adj}_{SP}$	-0.092***		-0.124***		-0.114**		-0.113	
	(-4.01)		(-3.91)		(-2.09)		(-1.36)	
$Salegr^{adj}_{NSP}$		-0.040		-0.092		-0.139		-0.155
		(-1.12)		(-1.42)		(-1.53)		(-1.54)
$FLS\%$	2.732***	2.712***	4.154***	4.053***	7.573***	7.439***	10.993***	10.616***
	(3.80)	(3.72)	(4.59)	(4.23)	(5.59)	(5.26)	(5.90)	(5.46)
$LMneg$	65.168***	66.233***	96.701***	97.412***	118.838***	118.212***	205.192***	204.881***
	(5.77)	(5.95)	(7.01)	(7.03)	(5.47)	(5.47)	(6.59)	(6.57)
$LMpos$	61.058**	63.080**	113.490***	113.081***	179.438***	176.006***	243.770***	239.680***
	(2.06)	(2.11)	(3.19)	(3.16)	(3.17)	(3.13)	(2.75)	(2.73)
logbm	-0.296***	-0.297***	-0.323**	-0.322**	-0.585**	-0.581**	-0.682**	-0.678**
	(-3.10)	(-3.15)	(-2.07)	(-2.07)	(-2.45)	(-2.45)	(-2.39)	(-2.41)
logsize	-0.883***	-0.886***	-1.145***	-1.146***	-1.773***	-1.770***	-2.352***	-2.347***
	(-21.23)	(-21.22)	(-22.67)	(-22.66)	(-23.46)	(-24.00)	(-21.52)	(-21.75)
turnover	0.879***	0.878***	1.299***	1.293***	1.935***	1.939***	2.694***	2.702***
	(8.72)	(8.79)	(9.84)	(9.96)	(11.41)	(11.32)	(10.20)	(10.24)
1_{nasdaq}	0.025	0.020	0.036	0.040	0.100	0.103	0.249	0.251
	(0.24)	(0.19)	(0.25)	(0.28)	(0.53)	(0.54)	(0.95)	(0.94)
Year×Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	60,893	61,260	60,893	61,260	60,893	61,260	60,893	61,260
Adj. R^2	0.153	0.153	0.173	0.173	0.176	0.176	0.161	0.161

Table 9: Analyst Revenue Forecast Revision and Future Sales Growth Probability

This table reports regressions of analysts' one-year-ahead annual revenue forecast revision around disclosure on two measures of sales growth probability. The first two columns use full sample and the last three columns use downward revisions only. $Revision_{30d}$ is the percentage change of analyst's revision on firm's annual revenue forecast around 10-K disclosure and earnings announcement, up to 30 days after 10-K filing date. $Salegr^{adj}_{SP}$ ($Salegr^{adj}_{NSP}$) is the estimated growth probability using Specific (Non-Specific) forward-looking sentences and further demeaned by fiscal year and industry. *Similarity* is the jaccard similarity of item 7 of 10-K filing to previous year's, $LMneg$ ($LMpos$) is the percentage of negative (positive) words in the whole 10-K filing, *Leverage* is the book leverage ratio, *logbm* is the log of book-to-market ratio, *logsize* is the log of market capitalization, *Surprise* is the percentage difference of reported revenue and analysts' consensus before earnings announcement. The sample period is from 1998 to 2022. Industry by year and firm fixed effects are included. T-statistics in parentheses are based on standard errors double clustered by firm and industry by year. */**/*** indicate significance at the 10/5/1% level.

	Full Sample		Downward Revisions ($Revision_{30d} < 0$)		
	(1)	(2)	(3)	(4)	(5)
	$Revision_{30d}$	$Revision_{30d}$	$Revision_{30d}$	$Revision_{30d}$	$Revision_{30d}$
$Salegr^{adj}_{SP}$	0.068		0.081*		0.096**
	(1.49)		(1.84)		(2.09)
$Salegr^{adj}_{NSP}$		-0.071*		-0.014	-0.047
		(-1.86)		(-0.36)	(-1.12)
Similarity	0.538	0.625	1.154**	1.173**	1.168**
	(1.11)	(1.29)	(2.38)	(2.41)	(2.40)
$LMneg$	-4.273	-6.019	-23.503	-24.774	-24.259
	(-0.21)	(-0.29)	(-1.14)	(-1.20)	(-1.18)
$LMpos$	99.671*	102.482*	57.443	57.483	58.084
	(1.88)	(1.92)	(1.05)	(1.05)	(1.07)
$FLS\%$	-0.380	-0.569	-5.984***	-6.124***	-6.043***
	(-0.28)	(-0.42)	(-3.98)	(-4.08)	(-4.00)
Leverage	0.138	0.153	1.114*	1.120*	1.111*
	(0.26)	(0.29)	(1.94)	(1.95)	(1.94)
Logbm	-0.554***	-0.563***	-0.107	-0.112	-0.108
	(-5.47)	(-5.59)	(-1.07)	(-1.13)	(-1.09)
Logsize	0.829***	0.833***	0.894***	0.901***	0.894***
	(7.21)	(7.23)	(7.58)	(7.70)	(7.59)
Surprise	0.112***	0.111***	0.034***	0.034***	0.034***
	(8.67)	(8.63)	(4.77)	(4.76)	(4.78)
β_{SP} - β_{NSP}					0.143
t-statistic					2.31
p-value ($\beta_{SP} > \beta_{NSP}$)					0.0439
Industry-Year FE	Yes	Yes	Yes	Yes	Yes
Firm FE	Yes	Yes	Yes	Yes	Yes
Observations	118,017	118,198	58,363	58,471	58,362
Adj. R^2	0.261	0.261	0.450	0.450	0.450

Table 10: Fama-MacBeth Regressions

This table documents results from Fama-MacBeth Regressions of the form $r_{tj} = \beta' X_{t-1,j} + \epsilon_{tj}$. Cross-sectional regressions are run from May to November each year from 1999 to 2023. The characteristics $X_{t-1,j}$ include $Salegr_{SP}^{adj}$, the log of market capitalization ($logsize$), the log of the book-to-market ratio ($logbm$), gross profitability ($GProf$), Asset Growth (ATg), momentum ($r_{12,1}$), and short-term reversals ($r_{1,0}$). $Salegr_{SP}^{adj}$ is computed as before. Independent variables are winsorized at 1 percent level. The t-statistics are in brackets and calculated using 6 periods of Newey-West lags.

	Regression of the form $r_{tj} = \beta' X_{t-1,j} + \epsilon_{tj}$								
Coef.	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$Salegr_{SP}^{adj}$	0.16 [3.87]	0.12 [3.59]	0.15 [3.92]	0.14 [3.53]	0.17 [4.09]	0.14 [3.55]	0.16 [4.13]		0.12 [4.25]
$logsize$		0.08 [1.30]						0.06 [0.84]	0.06 [1.06]
$logbm$			-0.00 [-0.03]					0.11 [1.00]	-0.00 [-0.00]
$GProf$				0.66 [2.25]				0.71 [2.92]	0.33 [1.03]
ATg					-0.50 [-3.27]			-0.61 [-4.16]	-0.51 [-3.40]
$r_{12,1}$						0.61 [2.15]		0.67 [2.09]	0.54 [1.95]
$r_{1,0}$							-0.68 [-0.98]	-1.71 [-3.05]	-1.30 [-2.04]

Appendix A. Specific Forward-looking Sentences from 2014 Agios 10-K Report

This table displays Specific Forward-looking Sentences and corresponding estimated Growth Probability from Agios Pharmaceuticals's 10-K for the fiscal year of 2014.

Growth Probability	Specific Forward-looking Sentence
0.09	We are also unable to predict when, if ever, material net cash inflows will commence from AG-221, AG-120, AG-348, or any of our other product candidates.
0.14	We intend to begin a global registration program for AG-221 in year t + 1 for IDH2-mutant positive hematologic malignancies.
0.14	We anticipate that our expenses will increase significantly as we continue to advance and expand clinical development activities for our lead programs, AG-221, AG-120 and AG-348; continue to discover and validate novel targets and drug product candidates; expand and protect our intellectual property portfolio; hire additional commercial, development and scientific personnel; and continue to operate as a publicly-traded company.
0.14	We expect to continue to incur significant expenses and operating losses over the next several years.
0.14	In the future, we will seek to generate revenue from a combination of product sales and upfront payments, extension payments, cost reimbursements, milestone payments, and royalties on future product sales in connection with our Celgene collaboration or other strategic relationships.
0.17	Our commercial revenues, if any, will be derived from sales of medicines that we do not expect to be commercially available for many years, if at all.
0.18	Financial Operations Overview Revenue Through December 31, year t, we have not generated any revenue from product sales and do not expect to generate any revenue from product sales in the near future.
0.19	Celgene would lead and fund global development and commercialization of development candidates for which they exercise their option to obtain a co-commercialization license, and we would retain development and commercialization rights in the United States for development candidates for which we exercise our option to retain a split license.
0.22	Based on these findings, we plan to initiate multiple expansion cohorts in the first half of year t + 1.
0.24	We intend to initiate a global registration program for AG-120 in IDH1-mutant positive hematologic malignancies by early year t + 2.
0.26	We anticipate that our general and administrative expenses will increase in the future to support continued research and development activities, potential commercialization of our product candidates and increased costs of operating as a public company.
0.29	On all programs, we are eligible to receive up to 120 million in milestone-based payments as well as royalties on any sales.
0.31	Accordingly, we will need to obtain substantial additional funding in connection with our continuing operations.
0.32	Under these agreements, as of December 31, year t we are obligated to pay up to 40.2 million to these vendors.

Continued on next page

continued from previous page

Growth Probability	Specific Forward-looking Sentence
0.32	We may also receive future milestone or royalty payments under the Celgene collaboration agreement.
0.33	The extension marks the final year for the discovery phase and Celgene will maintain its exclusive option to drug candidates that emerge from our cancer metabolism research platform through April year t + 2.
0.34	In addition, if we obtain marketing approval for any of our product candidates, we expect to incur significant commercialization expenses related to product sales, marketing, manufacturing and distribution to the extent that such sales, marketing and distribution are not the responsibility of Celgene or other collaborators.
0.35	Furthermore, we expect to continue to incur additional costs associated with operating as a public company.
0.35	Until such time, if ever, as we can generate substantial product revenues, we expect to finance our cash needs through a combination of equity offerings, debt financings, collaborations, strategic alliances and licensing arrangements.
0.37	Upon Celgene s exercise of its exclusive option under the terms of our agreement, Celgene would lead development and commercialization outside the United States for AG-120, and we and Celgene would equally fund the global development costs of AG-120 that are not specific to any particular region or country.
0.38	We are also required to make payments in amounts ranging from 7.0 to 25 for non-royalty income received from any sublicense of the rights granted to us under such agreements.
0.40	We expect that our existing cash, cash equivalents and marketable securities as of December 31, year t, together with 3.8 million in anticipated refundable income taxes, anticipated interest income, the 20.0 million anticipated from Celgene as a result of its exercise of its option in December year t to extend the discovery term of our agreement for an additional year and anticipated expense reimbursements under our collaboration agreement with Celgene will enable us to fund our operating expenses and capital expenditure requirements until at least late year t + 3.
0.40	We will also receive an additional 20.0 million extension payment as a result of Celgene electing to extend the discovery phase until April year t + 2.
0.41	We will receive a 20.0 million payment as a result of the extension, which we expect to receive in the second quarter of year t + 1.
0.42	We expect to receive additional consideration under our collaboration agreement with Celgene related to certain development services to be performed.
0.42	Additionally, we anticipate increased costs associated with being a public company including expenses related to services associated with maintaining compliance with exchange listing and Securities and Exchange Commission requirements, insurance, and investor relations costs.
0.44	Funding requirements We expect our expenses to increase in connection with our ongoing activities, particularly as we continue the research, development and clinical trials of, and seek marketing approval for, our product candidates.
0.44	We expect research and development costs to increase significantly for the foreseeable future as our product candidate development programs progress.
0.46	We expect to provide final results from the MAD study in year t + 1 and to initiate a phase 2 study of AG-348 in patients with PK deficiency in the first half of year t + 1.

Continued on next page

continued from previous page

Growth Probability	Specific Forward-looking Sentence
0.46	We will make separate milestone payments when we accumulate net profits of 5.0 million, 50.0 million and 250.0 million, respectively, from sales of the product.
0.47	We expect to provide the first data from the natural history study in year t + 1.
0.49	Celgene would be eligible to receive royalties on any net sales in the U.S. We would be eligible to receive royalties on any net sales outside the U.S. and up to 120.0 million in payments on achievement of certain milestones.
0.49	In addition to our existing cash, cash equivalents and marketable securities, we are eligible to earn a significant amount of milestone payments and are entitled to cost reimbursement under our collaboration agreement with Celgene.
0.50	Under the agreement, the applicable party will pay to the other party a royalty based on worldwide net sales of products.
0.51	We are obligated to pay the licensor up to 100,000 in milestone payments, contingent upon the issuance of certain patents.
0.55	Celgene would be responsible for development and commercialization costs specific to countries outside the United States, and we would be responsible for development and commercialization costs specific to the United States.
0.57	We have worldwide development and commercial rights to AG-348 and expect to fund the future development and commercialization costs related to this program.
0.67	The license agreements require us to pay ongoing annual maintenance payments, initially totaling 45,000 per year and increasing to 70,000 per year beginning in year t + 2, as well as reimburse certain patent costs previously incurred by the licensors, as applicable.

Appendix B. Orthogonalization of Idiosyncratic and Systematic Forward-looking Growth Measures

Table B1: Relative Predictive Power of two Orthogonolized Forward-looking Growth Measures on Stock Returns

This table reports standard event study regressions. r_{4d} is the event window buy-and-hold return from $t-1$ to $t+2$ minus CRSP value-weighted market return. $Salegr^{\top}_{SP}$ ($Salegr^{\top}_{NSP}$) is the estimated growth measure $Salegr^{adj}_{SP}$ ($Salegr^{adj}_{NSP}$) orthogonolized on $Salegr^{adj}_{NSP}$ ($Salegr^{adj}_{SP}$) by year. $LMneg$ ($LMpos$) is the percentage of negative (positive) words in the whole 10-K filing, lev is the book leverage from the most recent filing, $logbm$ is the logrithm of book-to-market ratio, $logsize$ is the logrithm of size, $\Delta logsale_t$ is the log difference of sales growth compared to last fiscal year. Industry-by-fyear and firm fixed effects are included. T-statistics in parentheses are based on standard errors double clustered by firm and fyear. */**/*** indicates significance at the 10/5/1% level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	r_{4d}		r_{30d}		r_{120d}		r_{180d}	
$Salegr^{\top}_{SP}$	0.104**		0.282**		0.671***		1.156***	
	(2.07)		(2.35)		(3.78)		(5.14)	
$Salegr^{\top}_{NSP}$		0.036		0.056		0.071		0.060
		(0.76)		(0.48)		(0.34)		(0.23)
$LMneg$	-13.584	-15.063	-11.395	-15.851	-32.042	-43.341	-52.277	-72.418
	(-1.20)	(-1.36)	(-0.35)	(-0.49)	(-0.40)	(-0.55)	(-0.69)	(-0.99)
$LMpos$	-53.307	-53.955	-121.704	-124.675	-188.917	-197.839	-305.653	-322.858
	(-1.58)	(-1.61)	(-1.61)	(-1.67)	(-1.10)	(-1.17)	(-1.57)	(-1.67)
$logbm$	0.211*	0.210*	0.664*	0.659*	1.424**	1.411**	2.183**	2.160**
	(1.83)	(1.80)	(1.89)	(1.87)	(2.44)	(2.40)	(2.68)	(2.63)
$logsize$	0.206***	0.210***	0.345*	0.356*	0.630*	0.657**	1.180**	1.227**
	(3.90)	(3.96)	(1.78)	(1.84)	(1.95)	(2.04)	(2.47)	(2.56)
$turnover$	-0.405***	-0.403***	-1.001**	-0.995**	-2.429***	-2.418***	-3.277***	-3.258***
	(-2.96)	(-2.94)	(-2.32)	(-2.30)	(-3.27)	(-3.26)	(-3.40)	(-3.38)
1_{nasdaq}	0.002	0.004	0.066	0.073	0.206	0.224	0.785	0.816
	(0.01)	(0.03)	(0.20)	(0.22)	(0.35)	(0.38)	(1.18)	(1.21)
fyear×quarter	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
industry	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	60,817	60,817	60,817	60,817	60,817	60,817	60,817	60,817
Adj. R^2	0.014	0.013	0.044	0.044	0.048	0.048	0.054	0.053

Table B2: Relative Predictive Power of two Orthogonolized Forward-looking Growth Measures on Stock Market Response

This table reports standard event study regressions. $|r_{4d}|$ is the absolute value of event window buy-and-hold return from $t-1$ to $t+2$ minus CRSP value-weighted market return. $Salegr^{\top}_{SP}$ ($Salegr^{\top}_{NSP}$) is the estimated growth measure $Salegr^{adj}_{SP}$ ($Salegr^{adj}_{NSP}$) orthogonolized on $Salegr^{adj}_{NSP}$ ($Salegr^{adj}_{SP}$) by year. $logbm$ is the logrithm of book-to-market ratio. $logsize$ is the logrithm of size. $\Delta logsale_t$ is the log difference of sales growth compared to last fiscal year. Industry-by-fyear and firm fixed effects are included. T-statistics in parentheses are based on standard errors double clustered by firm and fyear. */**/*** indicates significance at the 10/5/1% level.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	$\|r_{4d}\|$		$\|r_{10d}\|$		$\|r_{30d}\|$		$\|r_{60d}\|$	
$Salegr^{\top}_{SP}$	-0.085***		-0.089***		-0.048		-0.029	
	(-3.84)		(-3.76)		(-1.16)		(-0.35)	
$Salegr^{\top}_{NSP}$		-0.009		-0.053		-0.108		-0.148
		(-0.25)		(-0.82)		(-1.33)		(-1.64)
$FLS\%$	2.776***	2.788***	4.262***	4.206***	7.716***	7.560***	11.126***	10.899***
	(3.89)	(3.83)	(4.70)	(4.37)	(5.79)	(5.31)	(5.84)	(5.61)
$LMneg$	65.983***	67.376***	98.491***	99.637***	120.318***	120.348***	207.525***	206.928***
	(5.77)	(5.92)	(6.98)	(7.11)	(5.53)	(5.55)	(6.63)	(6.63)
$LMpos$	62.279**	63.315**	115.036***	115.286***	181.237***	179.842***	244.762***	242.309***
	(2.07)	(2.08)	(3.16)	(3.15)	(3.14)	(3.18)	(2.71)	(2.73)
$logbm$	-0.295***	-0.293***	-0.320**	-0.319**	-0.579**	-0.581**	-0.672**	-0.675**
	(-3.10)	(-3.08)	(-2.05)	(-2.05)	(-2.44)	(-2.45)	(-2.38)	(-2.39)
$logsize$	-0.885***	-0.889***	-1.148***	-1.152***	-1.777***	-1.779***	-2.356***	-2.357***
	(-21.21)	(-21.37)	(-22.60)	(-22.89)	(-23.45)	(-23.93)	(-21.45)	(-21.53)
$turnover$	0.879***	0.878***	1.296***	1.294***	1.934***	1.933***	2.689***	2.688***
	(8.74)	(8.73)	(9.88)	(9.85)	(11.36)	(11.30)	(10.19)	(10.17)
1_{nasdaq}	0.020	0.018	0.034	0.033	0.095	0.096	0.248	0.250
	(0.20)	(0.17)	(0.24)	(0.22)	(0.51)	(0.51)	(0.95)	(0.95)
fyear×quarter	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
industry	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N	60,817	60,817	60,817	60,817	60,817	60,817	60,817	60,817
Adj. R^2	0.153	0.153	0.173	0.173	0.176	0.176	0.161	0.161