Uncovering Information: Can AI Tell Us Where to Look?*

Anna Costello Bradford Levy Valeri Nikolaev

First Draft: April 2023
Current Draft: December 2024

Abstract

Amidst the sea of content in the digital era, identifying information that is genuinely relevant is a challenging task. This study uses large language models to uncover new, or surprising, information in textual data. Specifically, we train LLMs using data from the financial domain to model investors' prior beliefs over narrative content and, thus, identify new information that a company communicates. Our models, which are tailored to each firm and time, allow us to determine when a sentence or a paragraph in a corporate filing contains a surprise. The models are trained with fixed knowledge cutoffs to eliminate lookahead bias. We also generate a summary measure of the amount of information released in each document, and we show that this measure explains a large portion of market response to disclosed information and predicts firms' future returns. In sum, our study highlights the importance of identifying new and salient information in today's complex, data-driven landscape.

JEL Classifications: C45; D8; D84; G14; M41

Keywords: Beliefs; AI; large language models; GPT; information in markets; market efficiency

*We thank Ralph Koijen, Stefan Nagel, and seminar participants at Harvard University and the University of Chicago for helpful comments and suggestions. The authors gratefully acknowledge financial support from the University of Chicago Booth School of Business and the Center for Applied AI.

1. Introduction

Information that is important to various stakeholders, including new or prospective investors, labor market participants, and product markets, is often buried in the sea of narrative text released by a firm, and finding this relevant information is often like trying to find a "needle in a haystack." The volume and various forms of unstructured narrative disclosures has grown substantially over time, giving rise to the need for more efficient ways to process such information. In this paper, we combine the principles of information theory with recent advancements in generative AI technology to help stakeholders quickly and easily sort through unstructured disclosures in order to identify what is important.

Measuring information in narrative disclosures presents a tall challenge. Take, for example, the average 10-K containing over 60,000 words. Unveiling its information requires a comprehensive analysis of the words, sentences, and their broader context both within the document and within the history of disclosures that came before it. Prior studies tackle this challenge in a few ways. For example, several studies measure properties of the text like "tone" (e.g., Li (2008)) or the proportion of words belonging to a reference dictionary (Loughran and McDonald, 2011). While straightforward, these approaches rely on the researcher's subjective classification of words and their meaning, and they offer little scope for assessing whether a disclosure is surprising, thereby updating investors' prior beliefs. Other work analyzes changes in text over time by, for example, comparing a firm's newly released 10K to the prior 10K released by the firm in the prior year (Cohen et al. (2020)). While insightful, the sheer volume of information released by the firm over the course of the year makes an apples-to-apples comparison from one disclosure type to another difficult to implement.

Our paper addresses these challenges by building a measure of information that is context-specific, and it accounts for every type of disclosure released by the firm and filed in EDGAR. To detect and measure newsworthy information within financial disclosures, we start from the

premise that surprising, or "new" disclosures are more informative than stale information that has previously been released by the firm. This logic follows the literature on asset pricing in capital markets, where a disclosure is informative to the extent that it updates beliefs about firms' cash flows or risks associated with these cash flows (Fama, 1970). When considering *numerical* disclosures, developing such a measure is relatively straightforward. For instance, much of the prior literature quantifies the news in earnings based on the surprise, defined as the difference between realized and expected earnings (Ball and Brown, 1968). Similarly, the approach we follow in this paper develops a measure of the surprise of a piece of narrative text, relative to all narrative text that came from any previous disclosure released by the firm. The flexibility of our measure allows us to assess the surprise of a piece of information, relative to all disclosure that came before it.

We rely on a foundational idea from the information theory literature pioneered by Shannon (1948), which shows that when quantifying new information, unlikely content is more informative than observing content that the receiver believes will occur with high probability. Shannon's analysis implies that one can calculate a measure of new information in a text as the sum of the log probabilities of each token (word), conditional on the preceding context (prior tokens).

Calculating the probability that a word will appear in a text requires an estimate of the conditional probability over the words that could occur, conditional on prior words. Notably, large language models (LLMs) have been developed specifically for this task and demonstrated the ability to do so very accurately (Radford et al., 2019). As such, we begin by pretraining an LLM *from scratch* on a cross-section of firms' narrative disclosures prior to 2007.[1] We then continue pretraining the cross-sectional model on each firm's individual time series of disclosures prior to 2007 to yield a firm-specific LLM for each firm in the sample. Finally, we iteratively apply each firm-specific LLM to the firm's new disclosures to

[1] We eliminate numerical tables from the training data in order to focus the analysis on narrative content.

measure the information content in them and then update the firm-specific LLM by continued pretraining on the new disclosure. By pretraining from scratch with a fixed knowledge cutoff of 2007 and iteratively updating each firm-specific LLM, we eliminate concerns regarding look-ahead bias (Sarkar and Vafa, 2024). Notably, this measure is remarkably flexible in that it allows for capturing information at the word, sentence, paragraph, section, and document levels. It thus enables us to pinpoint the location in a potentially very lengthy text where significant amounts of new information reside (see Figure 2 for an example of our method applied to the disclosure which first mentions the Apple iPhone).

Our sample includes *all* disclosures filed on EDGAR by companies during the period 1996-2023 (Wang and Levy, 2024). Figure 1 depicts how these disclosures are used in our model training and inference process. Our pretraining data includes all disclosures filed prior to 2007 consisting of roughly 35.5 billion tokens, whereas our inference (out-of-sample) data includes disclosures from 2007 and thereafter. The inference data cover a sample of 500 firms stratified by firm size in 2006, and consists of nearly 278,000 filings containing roughly 1.7 billion tokens.

We use our method to compute a measure of new, or surprising, textual information (henceforth we refer to this generically as *Information*, for simplicity). We compute both a disclosure-wide measure of information, which is an average for a particular firm's unique filing (e.g., their 2024 10K), as well as more granular measures of new information within a particular filing. With our flexible measures of new information in hand, we first turn to descriptive evidence of where novel news is most prevalent. Figure 3) shows that much of the new information released by a firm is located in exhibits - which often contain new procurement, debt, and other types of contracts. The figure also shows that current reports (Form 8-K) contain much more information than annual and quarterly reports. Further, while prior work typically focuses on the news in earnings announcements (Item 2.02 of the 8-K) other types of events such as changes in accountants, bankruptcy, and warnings that stakeholders should not rely on previously issued financial statements are much more

newsworthy, according to our measure. These results point to where investors should focus their attention when seeking new information within lengthy text documents.

In addition to measuring where information is released, we also consider *when* it is released (see Figure 4). Using our measure, we see that the periodic nature of earnings announcements (EAs), quarterly reports, and annual reports leads to a "lumpy" and infrequent release of information, amounting to 10.2%, 14.3%, and 20.3% of all content in the highest quartile of information, respectively. In contrast, the other 55.2% of high information content arrives almost continuously via non-EA current reports (Form 8-K) and other filings. Turning to the bottom panel of Figure 4, we explore the proportion of market returns realized throughout the reporting year. The distribution of market returns largely mirrors the timing of information content in the top panel. Thirteen percent of absolute returns over the reporting year can be attributed to EAs and 10-K/Qs, 22.1% to non-EA 8-Ks and other filings, and the remainder to other sources of volatility. The results of this figure highlight a key insight of our paper: information that is potentially decision-relevant is released by the firm continuously throughout the year, and prior work focusing only on a select set of filings potentially misses a large chunk of this news. This highlights the importance and novelty of our measure of information which allows for flexibly capturing up-to-date "new information," regardless of form type, structure and timing.

Having used each firm and time-specific LLM to compute a measure of information in each disclosure, we next assess the extent to which our measure captures information within the framework of Fama (1970). If prices impound all publicly available information, and our measure captures this information, then we expect the measure to explain short window price reactions. Consistent with our measure capturing decision-relevant information, we find that our measure explains the majority of the market reaction. For example, the event-day increase in absolute return associated with a current report (Form 8-K) is 67.1% relative

to the sample mean.[2] After controlling for our measure, we find that current reports in the lowest decile of information are associated with an 17.9% increase in absolute return, while those in the highest decile of information are associated with a 98.6% increase. Additionally, we find that the adjusted R^2 increases roughly four-fold after including our measure in the regression. We find similar patterns for trading volume and other form types, consistent with the notion that narrative content is responsible for the majority of the information in the financial disclosures we consider.

In our second set of tests, we consider the interaction between information and sentiment. Specifically, a large literature finds that the market reacts to sentiment (see, e.g., Loughran and McDonald (2016) for a review). Our insight is that sentiment should only explain market reactions if it updates investors' prior beliefs. For example, consider a firm that previously announced it is entering receivership. Then consider a subsequent disclosure, where the firm restates that announcement for context and discloses new information alongside it. An efficient market which has already priced the original disclosure should not be a function of the portion of the restated disclosure–unless the fact that the firm is restating it contains information incremental to the original disclosure. Along these lines, we expect an interaction between the sentiment and information of content, which one can think of as an information-weighted measure of sentiment.

Consistent with prior literature (e.g., Loughran and McDonald (2011)), we find that the abnormal return to Form 10-K/Q filings in the lowest decile of sentiment is roughly 22 bps. However, this reaction is not statistically distinct from the reaction to forms in less negative deciles of sentiment. For Form 8-K filings, the difference from the lowest (most negative) to highest (most positive) decile of sentiment is 0.66% for Forms 10-K/Q (Forms 8-K). However, when we weight the sentiment of each word by its information, we find that the difference is

[2] In Table 5 Panel B, the coefficient on *Event*[0] represents the average incremental absolute return relative to the non-event window. The percent change of 67.1%, and others throughout the paper, are thus computed using this incremental change relative to the sample mean: 67.1% = 1.39 / 2.07

4.84% (3.87%) for Forms 10-K/Q (Forms 8-K). These results are consistent with the notion that the market reacts to *informative* sentiment rather than sentiment more generally. They also highlight the importance of (precisely) measuring both information and sentiment when processing a new disclosure.

In our third set of tests, we explore the implications of limited attention on what is considered "news." Our earlier analyses suggest that EAs and 10-K/Qs are an important source of information which tends to arrive at regular intervals (see Figure 4). In contrast, 8-Ks and other forms are also an important source of information which tend to arrive *sporadically* thereby requiring investors to continuously monitor firm disclosure channels. To explore the implications of being constrained to process particular subsets of disclosures, e.g., only annual reports, we take the firm-specific models pretrained on data prior to 2007, and iteratively apply and update them as was done previously, except that the models are only updated when a given disclosure is from a particular subset: annual reports, annual and quarterly reports, or current reports.

Across our tests of market reaction and sentiment, we find that relying solely on annual reports or annual and quarterly reports is associated with *perceived* under-reaction to what is "news" to these investors. For example, in the case of investors who only consume annual and quarterly reports, the absolute return associated with 10-K/Q content in the highest decile of information–to these investors–is roughly 20.3% lower than that associated with the lowest decile of information. In contrast, investors relying solely on current reports generally see market reactions which directionally match their beliefs but are somewhat muted relative to measuring "news" using all disclosures. This evidence is consistent with our earlier findings that current reports are an important disclosure channel which requires near constant monitoring.

In our final set of tests we return to the question of which content is not only informative but also decision relevant, i.e., useful for predicting future firm performance. Following prior

literature, we predict the binary market sentiment subsequent to the release of disclosures. Across a variety of methods we find that weighting content by our information measure leads to more accurate predictions. Further, when we allow models to learn which pieces of content are most useful for predictions, we find that a roughly 80% overlap between this content and what our measure labels as high information. This is consistent with (i) our measure capturing information, and (ii) the content on EDGAR being largely relevant to future performance.

The primary contribution of our study is the development of a method for identifying information within unstructured data, e.g., narrative disclosures. As such, our study and findings are of interest to a broad audience beyond accounting and finance. We show that LLMs can be used to form priors over narrative content, which can then be used to identify information in new content. Our study is the first to do this in the context of financial reporting, and it significantly adds to prior work on textual analysis in this setting. Specifically, our study recognizes that text is multidimensional, often unstructured, and difficult to directly compare across documents. Our parsimonious measure of information accounts for all of these complexities.

Beyond the methodological nature of our paper, we are also able to provide several novel insights. First, we document that while there is a large literature focused on EAs, annual, and quarterly reports, there exist other forms that contribute significantly to firms' information environments and stock volatility. In this sense, we are able to account for an additional source of volatility and highlight challenges posed by these disclosure channels to attention constrained investors.

Our work is also related to recent papers which use the output of LLMs to predict returns. These studies include asking an LLM whether a stock will rise or fall (Lopez-Lira and Tang, 2023), generating summaries of lengthy content (Kim et al., 2023), and extracting dense representations of content (Chen et al., 2022). A common theme across these studies is that

an LLM is used to transform a source document into another representation. As a result, it is not immediately clear why these representations predict returns–although Chen et al. (2022) provide evidence that LLMs are better at unraveling the sentiment of text. In contrast to these studies, our work preserves the source document and uses an LLM to overlay a measure of information (see, e.g., Figure 2) thereby "uncovering the information" and eliminating the possibility that part of the output is a hallucination–one of the primary barriers to using LLMs in fields such as finance and law where the veracity of information is crucial (Ji et al., 2023). In contemporaneous work, Sarkar and Vafa (2024) highlights that commercial LLMs appear to contain lookahead bias which could be a source of return predictability. In our paper, we train LLMs from scratch with fixed knowledge cutoffs such that all of our tests are out of sample relative to the training data.

Our paper is also related to the large literature on costly information processing. The closest paper in this area is Cohen et al. (2020), who study annual and quarterly reports. Specifically, they measure the similarity of text across these documents as a measure of informativeness. We extend this idea by forming priors over *all* disclosures. We find evidence consistent with Cohen et al. (2020) in the sense that (i) annual and quarterly reports do appear to be informative disclosures, and (ii) investors who only condition over subsets of firms' disclosures will observe market reactions which are not inline with their expectations.

The remainder of our paper proceeds as follows. Section 2 lays out the theoretical foundations of our measure. Section 3 discusses our sample, data, and model training. Section 4 presents empirical tests of our measure. Section 5 concludes.

2. Theoretical Motivation

Information content of corporate disclosures and the associated market reactions are among the most fundamental questions within the accounting and finance literature. This literature goes back to seminal studies by Ball and Brown (1968), Beaver (1968), and Fama

et al. (1969). The primary result from these studies is that security prices are quick to incorporate corporate news upon arrival and hence reflect publicly available information in a timely manner. For the most part, however, the studies examining information content over the past five decades focus on quantitative information, e.g., earnings or stock returns, and use the notion of *surprise*, i.e., the difference between the realization of a random variable and its expectation, to quantify the amount of new information. More formally, the surprise is defined as:

$$s_{t+1} = x_{t+1} - \mathbb{E}[\tilde{x}_{t+1}|\Omega_t]. \tag{1}$$

For example, in Ball and Brown (1968) $\tilde{x}_t$ is a firm's net income. More recent work has also used this approach in attempt to quantify narrative content, e.g., in Tetlock et al. (2008) $\tilde{x}_t$ is the fraction of negative words in firm-specific news stories, and in Loughran and McDonald (2011) it is the fraction of words in a firm's 10-K belonging to a given dictionary, e.g., litigious words.

In contrast to numeric data, measuring the amount of new information in textual communications presents a formidable challenge. First, narrative information is highly multidimensional and becomes the notion of a surprise, which is challenging to implement directly. Second, calculating expected values for textual data is not a well-defined problem.

One way the literature tacked these challenges is by assessing the similarity between documents (annual reports) from period to period, e.g., based on bag of words representation. This approach has been adopted by several recent studies (e.g., Brown and Tucker, 2011; Cohen et al., 2020). Similarly, Loughran and McDonald (2011) use the prior period's 10-K as a proxy for the expected sentiment (a unidimensional construct) in the current period's 10-K.

One problem with this approach is that it does not identify the new information per se but instead suggests that it appears present in the filing (relative to the prior period). Thus, it is not particularly helpful if one is interested in identifying and processing the new

information in its own right. Furthermore, the lack of similarity does not imply the presence of new information. Specifically, dropping irrelevant text from an older 10-K affects similarity in the same way as adding new information to the current 10-K. More importantly, the approach requires two consecutive 10-Ks and cannot be easily adapted to compare annual and quarterly reports due to their different nature and level of detail. It is plausible that disclosures, such as 8-Ks (press releases) and 10-Qs that precede the current 10-K, among others, had already communicated the information, which is repeated in the annual report. In this case, information is already reflected in prices and investors' expectations and hence does not meet the definition of new information.

This discussion suggests that measuring the degree of new information in corporate filings requires a fundamentally different approach. In this paper, we aim to address this challenge by combining the primitives from the information theory and recent developments in natural language modeling.

2.1. Measuring Information

In a foundational paper within the field of information theory, a branch of applied mathematics, Shannon (1948) examines a model of communication in which a sender transmits a message to a receiver. In this setting, the information content of the message is the component that was previously unknown to the receiver. For example, if the entire message was known to the receiver a priori, i.e., the message was expected to be received with probability one, then the information content is zero. Thus, the information content depends not only on the message but also on the receiver's expectations about the message–just as in Eq. 1 above. Shannon's key insight was that witnessing a realization of an unlikely event is more informative than witnessing an event that is likely, which leads to the well-known definition of information:

$$I(x) = -\log p(x) \tag{2}$$

where $p(\cdot)$ is the probability density function of x.

Shannon information, sometimes referred to as self-information, is closely linked to the notion of a surprise used in the prior literature. To illustrate this, consider a continuous random variable $\tilde{x}$ that follows a Gaussian distribution. Substituting the density of $\tilde{x}$ into Eq. 2 yields the following relation for the information contained in a given realization x:

$$I(x) \propto (x - \mathbb{E}(\tilde{x}))^2. \tag{3}$$

Hence, the main difference between Eqs. 1 and 2 is that the former signs the "surprise" in observing a realization x, whereas the latter does not. In other words, Eq. 2 does not make a statement about the sentiment of the surprise, only its magnitude, but otherwise retains its intuitive and desirable properties.[3]

Our measure of information in a given textual disclosure follows directly from Shannon's definition of information. Any textual disclosure $\tilde{\mathcal{D}}$ can be represented as a sequence of random variables (tokens), $\tilde{\mathcal{D}} = \{\tilde{\tau}_1, ..., \tilde{\tau}_n\}$, which follow a joint probability distribution $p_{\Omega_t}(\cdot)$. These could be words in a sentence or could also be sentences in a document. Then, the information in a realized disclosure $\mathcal{D}$ is given by:

$$\begin{aligned} I_t(\mathcal{D}) &= -\log p_{\Omega_t}(\tau_1, ..., \tau_n) \\ &= -\log \prod_{i=1}^{n} p_{\Omega_t}(\tau_i | \tau_{i-1}, ..., \tau_1), \\ &= -\sum_{i=1}^{n} \log p_{\Omega_t}(\tau_i | \tau_{i-1}, ..., \tau_1), \end{aligned} \tag{4}$$

where Ω_t is a set of publicly available information and where we used the chain rule of conditional probabilities. This formulation is particularly useful as it implies that measuring the amount of new information in a text boils down to summing the log probabilities of

[3]While this example assumes that $\tilde{x} \sim \mathcal{N}(\mu, \sigma^2)$, other tractable and intuitive forms of the information content can be derived for members of the exponential family of distributions—a broad class of distributions which includes the Gaussian.

each word (token) conditional on the preceding words. These conditional probabilities are precisely what large language models, such as GPT, are designed to model.[4]

Along these lines, we leverage the GPT architecture to empirically estimate $p_{\Omega_t}(\cdot)$ from a collection of disclosures available at time t. For historical context, it is important to note that the fields of information theory and computational linguistics have a long history of constructing such estimates, going back to Shannon (1948), who presents a series of approximations to the English language. While the literature examined various methods for modeling $p_{\Omega_t}(\cdot)$, it was not until the recent revolutionary advances in language modeling that researchers were able to generate believable and readable text (Radford et al., 2019). These developments are responsible for the recent success of "Generative AI" in a wide range of domains. For these reasons, we focus our attention on modeling $p_{\Omega_t}(\cdot)$ using the LLM architecture detailed in Section 3.4.

3. Data, Model Training, and Inference

3.1. Sample Construction

One challenge we face in our study is the computational cost of training large language models. Specifically, our study requires (i) pretraining an LLM from scratch on a cross-section of firms' narrative disclosures, (ii) further pretraining the LLM from step (i) on each individual firm's time-series of disclosures to yield a firm-specific model for each firm in the sample, and (iii) iteratively applying and further pretraining the firm-specific model from step (ii) out-of-sample to measure the information in new narrative disclosures. The

[4]The formulation above is causal in the sense that the probability of a given token depends only the context on the left side of the token. Alternatively, one can write the joint distribution based on the tokens to the right of a given token, i.e., $p(\mathcal{D}) = p(\tau_n)p(\tau_{n-1}|\tau_n) \times ... \times p(\tau_1|\tau_2, ..., \tau_n)$. Large language models (LLMs), such as BERT (Devlin et al., 2019), take advantage of this idea by using the context on both sides of a given token, which is particularly helpful if the goal is to encode information (e.g., encode a sentence in a foreign language before translating it). We do not follow this approach because we are interested in identifying new information conditional on prior information.

computation associated with executing this sequence of tasks is largely a function of two choices: the LLM size and the number of firms in the sample.

In selecting an appropriate LLM size, i.e., the number of parameters, one must consider the volume of training data, compute budget, and the capacity of the LLM to model the distributional properties of firms' disclosures (Hoffmann et al., 2023). We focus on the latter since the volume of training data and compute budget are largely fixed. Existing evidence suggests that model size has decreasing returns to scale, i.e., the benefits from using larger and larger models become relatively small. Radford et al. (2019) show that doubling the model size from 345 million parameters to 762 million reduces perplexity–a measure of how well a probability model fits a sample–by 30.3%, while further doubling the model size from 762 million parameters to 1.54 billion reduces perplexity by just 14.4%. Although larger models are capable of generating text that humans would attribute to a human author (e.g., Brown et al., 2020)), we are primarily concerned with modeling the distributional properties of firms' disclosures, of which perplexity is a direct measure. Ultimately, it is an empirical question how large of a model is necessary, thus, we train models following the Pythia scaling suite using the 410M parameter size model as our baseline architecture (Biderman et al., 2023).

The other factor influencing our computation costs is the number of firms in our sample. The larger the number of firms, the greater the number of firm-specific LLMs that need to be trained and the volume of out-of-sample disclosures that need to be processed. We moderate our computation costs by drawing a random sample of firms stratified by decile of market capitalization. Specifically, we consider all firms that filed either a 10-K or 10-Q on EDGAR during the year 2006 and had a valid link between Compustat and CRSP. Additionally, we apply common filters for financial statement variables (positive total assets, sales, shares outstanding, price, and non-missing net income), stock characteristics (CRSP share code of 10 or 11, listed on AMEX, NASDAQ, or NYSE), and exclude financial firms. Next, we collect firms' market value of equity as of the end of 2006 from Compustat and assign each

firm to a market value decile. Finally, we randomly sample 50 firms from each decile of market value for a total sample size of 500 firms. Table 2, presents the effects of these filters on the firms eligible to be included in our sample. Given this sample of firms, we then collect firm fundamentals from Compustat, equity market data from CRSP, and intraday data from NYSE TAQ. We also collect firms' disclosures from EDGAR, as detailed below. Throughout our analyses, we winsorize all continuous variables by year at the 1% level.

3.2. Extracting Narrative Disclosure

Developing our measure requires a set of narrative disclosures in plain text format, i.e., a "corpus." To meet this need, we leverage the BeanCounter corpus (Wang and Levy, 2024), which includes plain text versions of *all* filings accepted by the SEC EDGAR system from 1996 through 2023. Table 1 presents descriptive statistics on the portion of BeanCounter used in our work. While prior work has largely focused on annual and quarterly reports, other forms are also major contributors of narrative disclosure such as current reports (Forms 8-K), proxy statements (Forms DEF 14A), and registrations of material M&A information (Forms S-4).

Each filing, e.g., a current report filed on Form 8-K, consists of the main filing and may include attachments. The main filing generally follows a standardized format based upon templates provided by the SEC.[5] Attachments can follow a variety of layouts and file types, e.g., HTML, plain text, images, PowerPoint files, etc. BeanCounter provides the main filing and each attachment as a separate document thereby enabling us to pinpoint the location of informative content.

[5]For example, the Form 10-K instructions and template are available here: https://www.sec.gov/files/form10-k.pdf

3.3. Tokenization

As discussed in Section 2, large language models require converting text into tokens, which are subsequently encoded as vectors. One could define tokens based upon entire words, e.g., "the" and "context" would each be a unique token. This would be akin to memorizing a dictionary and simply looking up the meaning of a given token in the dictionary. One challenge with this approach is that it leads to a very large vocabulary, perhaps unnecessarily large since many words are constructed by combining sub-pieces together where each sub-piece may share meaning across the vocabulary, e.g., the prefix "un" in words such as unpresumptuous and unselfish.[6]

Following prior literature, we use Byte-Pair Encoding (BPE) (Sennrich et al., 2016), which leverages commonality across words to build a vocabulary of a more manageable size. This approach involves finding a smaller number of unique sub-words in the corpus, along with the individual characters (or more generally "symbols,") that can be used to form the words in the target vocabulary. The algorithm begins by merging the pairs of adjacent symbols which are most commonly found next to one another and repeats this process until the vocabulary reaches a target size.

We train a tokenizer from scratch using *only* the "in-sample" disclosures within our corpus, i.e., those filings from 1996 through 2006. The inputs to the training process are largely the same as those in Radford et al. (2019) with the exception that we explicitly include tokens for the Arabic numerals $0, 1, ..., 9$. This results in a tokenizer with a vocabulary size of 50,270: 50,257 tokens learned via the training process, 10 tokens Arabic numerals, and three special tokens for beginning of sentence, end of sentence, and unknown (word that is not in the

[6] At the opposite end of the spectrum, one can define tokens based on the 26 characters in the English alphabet, plus the 10 Arabic numerals, some punctuation characters, say . and !, and the white-space character " ". This would lead to a set of 39 unique tokens. This set of tokens would be referred to as the "vocabulary" of our model. With just these 39 unique tokens much of English text could be modeled, doing so however would require learning context specific representations for every single token, i.e., the meaning of "t" when used in the sequence " the " versus " context."

vocabulary).

3.4. Model Training

In this section, we detail our approach to empirically estimating investors' prior beliefs about a disclosure conditional on the information available to investors at time t, i.e., $p_{\Omega_t}(\mathcal{D})$. Recall that the definition of our measure of information (Eq. 4) is:

$$I_t(\mathcal{D}) = -\log \prod_{i=1}^{n} p_{\Omega_t}(\tau_i|\tau_{i-1}, ..., \tau_1),$$

Let a collection of disclosures available at time t be indexed by $J = 1, 2, ..., m$. Then, the information set at time t can be approximated as $\widehat{\Omega}_t = \{\mathcal{D}_j | j \in J\}$, where each disclosure $\mathcal{D}_j$ is a collection of tokens $\tau_{j,1}, ..., \tau_{j,n_j}$. Let $\widehat{p}_{\Omega_t}(\cdot;\theta)$ be a neural network parameterized by θ. We obtain our empirical estimate of investors' prior beliefs by choosing θ such that

$$\theta := \arg\min_{\theta} \sum_{j=1}^{m} \sum_{i=1}^{n_j} -\log \widehat{p}_{\Omega_t}(\tau_{j,i}|\tau_{j,i-1}, ..., \tau_{j,i-k}; \theta) \tag{5}$$

where $\widehat{p}_{\Omega_t}(\cdot;\theta)$ is a GPT-style model and k is the size of the context window which we set to 2,048. This objective function is the standard causal language modeling objective, i.e., the context window consists only of tokens which precede τ_i, and is equivalent to maximizing the log-likelihood.

While we use the same *architecture* as Radford et al. (2019), **we train a new model from scratch** such that the model has *only* been conditioned on the historical information contained in Ω_t. This differs sharply from other studies which fine-tune pretrained models, e.g., FinBERT (Araci, 2019), and addresses concerns of look ahead bias in LLMs (see, e.g., Sarkar and Vafa, 2024). Such distinction is critical when detecting new information as one must ensure that the model does not condition its predictions on the information revealed in the future.

3.4.1. Cross-sectional Model

We begin by pre-training a single "cross-sectional" LLM on the narrative content extracted from the disclosures of all firms in our sample filed on EDGAR from 1996 through the end of 2006. The training data from this time period totals 35.5B tokens, roughly 5 times what was used to train GPT-2 (Radford et al., 2019) (see Table 1 Panel A for more descriptives). Given our baseline model size of 410M parameters, a single epoch through this data corresponds to roughly 80 tokens per parameter which is towards the "inference optimal" end of the spectrum and sufficient for training a model of this scale Touvron et al. (2023). Our cross-sectional model results from minimizing Eq. 5 over these 35.5B tokens using SGD with a batch size of 2M tokens and a learning rate of 1e-3 ramped up over the first 100 batches and decayed following a cosine schedule to 1e-5. All other hyperparameters follow those detailed in Biderman et al. (2023).

The primary benefit of using a decoder architecture is the ability to generate text by sampling from the prior modeled by the LLM. In Appendix B, we exploit this feature of our LLM to explore the extent to which the training procedure has resulted in an LLM that reasonably models the distributional properties of narrative content. Specifically, we prompt the model with text that is commonly found at the start of the business description in Form 10-K. Table B1 presents this prompt.[7] Table B2 presents four examples of text generated by our cross-sectional LLM. We note that the model is capable of generating reasonably coherent text spanning various industries and interleaving numbers and dates as a reader might expect. We view this as additional evidence of a successful training process.

[7]Formatting such as white space and indentation are preserved in our training data. Thus, when prompting the model we intentionally use similar formatting.

3.4.2. Firm and Time-Specific Models

Our ultimate goal is to construct a firm and time-specific estimate of investors' prior beliefs which we can use to compute our measure of information. We obtain such an estimate by continued pretraining of the cross-sectional model on a given firm's time series of narrative disclosures. In words, we initialize a new model using the weights from the cross-sectional model detailed in 3.4.1 and continue the pretraining process using a single firm's disclosures prior to 2007. Beginning in 2007, we iteratively process new disclosures in time order by first measuring the information in the new disclosure and then continue pretraining the firm-specific model on the new disclosure. This process is carried out iteratively through 2023. At the end of this process, we have a time series of our information measure from 2007 through 2023 where the prior belief used to measure information was conditional on all of the firm's disclosures immediately prior to each disclosure in the time series.

Table 1 Panels B and C presents descriptive statistics on the sample of disclosures used for this part of our model training. These data include roughly 2.5B tokens from disclosures produced by firms in our random sample. In contrast to Panel A, where content from 10-K/Qs and 8-Ks contributed roughly 30% of the data, these filings are now responsible for more than 50% of the content and the major form types are typical of operating companies, e.g., Form 485BPOS is typically filed by investment management companies and no longer contributes a substantial amount of content to the data. This difference in the composition of the training data is one justification for our approach of developing firm-specific models.

Gupta et al. (2023) develop guidelines for continued pretraining from empirical experiments and show that under certain conditions further pretraining will cause the model to "forget" older training data and overfit newer data. We balance updating of the models to reflect new content against retaining older content by measuring perplexity during continued pretraining on a holdout set of data. Since perplexity is a measure of how well a probability model fits a set of data, this enables us to monitor overfitting on the new data. We cease

pretraining when either of two stopping criteria are met: (i) the average perplexity on the holdout data over the last 10 batches drops below the final perplexity of the cross-sectional model, or (ii) 10 epochs through the firm-specific data. We find that for the majority of new disclosures, criteria (i) halts the continued pretraining. Following Gupta et al. (2023), we continue pretraining at a learning rate of 50% of the maximum from the pretraining phase (1e-3) without any warmup and decay the learning rate to 10% of the maximum.

As we did with our cross-sectional LLM, we prompt these firm-specific LLMs and generate text by sampling from them. Table B3 presents examples for Apple, Cummins, Pfizer, and Weyco Group (a microcap). The diversity of the generated text is striking, with the firm-specific LLMs referencing relevant products, previous disclosures, terms of agreements with suppliers, and financial performance. We manually inspect these samples relative to the firms' actual disclosures. The text generated for Cummins refers to a particular footnote of a previous disclosure. Examining Cummins's time-series of disclosure we find that this is an accurate reference. For Pfizer, we note that the text references two drugs called Alond and Exubera. The former does not appear to exist howver the latter was indeed a Pfizer drug which was withdrawn from the market due to low sales. Collectively, we view the richness of these prompts as further evidence that our training and fine-tuning processes result in LLMs that are reasonable models of the distributional properties of narrative disclosure.

4. Empirical Tests

From Table 1 Panel C, we can see that Forms 8-K, 10-Q, and 10-K contribute the most narrative content to our sample (71%). The next two largest contributors are responsible for 9% and 2% suggesting a long tail of forms comprising the remaining 29% of content. For this reason, our empirical tests largely focus on Forms 8-K and 10-K/Q.

4.1. Measuring Information

Given a disclosure $\mathcal{D}$ produced by a firm f at time t, we apply the firm and time-specific prior to quantify the information of the i-th token:

$$I(\tau_i) = -\log \widehat{p}_{\Omega_{f,t}}(\tau_i|\tau_{i-1}, ..., \tau_{i-k}; \theta_{f,t}). \tag{6}$$

This approach allows us to identify the location of comparatively more informative tokens with a high degree of precision. For example, consider the announcement on January 17, 2007, disclosed in Apple's 8-K filing, where the company first mentions the introduction of its iPhone. As illustrated in Figure 2, the sentences shaded in green highlight the degree of "news" relative to the estimated Apple-specific prior. While the disclosure contains numerous pieces of information, such as record revenue, the most informative sentence–according to the estimated prior–is the sentence in which Apple mentions launching the iPhone and Apple TV.

Table 3 presents the distribution of our information measure separately for the set of main filings and exhibits in our sample. For a given main filing, we compute the filing's information as the mean of the token-by-token measurements from Eq. 6. For a set of exhibits attached to a given filing, we compute mean of the token-by-token measurements across all exhibits, i.e., the weight given to each token in an exhibit is equal regardless of the length of the exhibit in which it is located. We then compute distributional statistics across these main filing and exhibits-level measurements. We find that, on average, the most informative narrative content is in the exhibits attached to current reports.

4.2. Location of Information

As a next step, we provide descriptive evidence on the location of information in filings. Specifically, we label a token τ_i as "high information" if $I(\tau_i)$ is in the top quartile of $I(\cdot)$ across all filings, i.e., for each token we create an indicator variable $HighInfo$ equal to one

if the information of that token is in the top quartile of information and zero otherwise. We then take the mean of $HighInfo$ over a various splits of the data. If content deemed high information was uniformly distributed throughout disclosures, then we should see that these means are generally statistically indistinguishable from 25%.

Both within and across filing types we find significant heterogeneity. For example, Figure 3 Panel A presents results for 10-K/Q filings and provides two key insights. First, exhibits tend to contain roughly 150% *more* content labeled as high information than the main filing. Given that prior literature, e.g., Dyer et al. (2017), has focused on the main filing portion of 10-K/Qs this is perhaps one reason why these reports appear to have become less informative over time. Second, within the main filing, the business description is the most likely to contain content labeled as high information (22%) followed by legal proceedings (14%) and the management's discussion and analysis (13%). These results are intuitive since the SEC requires firms to update the business description section with any developments since the beginning of their fiscal year, and are consistent with prior literature highlighting that many pieces of information required to create better estimates of fundamental performance are dispersed throughout filings (see, e.g., Rouen et al., 2021).

Panel B presents results for 8-K filings and highlights that 8-Ks, on average, contain more content labeled as high information–consistent with 8-Ks providing timely information regarding recent material events. Additionally, while much research has been devoted to studying earnings announcements, i.e., Form 8-K Item 2.02, filings containing this item tend to contain the smallest volume of informative narrative content relative to other 8-K items (roughly 27%). We find that other items such as changes in accountants, bankruptcy, and warnings that stakeholders should not rely on previously issued financial statements, tend to contain the highest volumes of content labeled as high information (60% or more). Notably, the labeling of content as high information was based on the distribution of our measure across *all* filings. Thus, these results show that the majority of content labeled as high information arrives via Form 8-K. While we are unaware of prior work documenting this

empirical pattern, the result is intuitive given that Form 8-K is meant to provide stakeholders with timely information regarding firms' operational status *as material events occur* whereas Forms 10-K/Q are meant to provide updates at regular, but likely delayed, intervals.

4.3. Timing of Information Release

A large literature has examined the movement of stock prices and the underlying drivers of stock volatility (perhaps most famously, Shiller, 1981). Along these lines, our previous results document that the majority of the content we label as high information arrives via Form 8-K. This begs the question of the timing of this information and its contribution to daily stock volatility. We explore this question by (i) measuring the daily arrival of content labeled as high information vis a vis the approach mentioned in Section 4.2 and (ii) attributing daily stock volatility to disclosures based on the timing of their arrival. For (ii), we measure volatility using absolute abnormal returns which is then allocated to disclosures based on non-overlapping [-1,1] event time windows around their release. For both (i) and (ii), we measure the total volume of high information content released and volatility realized over the reporting year. We then scale the cumulative amount of information and volatility by these totals, respectively. As a result, if all information or volatility was accounted for, then at the end of the reporting year the cumulative quantities would equal 100%.[8]

Figure 4 presents the cumulative information flow over the reporting year measured using the two approaches outlined above. For quarterly and annual reports, the arrival of information and volatility mechanically jumps at the end of each reporting quarter since observations are aligned by reporting quarter. Earnings announcements, which often precede the release of more complete financial information presented in Forms 10-K/Q, begin to arrive just before the end of the reporting quarter but a largely confined to the last 20% of

[8]One drawback of this approach is that noise in the abnormal returns will not "cancel out." For this reason, we choose to measure the volatility attributable to a particular disclosure using the tight [-1,1] window around its release.

the quarter. In contrast, the information contained in Forms 8-K and other filings arrives nearly continuously over the reporting quarter and is responsible for a significant portion of both the content we label as high information as well as stock volatility. These results speak to the importance of these comparatively less studied forms and the challenges faced by attention constrained investors who must continuously monitor these channels of unscheduled disclosures.

4.4. Market Reaction

In this set of tests, we use a short-window daily event study to examine whether the quantity of narrative information that we measure can explain the market reaction to disclosures. Specifically, if our measure indeed captures new information, then we expect to see a more pronounced market reaction around the disclosure date when our measure suggests a higher information content.

Figure 5 plots the average daily values of *AbsoluteReturn* and *Volume* in the [-10, 10] day window around the filing date separately for filings that are in the top and bottom quartiles of *Info*. Results are presented separately for annual and quarterly reports, the management's discussion and analysis (MD&A), and current reports (8-Ks). Across all panels, there is a distinct increase in the absolute return and volume on the filing date. More importantly, the increase appears markedly greater for the top quartile of *Info* sub-sample.

We also expect that the market reaction on the disclosure day will increase with our measure. To investigate this, Figure 6 plots the average value of *AbsoluteReturn* and *Volume* for $t = 0$ as a function of *Info*. We find that both measures of market reaction tend to be increasing functions of *Info*. We note that this non-monotonicity largely disappears for *Volume*.

4.4.1. Baseline Daily Event Study of Market Reaction

To test whether there is a statistically significant relation between *Info* and the market reaction, we estimate the following regression pooling across all firm-dates in the [-10, 10] day window around the filing date:

$$y_{f,t} = \beta_1 Event[0] + \beta_2 Event[0] \times Info_{f,j} + \boldsymbol{\gamma}\mathbf{X}_{f,t} + \delta_t + \lambda_{f,t} + \epsilon_{f,t} \tag{7}$$

where $y_{f,t}$ is either *AbsoluteReturn* or *Volume*, for firm f on date t. *Event*[0] is an indicator variable equal to one on the first date the market could react to the filing and zero otherwise.[9] $Info_{f,t}$ is our measure of the information contained in the j-th disclosure filed by firm f. We also present a benchmark model where we proxy for information content with document *Length* (Bonsall et al., 2017), which is the count of characters in the narrative content in the filing deciled and scaled to range from zero to one. $\mathbf{X}_{f,t}$ is a vector of control variables for firm f on date t (Defined in Table 4 and enumerated in Table 5). δ_t is a date fixed effect. $\lambda_{f,t}$ is a firm fixed effect for our 10-K/Q sample and a firm-year-quarter fixed effect for our 8-K sample. β_1 and β_2 are our coefficients of interest. β_1 represents the average market reaction to filings in the lowest decile of *Info* while β_2 represents the incremental market reaction associated with moving from the lowest to highest decile of *Info*. Throughout our analyses, we estimate regressions using OLS and report two-way standard errors clustered by firm and date, thus allowing for arbitrary correlation across time within a given firm and across firms within a given date (Cameron et al., 2006).

Table 4 presents descriptive statistics for the variables used in our daily event study partitioned by inclusion in our random sample. Panel A presents statistics for the year of our sampling and Panel B for the years following our sampling. Across partitions and panels

[9]This is the filing date for filings accepted either prior to markets opening or intraday and the date of the following trading session for filings accepted after markets close.

we find that few firm characteristics are statistically distinct between our sample and all firms. This suggests that our sample is not systematically different along these characteristics at the time of sampling nor in the years that followed sampling.

Table 5 presents results from estimating Eq. 7. Panel A (Panel B) presents results for our sample annual and quarterly (current) reports. Across all measures and form types, we find evidence of statistically and economically significant market reactions on the disclosure event day. For example, within our sample of annual and quarterly reports, estimates of β_1 for *AbsoluteReturn* and *Volume* suggest that moving from the lowest to highest decile of *Info* is associated with 41.5% and 98.7% increases in the market reaction, respectively. These increases are highly significant and in the case of *Volume*, completely subsume the event-day indicator. We also observe that the amount of variation in the market reaction explained by the regression models increases after the inclusion of our measure–absolute increases in R^2 range from 0.1% to 1.2%.

One can interpret the magnitude of β_2 relative to β_1 as the proportion of the market reaction explained by the information in narrative content. For example, if our measure was perfect and the information in narrative content was responsible for 90% of the market reaction to the announcement, then we would expect β_2 to be roughly nine times larger than β_1. For an imperfect measure of textual information content, however, β_1 is expected to absorb a portion of the market reaction attributable to narrative content. In this sense, the magnitude of β_2 relative to β_1 puts a lower bound on the portion of the market reaction attributable to information in narrative content. Across all but one specification, we find that β_2 is at leats 4.86 times larger than β_1. These results are consistent with the notion that the market reaction is *primarily* explained by the information provided in narrative content.

4.5. Informative vs. Uninformative Sentiment

In our second set of tests, we examine whether the market reaction to the sentiment of the filing depends on the disclosed textual information content. A large literature in finance and

accounting studies how markets respond to disclosures sentiment (see, e.g., Loughran and McDonald, 2016, for review). This literature generally finds that greater negative sentiment explains stock price movements. However, the sentiment of disclosure will only explain returns to the extent it is associated with the revelation of information. For example, the negative sentiment of MD&A is unlikely to trigger market reactions if the same content was previously communicated in prior press releases–notably, a market reaction to such content would be a market inefficiency unless the reprinting of the content is in and of itself informative. We thus examine the extent to which there is an interaction between sentiment and information when explaining contemporaneous stock returns.

We measure sentiment using the negative and positive dictionaries of Loughran and McDonald (2011). For each word in a disclosure, we sign the sentiment as negative one, one, or zero based on whether the word occurs in the negative, positive, or neither dictionary, respectively. Following the terminology of Loughran and McDonald (2011), we refer to this signed sentiment as *FinNeg*. We then interact *FinNeg* ***on a word-by-word basis*** with the information measured in Eq. 6. Given this word-by-word interaction between sentiment and information, we follow the same general approach used to calculate *Info*, i.e., we take the mean over each disclosure, sort the measure into deciles, and scale it to range from minus one to one. We refer to the resulting information-weighted sentiment measure as *IWS*.

To the extent our measure is effective at capturing the information content of narrative disclosures, we expect to observe two empirical phenomena. First, the *AbnormalReturn* associated with the filing should be a monotonically increasing function of the information-weighted sentiment, *IWS*. Second, we expect that the slope of the relation between *AbnormalReturn* and *IWS* is greater for filings that contain more information according to our measure. Figure 7 plots the average value of *AbnormalReturn* for $t = 0$ as a function of *IWS*. Across all form types, we find that *AbnormalReturn* is generally an increasing function of *IWS* and that the slope of this relation is steeper for more informative filings.

To test for a statistically significant relation between *AbnormalReturn* and *IWS*, we estimate the following regression pooling across all firm-dates in the [-10, 10] day window around the filing date:

$$AbrnomalReturn_{f,t} = \beta_1 Event[0] + \beta_2 Event[0] \times IWS_{f,t} + \boldsymbol{\gamma}\mathbf{X}_{f,t} + \delta_t + \lambda_{f,t} + \epsilon_{f,t} \quad (8)$$

where *AbnormalReturn* as the daily residual from the expected returns model of Fama and French (2015) and Carhart (1997) estimated over the [-70,-11] day window relative to the event, expressed as a percent. In addition to *IWS*, we also estimate Eq. 8 using equal-weighted sentiment deciled and scaled to range from negative one to one, *Sentiment*. As before, β_1 and β_2 are our coefficients of interest. However, β_1 now represents the average abnormal return to filings in the lowest decile of *IWS* while β_2 represents the incremental abnormal return associated with moving from the lowest to highest decile of *IWS*.

Table 6 Panel A (Panel B) presents results for annual and quarterly (current) reports. Our baseline results in columns (1) and (2) across panels are generally consistent with prior literature–Forms 10-K/Q in the most negative decile of sentiment are associated with negative abnormal returns of around 22 bps. After information weighting, in column (3) we find that the difference in abnormal market reaction from the most negative to most positive deciles of sentiment grows to 4.84% and 3.87% for filings on Form 10-K/Q and 8-K, respectively. Finally, in column (4), we include both *Sentiment* and *IWS* in the regression. We find that our inferences in terms of statistical and economic significance are generally unchanged.

Collectively, the evidence in Table 6 highlights the interaction between ***information*** and sentiment. Examining solely 10-K filings, Loughran and McDonald (2011) find a roughly 25 bps spread between the extreme quintiles of negative sentiment (p. 51), similar to our findings. However, when we account for the interaction between information and sentiment, i.e., consider only sentiment over the informative portion of the document, the effect is more

than an order of magnitude greater, i.e., roughly 484 basis points. This is consistent with the notion that (i) our measure is capturing the *information* in narrative content rather than some other latent time-varying firm characteristic, and (ii) the sentiment of non-informative content should not influence prices

4.6. Limited Attention and Alternative Information Sets

Our analysis thus far has been designed to represent an investor who updates their beliefs following each and every disclosure produced by a firm. However, a large literature highlights that the ability of investors to process information is constrained. Thus, our analyses are unlikely to be reflective of any single individual. In our next set of tests we explore three other representative investors who only update their beliefs over specific subsets of disclosures: annual reports, annual and quarterly reports, and current reports. Each of these representative investors is meant to capture the perceptions of an investor who can only process a subset of the information produced by firms, e.g., just annual reports.

We generate content subset-based priors following the procedure outline in Section 3.4.2 with the exception that we only update the model when a given disclosure is part of the relevant subset, e.g., an annual report. We continue to measure information as described in Section 4.1 for each disclosure released by a firm. For example, consider an investor who only "reads" annual reports. When an 8-K is released, we measure the information in the 8-K relative to a model conditioned over all prior annual reports. If another 8-K is subsequently released, we again measure information relative to all prior annual reports *without* having updated the model based on the prior 8-K. As a result, these investors may view pieces of content previously disclosed in filings outside of their subset as "news" even though such content is not news relative to an investor who conditioned over *all* disclosures.

On the one hand, if it is the case that conditioning beliefs over a particular subset of disclosures is sufficient for identifying "news" then we expect to find results similar to those in Tables 5 and 6 even when beliefs are formed over solely that subset. On the other hand, if

investors must process all disclosures to get an accurate picture of the "news" in disclosures, then measures of information computed using subsets of disclosure will not be associated with the market reaction.

Table 7 presents results from re-estimating Eqs. 7 and 8 using *Info* as measured by models trained on various subsets of the content produced by firms. For the measures relying only on annual and/or quarterly reports, we generally find evidence that the content labeled as high information is associated with an "under-reaction" from the perspective of the representative investor–in the case of *AbsoluteReturn* there is a statistically and economically significant decrease in the absolute return and in the case of *Volume* the relation is not significant. In contrast, when information is measured using models conditioned over current reports alone, we find results which are similar to those based on measures conditioned over all disclosures. In the case of *AbnormalReturn*, the coefficients on *IWS* and R^2 values are somewhat smaller than our earlier tests. These results are consistent with the notion that Form 8-K is a primary source of information and one of which attention constrained investors should be mindful.

4.7. Decision-relevant Information

Thus far, our measure of information has not relied on returns in anyway. One benefit of this approach is that the measure can be computed for any piece of content so long as a prior over content is available. However, it is likely that some of the content produced by firms may be irrelevant for particular decisions. Along these lines, our final set of tests explores identifying content which is not only informative but informative about future returns. We follow the prior literature by predicting binary market sentiment subsequent a disclosure (see, e.g., Chen et al., 2022).

We use the sentiment dictionaries of Loughran and McDonald (2011) as a baseline, converting our measure *Sentiment* into a binary prediction based on a threshold of zero. We then compare four alternative approaches to this baseline. First, we take our information weighted

sentiment measure *IWS* and convert it into a binary prediction based on a threshold of zero. For the remaining three approaches, we generate representations, i.e., embeddings, of each disclosure using Llama-3.1-8B. This approach is similar to that used in recent studies such as Chen et al. (2022), however, in our setting the pieces of content are substantially longer. Specifically, Chen et al. (2022) note that 60% of the articles they consider are shorter than 512 tokens–the maximum context length of one of the LLMs used in their study–whereas in our setting less than 16% of disclosures are this short and the average length is 14,409 tokens. Thus, the three remaining approaches explore ways to create representations that are maximally predictive of future returns.

When faced with documents longer than the context length of an LLM, it is common to divide the document into smaller chunks which are passed through the model individually and then aggregated into a single document-level representation. One of the most common approaches for combining the chunk-level representations is known as "mean-pooling," i.e., taking the average along each embedding dimension across chunks. Figure 8 presents this approach graphically: we split lengthy documents according to punctuation into chunks of no more than 1,024 tokens, pass them through an embedding model, and mean pool across token-level representations to yield chunk-level representations.

Our first method for creating disclosure-level representations from the chunk-level representations is mean-pooling, denoted as $Llama_{MP}$. While mean-pooling applies equal weight to each document chunk, an alternative in the spirit of our measure of information is to weight each chunk by the information contained in the chunk. Along these lines, we take our token-by-token measure of information and aggregate it to the chunk level. We then normalize by the total information in the disclosure and use the resulting weights to compute an disclosure-level representation. This approach is denoted as $Llama_{IP}$. We note that these representations are actually a combination of two models: the embedding from a static Llama-3.1-8B model and the information measures from our firm and time-specific LLMs.

Our final approach leverages an attention mechanism to *learn* how to aggregate the chunk-level representations into a document representation which is maximally predictive of future returns. This approach has at least two desirable characteristics. First, the attention mechanism is able to learn how much weight to put on each chunk-level representation depending not only on the representation itself but also all other representations in the sequence. Since we train this model with a maximum sequence length of 1,024 representations, it is capable of identifying predictive patterns in content separated by as many as 1 million tokens. Second, we are able to train the model with an objective function that maximizes the mutual information between the aggregated representation and future returns.

The specific architecture we use, presented in Figure 9, is similar to Touvron et al. (2021). We replace the convolution stem and trunk with the Llama-3.1-8B model such that the input to the attention-based pooling layer is a sequence of chunk-level representations generated as shown in Figure 8. Additionally, we use the same basic architecture as the Llama-3.1 model for the attention pooling layer modifying it to allow for three types of additional information: (i) temporal position, (ii) document-level position, and (iii) document type. The first two types of information are incorporated via rotary position embeddings (Su et al., 2024) while the document type is incorporated via a learned embedding. During training we freeze the weights of the Llama model and train the attention-pooling layer using a CLS token for each return horizon. We train this model by minimizing the binary cross-entropy between the model predictions and the true realized returns.[10] Given a trained model, we then use the model to create document level representations. We denote this approach $Llama_{LA}$ for "learned aggregation."

To minimize the chances that our predictions are biased favorably by lookahead bias, we employ an iterative training process. Specifically, we train a learned aggregation (LA) model

[10]See Boudiaf et al. (2021) for a proof that minimizing the cross-entropy is equivalent to maximizing the mutual information.

using disclosures up until time t and returns up until $t + 12$ months. We then apply the LA model to new disclosures released after time $t + 12$ months and up until time $t + 15$ months, i.e., we apply the model to those disclosures released during the quarter after the latest date used for any of the model's training data. This yields a sequence of representations for the three representation based approaches: $Llama_{MP}$, $Llama_{IP}$, and $Llama_{LA}$. Following a similarly staggered training procedure, we then train linear classifiers to predict binary market sentiment using the representations as features. The training data for this step is a rolling five year window. Using the average prediction over the training data as our decision threshold we then make out-of-sample predictions for disclosures released over the subsequent quarter. We repeat this process iterative to generate predictions for the disclosures in our sample.

Table 8 presents results for the five methods of predicting binary sentiment. Columns (1) and (2) present our baseline (Loughran and McDonald, 2011). For filings on Form 10-K/Q, we find that this approach provides a meaningful lift of 1-2% over randomly guessing the market sentiment. This lift exists for horizons from one day up to 12 months. Results for Form 8-K are similar except that the return predictability largely disappears after six months although it is still directionally correct. Columns (3 and (4) present results after information weighting the sentiment. Directionally, this method is generally better than equal-weighted sentiment although the differences are not always statistically distinct from the baseline. The largest lift comes from predicting long-horizon returns to 10-K/Q filings and short horizon 8-K filings.

Considering the representation-based approaches, columns (5) and (6) show that simply mean pooling embeddings from lengthy documents is generally no better and sometimes *worse* than using the baseline of equal-weighted sentiment. Information pooling in columns (7) and (8) is generally similar or slightly better than information weighting sentiment. Lastly, the learned aggregation approach is presented in columns (9) and (10). Perhaps unsurprisingly, we find that this is the dominant approach, providing lift over the baseline

of 2-3% depending on the horizon and form type. These results provide several key insights. First, despite being straight-forward, the sentiment measure of Loughran and McDonald (2011) appears to work well both within the sample of filings it was designed for–Forms 10-K–as well as outside. Second, using a measure of information to weight pieces of content–whether those are tokens or chunk-level representations–appears to improve predictions. Lastly, when a particular statistic of interest is readily available–returns in this setting–then leveraging a model that can learn predictive patterns in lengthy disclosures also appears to improve predictions.

As a final vignette into the which content is identified as value-relevant by the LA model, we compare the overlap between content labeled as high information using our measure and the content most heavily attended to in the LA models. Specifically, we collection the attention weights placed on each chunk-level representation by each LA model when generating the disclosure-level representations. Since a given chunk of content can be attended to across multiple predictions, we average these weights to get a single attention weight for each chunk. We then compute how much of the content in the top quartile of *Info* is also in the top quartile of attention, i.e., highly value-relevant. Figure 10 presents this overlap by topic following Dyer et al. (2017). Across topics, we find that content labeled as high information tends to also be highly value-relevant. Specifically, there is generally at least an 80% or higher overlap between these two sets. This is further evidence that our measure is capturing not only information but value-relevant information–at least when applied to content on EDGAR.

5. Conclusion

Although narrative disclosure is a frequent form of communication by firms to outside stakeholders, most studies examining the reaction to new information have focused on quantitative information such as earnings. In this paper, we draw on the foundations of information

theory and the emergence of large language models to measure investors' prior beliefs about disclosure. We then use these priors to uncover new information contained in textual data and study the value of narrative content in capital markets.

We begin by training an LLM on all EDGAR disclosure prior to 2007, we then train firm and time-specific LLMs. Using these models as surrogates for investors' beliefs about future disclosures, we compute our measure on a granular basis and examine where new narrative information is comparatively more likely to occur. Across form types, we find that current reports–rather than annual or quarterly reports–are most likely to contain content that is high information. We also find that exhibits attached to Forms 10-K/Q and 8-K are more likely to contain high information content than the main portion of these filings.

To better understand *when* information is released, we create views of our measure and stock volatility within reporting years. We find that whether information is measured using our method or returns, a significant portion of the information released by firms arrives via Form 8-K and other filings. This evidence speaks to the value of studying other channels of disclosure beyond annual and quarterly reports, the challenges faced by attention constrained investors who must monitor for these sporadic information releases, and the sources of daily stock volatility.

Having used each firm and time-specific LLM to compute a measure of information in each disclosure, we assess the extent to which our measure explains the market reaction to new disclosures. Consistent with our measure capturing decision-relevant information, we find that our measure explains the majority of the market reaction. In our second set of tests, we consider the interaction between information and sentiment. Noting that sentiment should only explain market reactions if it updates investors' prior beliefs, when we weight the sentiment of each word by its information, we find that the amount of the abnormal return explained by our measure is an order of magnitude larger than measuring sentiment alone. Our results highlight the importance of measuring both information and sentiment

when processing a new disclosure.

In our third set of tests, we explore the implications of limited attention on what is considered "news." Specifically, we retrain our models restricting them to updating only over particular types of forms. We find that relying solely on annual reports or annual and quarterly reports is associated with *perceived* under-reaction to what is "news" to these investors. In contrast, investors relying solely on current reports generally see market reactions which directionally match their beliefs but are somewhat muted relative to measuring "news" using all disclosures. This evidence is consistent with our earlier findings that current reports are an important disclosure channel which requires near constant monitoring.

In our final set of tests we aim to identify content useful for predicting future firm performance. Across a variety of methods we find that weighting content by our information measure leads to more accurate predictions. Further, when we allow models to learn which pieces of content are most useful for predictions, we find that a roughly 80% overlap between this content and what our measure labels as high information. This is consistent with (i) our measure capturing information, and (ii) the content on EDGAR being largely relevant to future performance.

Collectively, our evidence provides at least three key insights. First, LLMs appear to be useful tools for modeling investors' priors over narrative content. Second, other information events aside from earnings announcements, annual reports, and quarterly reports–which have been studied in great depth–are a significant, if not more significant, source of information about firm news. Third, information contained in these other forms tends to arrive sporadically, rather than at regular intervals, posing challenges for attention constrained investors and contributing to daily stock volatility.

References

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models.

Ball, R. and Brown, P. (1968). An empirical evaluation of accounting income numbers. *Journal of Accounting Research*, 6(2):159–178.

Beaver, W. H. (1968). The information content of annual earnings announcements. *Journal of accounting research*, pages 67–92.

Biderman, S., Schoelkopf, H., Anthony, Q., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., Skowron, A., Sutawika, L., and Van Der Wal, O. (2023). Pythia: a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Bonsall, S. B., Leone, A. J., Miller, B. P., and Rennekamp, K. (2017). A plain english measure of financial reporting readability. *Journal of Accounting and Economics*, 63(2):329–357.

Boudiaf, M., Rony, J., Ziko, I. M., Granger, E., Pedersoli, M., Piantanida, P., and Ayed, I. B. (2021). A unifying mutual information view of metric learning: cross-entropy vs. pairwise losses.

Brown, S. V. and Tucker, J. W. (2011). Large-sample evidence on firms' year-over-year MD&A modifications. *Journal of Accounting Research*, 49(2):309–346.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.

Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2006). Robust inference with multi-way clustering. Working Paper 327, National Bureau of Economic Research.

Carhart, M. M. (1997). On persistence in mutual fund performance. *Journal of Finance*,

LII:57–82.

Chen, Y., Kelly, B. T., and Xiu, D. (2022). Expected returns and large language models. *Social Science Research Network*.

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

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Dyer, T., Lang, M., and Stice-Lawrence, L. (2017). The evolution of 10-k textual disclosure: Evidence from latent dirichlet allocation. *Journal of Accounting and Economics*, 64(2):221–245.

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

Fama, E. F., Fisher, L., Jensen, M. C., and Roll, R. (1969). The adjustment of stock prices to new information. *International economic review*, 10(1):1–21.

Fama, E. F. and French, K. R. (2015). A five-factor asset pricing model. *Journal of Financial Economics*, 116:1–22.

Gupta, K., Thérien, B., Ibrahim, A., Richter, M. L., Anthony, Q. G., Belilovsky, E., Rish, I., and Lesort, T. (2023). Continual pre-training of large language models: How to re-warm your model? In *Workshop on Efficient Systems for Foundation Models @ ICML2023*.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., De, D., Casas, L., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. V. D., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Rae, J. W., Vinyals, O., and Sifre, L. (2023). Training compute-optimal large language models.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Kim, A. G., Muhn, M., and Nikolaev, V. V. (2023). Bloated disclosures: Can ChatGPT help investors process information? *SSRN Electron. J.*

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, 45(2):221–247. Economic Consequences of Alternative Accounting Standards and Regulation.

Lopez-Lira, A. and Tang, Y. (2023). Can ChatGPT forecast stock price movements? return predictability and large language models. *SSRN Electron. J.*

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

Loughran, T. and McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4):1187–1230.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners.

Rouen, E., So, E. C., and Wang, C. C. (2021). Core earnings: New data and evidence. *Journal of Financial Economics*, 142(3):1068–1091.

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

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In Erk, K. and Smith, N. A., editors, *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Shiller, R. J. (1981). Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends? *American Economic Review*, 71(3):421–436.

Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., and Liu, Y. (2024). Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C).

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

Touvron, H., Cord, M., El-Nouby, A., Bojanowski, P., Joulin, A., Synnaeve, G., and Jégou, H. (2021). Augmenting convolutional networks with attention-based aggregation.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). Llama: Open and efficient foundation language models.

Wang, S. and Levy, B. (2024). Beancounter: A low-toxicity, large-scale, and open dataset of business-oriented text.

Appendices

B. Examples of Disclosure Generated by the LLMs

In this section we present examples of narrative disclosures generated by conditioning the LLMs we train on a particular prompt and then sampling from the model. First, we condition the models with a prompt that is similar to what the SEC recommends for the start of Item 1. Business Description, presented in Table B1. After conditioning we then sample from the model using beam search. Table B2 presents examples generated using our cross-sectional LLM described in Section 3.4.1. Table B3 presents examples generated using a sample of the firm-specific LLMs described in Section 3.4.2.

Table B1. Examples of Disclosures Generated by the LLMs

Item 1. Business Description
Our principal operating activities consist of*[...]*

Table B2. Examples of Disclosures Generated by the LLMs

Sample	Generated Text
1	*[…]*the design, development, manufacture, marketing and sale of digital radio systems. Our principal research and development resources are located in the United States, Belgium, Germany and Canada.*[…]*
2	*[…]*the production, storage, processing and sale of natural gas as well as associated operating, general, and administrative costs necessary to provide benefits to our customers and to operate our business, including our marketing of natural gas, storage and processing, gas gathering, and transportation services.*[…]*
3	*[…]*developing and marketing our proprietary drug delivery technologies, research and development of drug delivery technologies associated with drug compounds and other non-topharmaceutical products and the provision of related clinical, regulatory, manufacturing and other related services. For the year ended December 31, 2004, revenues represented $13.4 million or 95% of our expenses.*[…]*
4	*[…]*the operation of our network infrastructure, including those related to our satellite service; data centers and communications facilities, including those located in Europe, Latin America and Canada; sales and marketing; and administrative activities.*[…]*

Table B3. Examples of Disclosures Generated by the LLMs

Firm	Generated Text
Apple	*[…]*the design, development, marketing and sale of personal computers and related software, the activities of which are conducted throughout the Company's operating segments. The Company's products consist of the Macintosh line of computers, related peripherals and services, and the Mac OS® X operating system (the Macintosh). The Company sells its products worldwide through its online stores, its retail stores, its direct sales force, and third-party wholes*[…]*
Cummins	*[…]*the operations of our Engine Segment, which is described in Note 17 to the Consolidated Financial Statements included in our 2004 Annual Report on Form 10-K. The Engine Segment produces engines and parts for sale to customers in on-highway, commercial, institutional and mining applications. Our Power Generation Segment produces engines and parts for power generation equipment and is headquartered in Columbus, Indiana. We have manufacturing plants worldwide, including major operations*[…]*
Pfizer	*[…]*the discovery, development, manufacture, marketing and sale of prescription medicines for humans and animals as well as our over-the-counter products. We also purchase products from numerous suppliers. These purchases are generally made through various means including through our operating subsidiaries. Wherever possible, commercial relationships are maintained with our suppliers. Where possible, these relationships can be terminated early with 90 days' notice. We believe that we have good relationships with our suppliers. We made significant progress during 1999 in the development of our pharmaceutical products, manufacturing processes and commercialized products. We have successfully enhanced these systems and are implementing significant changes in our internal systems and processes including: • Alond, a treatment for schizophrenia, was launched in September 1999 to increase capacity and efficiency in our manufacturing operations • Exubera, our inhaled insulin medication for chronic obstructive*[…]*
Weyco Group	*[…]*the manufacture and sale of men's casual, dress and fashion footwear. The majority of foreign-sourced purchases are denominated in U.S. dollars. The Company also has a small number of retail stores in Europe and, accordingly, the Company's results are primarily impacted by the economic conditions and the retail environment in the United States. We have historically generated adequate cash flow from operations to meet working capital requirements. In 2003, cash flows from operations was $7.5 million, as compared to $6.7 million last year.*[…]*

Figure 1. Model Training and Inference

This figure presents the data partitions used for model training and inference. The dotted line encompasses the data used for pretraining the cross-sectional model described in Section 3.4.1. The dashed line encompasses the data used for continued pretraining of the cross-sectional model on a particular firm's disclosures, as described in Section 3.4.2. Firm-specific models are iteratively applied and updated using their respective disclosures from 2007 through 2023 to generate out-of-sample estimates of the information contained in each disclosure–indicated by the small dashed line and arrows pointing from one disclosure to the next.

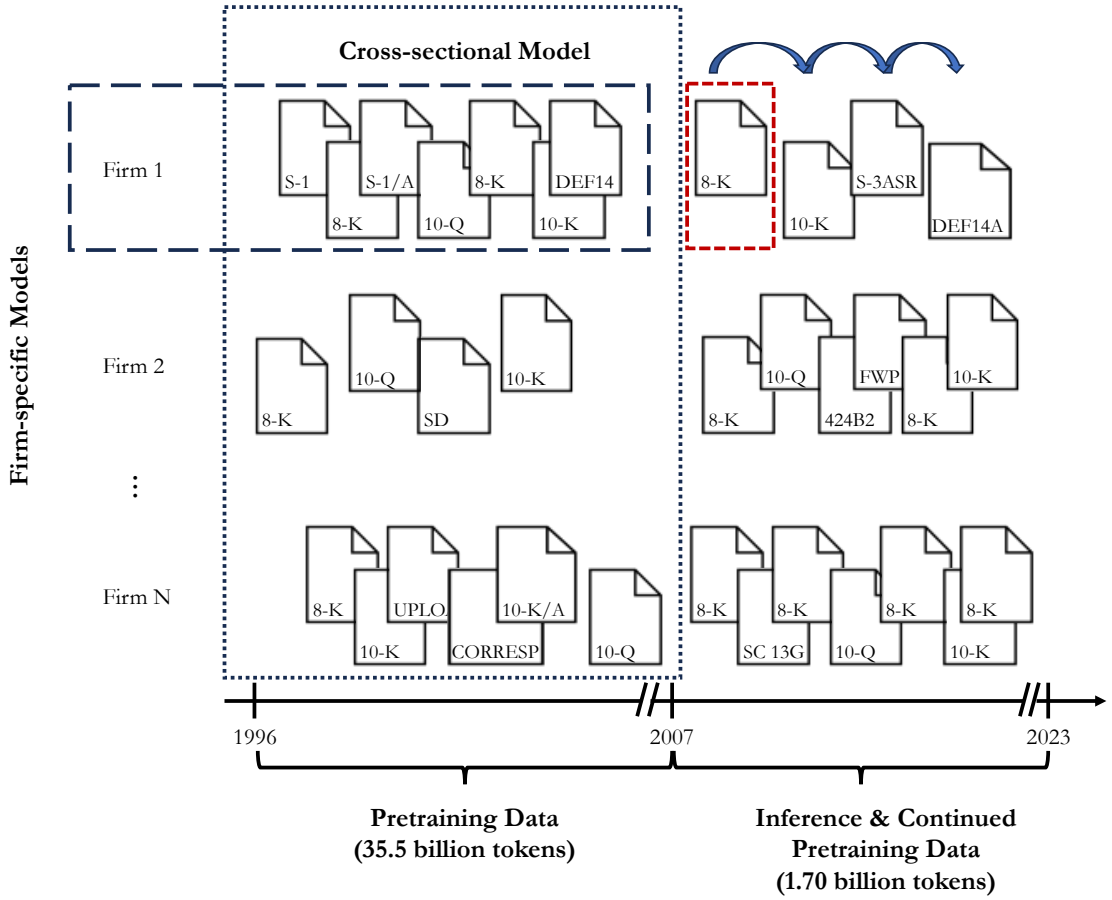


Figure 2. Example of Model Inference for First Mention of Apple iPhone

This figure presents inference results from a model fine-tuned on Apple's disclosures from 1996 through 2006. Each sentence is highlighted in green based on the mean of the token-by-token measurement of information, with darker green representing more information. The disclosure was filed with the SEC via EDGAR on January 17, 2007.

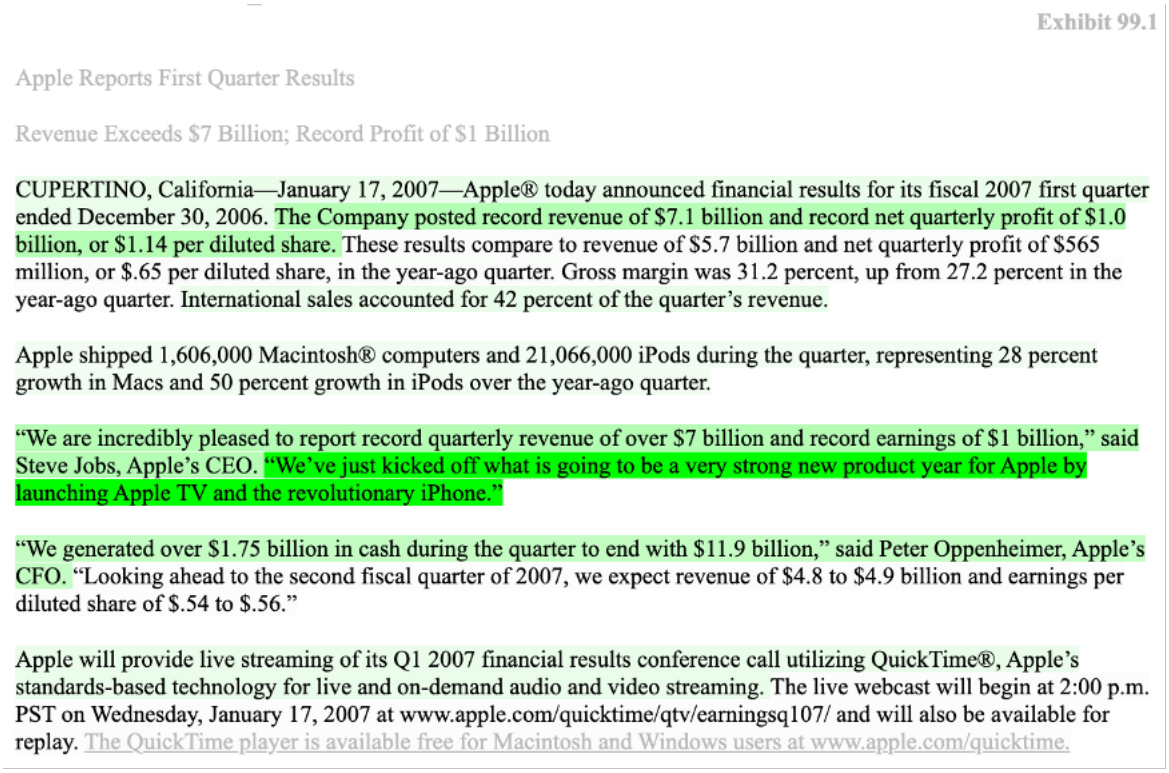


Exhibit 99.1

Apple Reports First Quarter Results

Revenue Exceeds $7 Billion; Record Profit of $1 Billion

CUPERTINO, California—January 17, 2007—Apple® today announced financial results for its fiscal 2007 first quarter ended December 30, 2006. The Company posted record revenue of $7.1 billion and record net quarterly profit of $1.0 billion, or $1.14 per diluted share. These results compare to revenue of $5.7 billion and net quarterly profit of $565 million, or $.65 per diluted share, in the year-ago quarter. Gross margin was 31.2 percent, up from 27.2 percent in the year-ago quarter. International sales accounted for 42 percent of the quarter's revenue.

Apple shipped 1,606,000 Macintosh® computers and 21,066,000 iPods during the quarter, representing 28 percent growth in Macs and 50 percent growth in iPods over the year-ago quarter.

"We are incredibly pleased to report record quarterly revenue of over $7 billion and record earnings of $1 billion," said Steve Jobs, Apple's CEO. "We've just kicked off what is going to be a very strong new product year for Apple by launching Apple TV and the revolutionary iPhone."

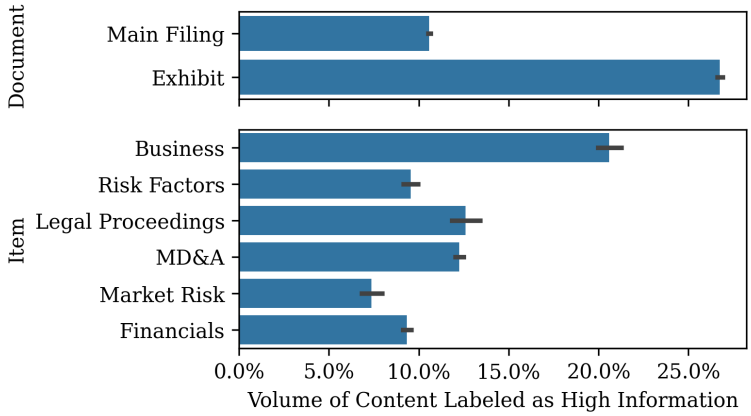
"We generated over $1.75 billion in cash during the quarter to end with $11.9 billion," said Peter Oppenheimer, Apple's CFO. "Looking ahead to the second fiscal quarter of 2007, we expect revenue of $4.8 to $4.9 billion and earnings per diluted share of $.54 to $.56."

Apple will provide live streaming of its Q1 2007 financial results conference call utilizing QuickTime®, Apple's standards-based technology for live and on-demand audio and video streaming. The live webcast will begin at 2:00 p.m. PST on Wednesday, January 17, 2007 at www.apple.com/quicktime/qtv/earningsq107/ and will also be available for replay. The QuickTime player is available free for Macintosh and Windows users at www.apple.com/quicktime.

Figure 3. Location of Information within Filings

This figure plots the volume of content in the highest quartile of information by various form types and common sections within each the form type. Panel A presents averages for Forms 10-K/Q separately by the main filing and exhibits as well as for common items. Panel B presents averages for Forms 8-K separately by the main filing and exhibits as well as for common items. Sample of 18,265 Form 10-K/Q filings and 49,452 Form 8-K filings.

Panel A. Annual and Quarterly Reports (Forms 10-K/Q)



Panel B. Current Reports (Forms 8-K)

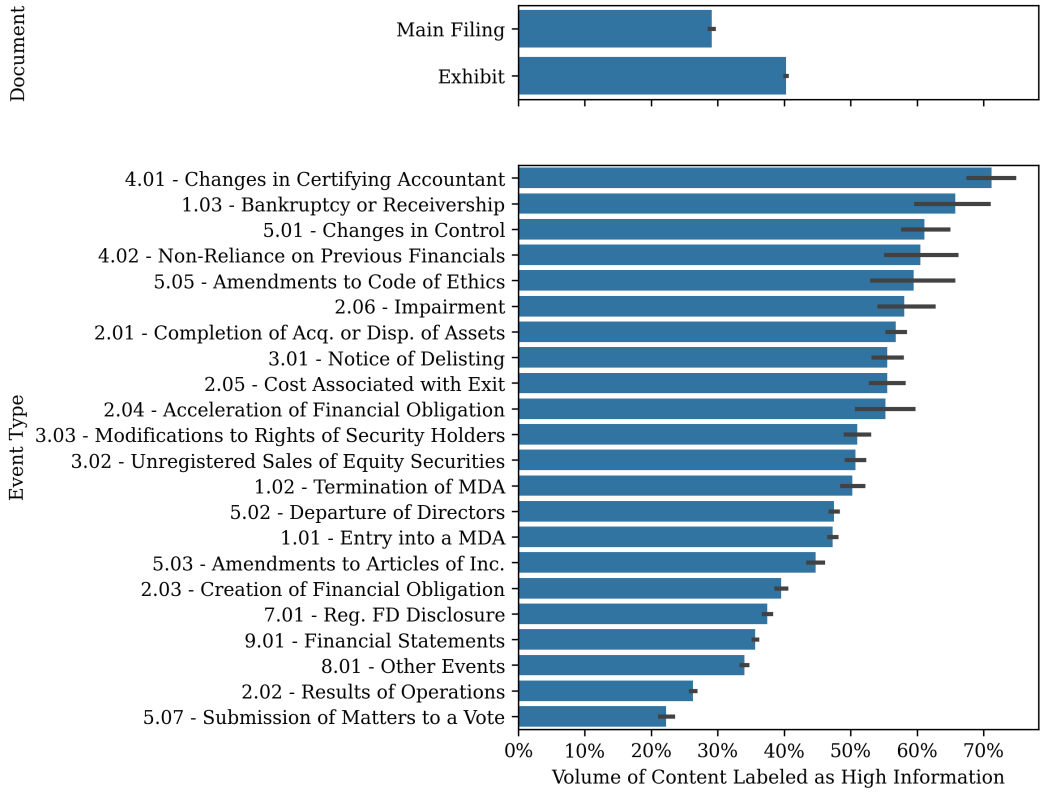


Figure 4. Information Flow Throughout the Reporting Year

This figure plots the flow of information throughout the reporting year. Information is measured using our LLM-based approach and using daily stock volatility. For the latter measure, we measure volatility using absolute abnormal returns and attribute volatility to filings based on a tight [-1,1] event window around the release of the filing. The cumulative information released during the reporting year is used to normalize the measures such that they sum to 100%. In the case of daily stock volatility, the cumulative amounts sum to less than 100% because we it is volatility not attributable to disclosure based on our approach. Sample of 4,656 Form 10-K, 13,609 Form 10-Q, 49,452 Form 8-K, and 49,857 other filings.

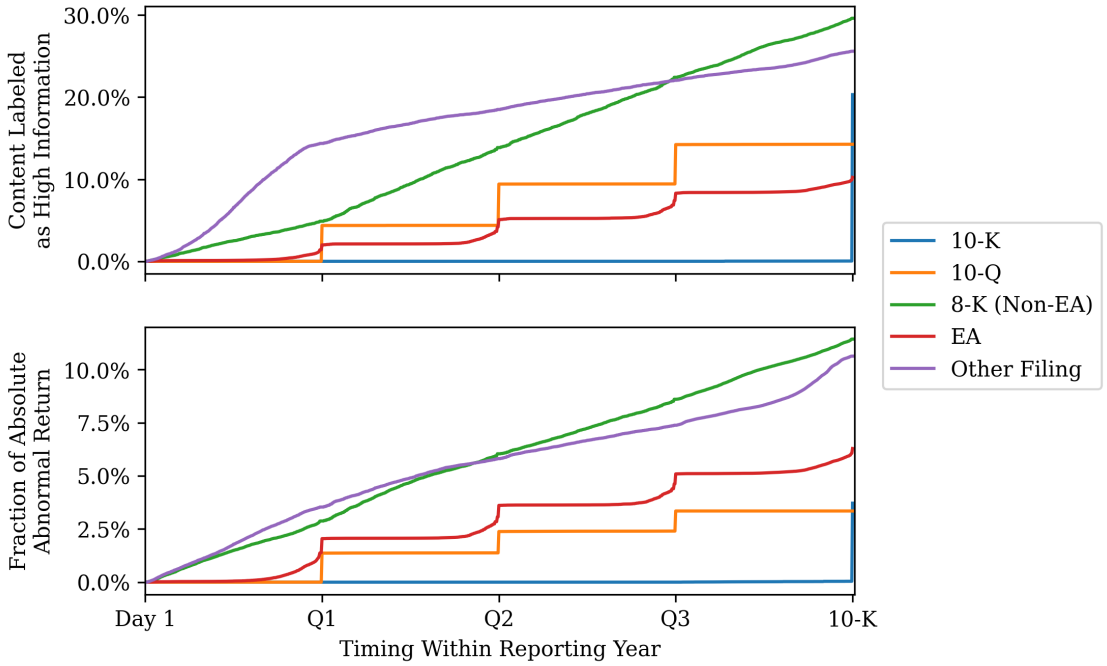


Figure 5. Market Reaction Event Study

This figure plots the average value of daily absolute returns and trading volume in the [-10, 10] window around the release of each filing, separately for filings in the top and bottom quartiles of *Info*. Day 0 represents the first day markets were open following acceptance of the filing on EDGAR. Specifically, Day 0 for filings accepted either prior to markets opening or intraday is the filing date, whereas Day 0 for filings accepted after markets close is the date of the following trading session. Sample of 188,457, 174,033, and 518,848 form-days in the [–10,+10] day window around 9,129 Form 10-K/Q filings, 8,430 MD&A sections extracted from 10-K/Q filings, and 24,726 Form 8-K filings.

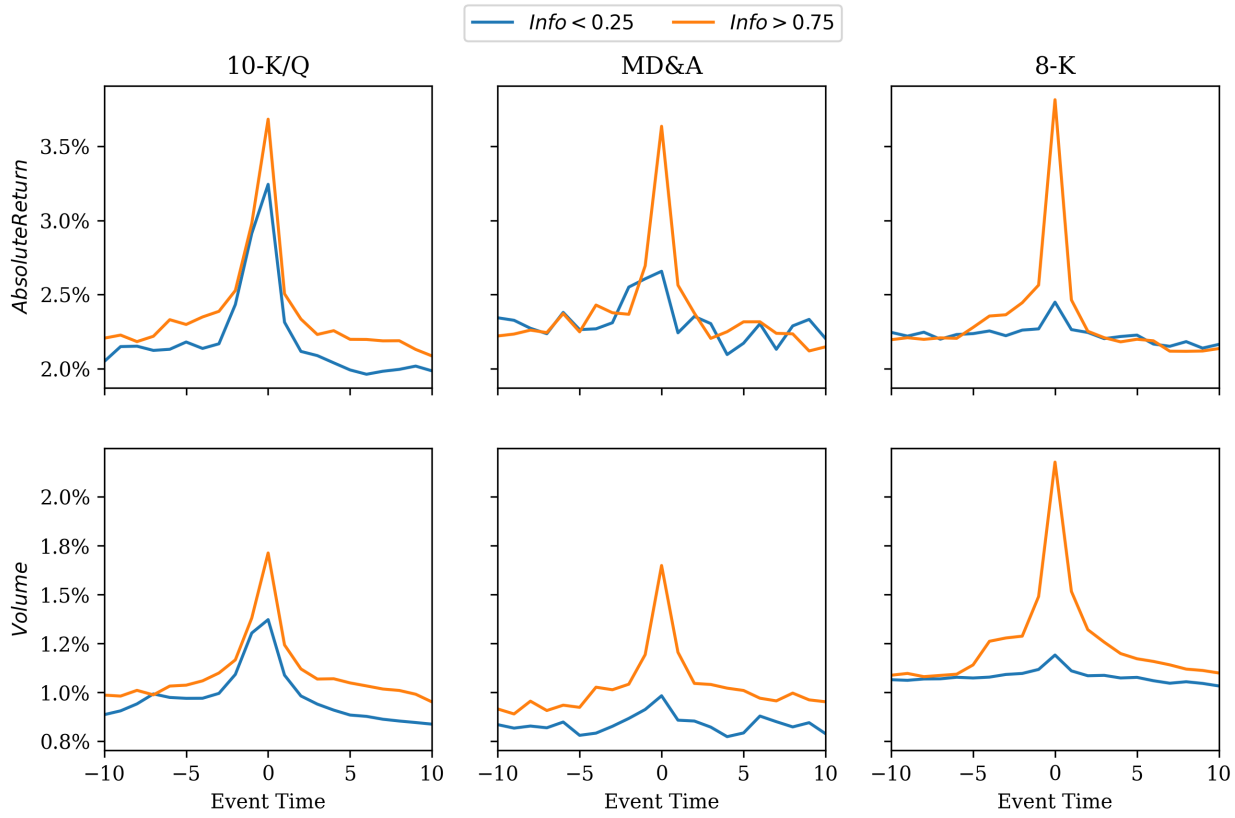


Figure 6. Market Reaction and Information

This figure plots the average market reaction to filings as a function of *Info*. The market reaction is measured on the first day markets were open following acceptance of the filing on EDGAR. Specifically, the market reaction to filings accepted either prior to markets opening or intraday is measured on the same day, whereas the reaction to filings accepted after markets close is measured the following trading session. Sample of 18,265 Form 10-K/Q filings, 16,861 MD&A sections extracted from 10-K/Q filings, and 49,452 Form 8-K filings.

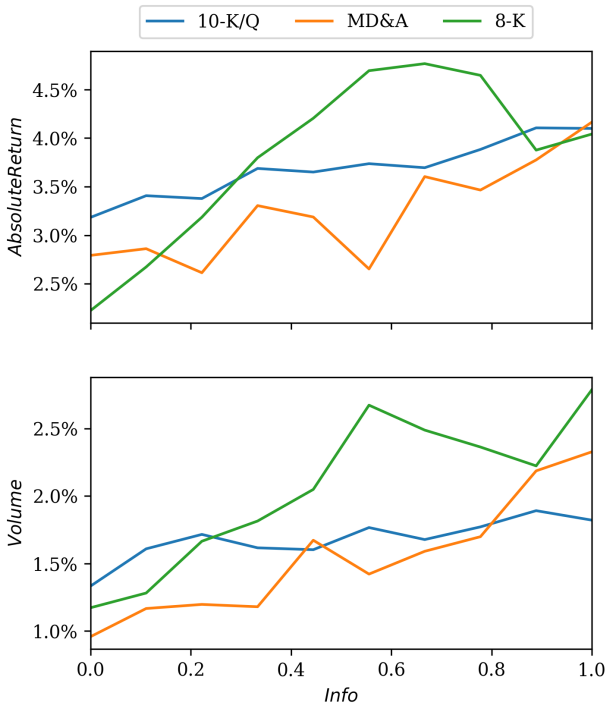


Figure 7. Market Reaction and Signed Information

This figure plots the average market reaction to filings as a function of *Sentiment*. The market reaction is measured on the first day markets were open following acceptance of the filing on EDGAR. Specifically, the market reaction to filings accepted either prior to markets opening or intraday is measured on the same day, whereas the reaction to filings accepted after markets close is measured the following trading session. Panel A (Panel B) presents the average value of *AbsoluteReturn* (*Volume*) by decile of Info. Sample of 9,129 Form 10-K/Q filings, 8,430 MD&A sections extracted from 10-K/Q filings, and 24,726 Form 8-K filings.

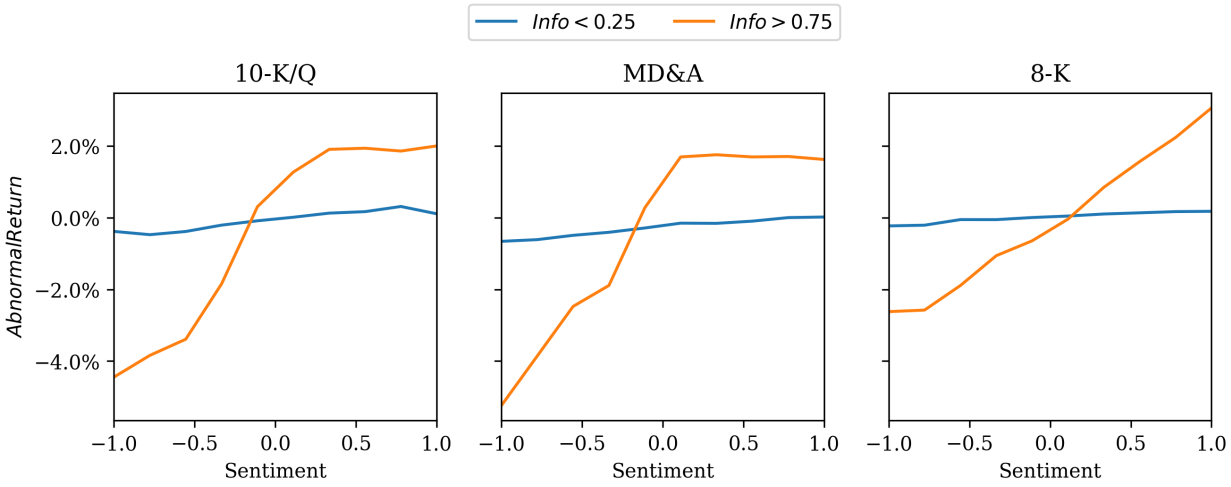


Figure 8. Document Embedding Overview

This figure presents our approach to generating a time-series of disclosure representations. We divide lengthy documents into chunks of roughly equal length, feed each chunk through an embedding model, and create a chunk-level representation by mean pooling over the output of the embedding model. We carry out this procedure for all disclosures a firm files on EDGAR to generate a time-series of chunk-level representations.

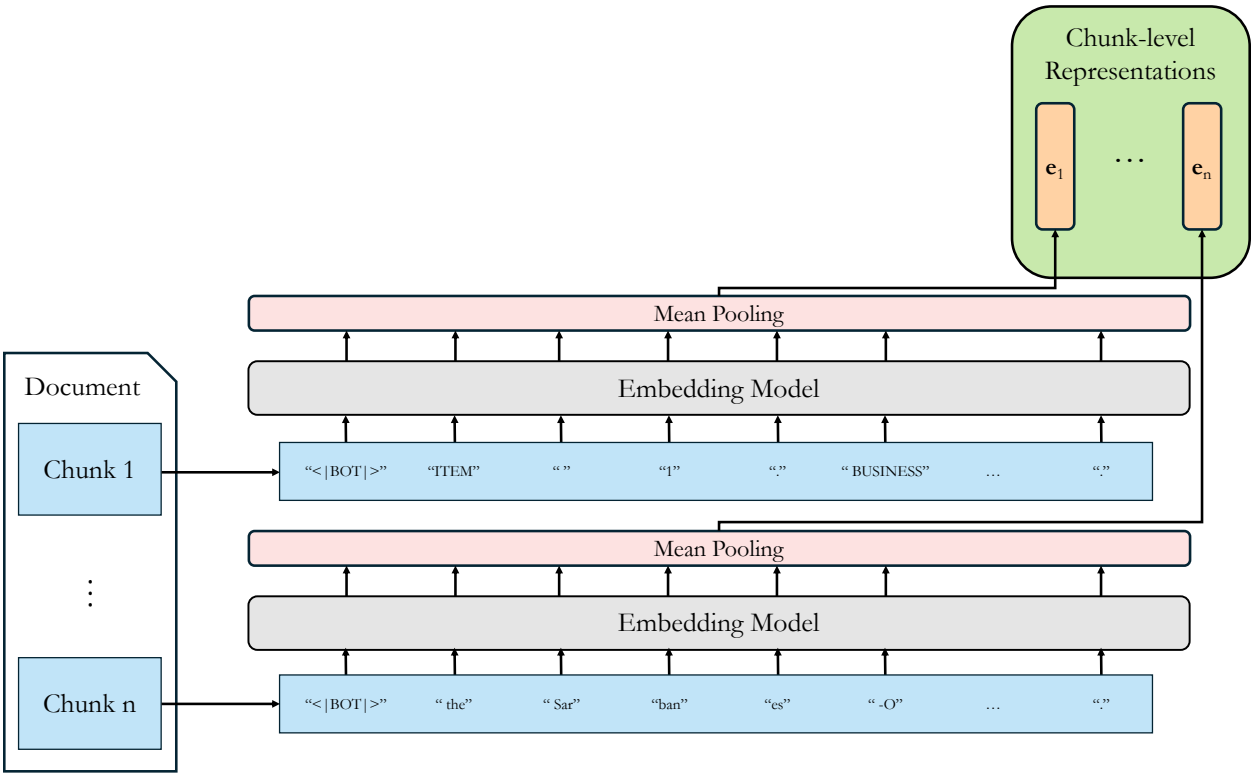


Figure 9. Learned Aggregation Overview

This figure presents our approach to learning representations of firms' time-series of disclosure which are predictive of future returns. We take a time-series of disclosure representations, use RoPE to add within document position encodings, use RoPE to add global temporal encodings relative to the newest filing in the time-series, and add a learned document type embedding. We prepend to the time-series a set of learnable class (CLS) tokens, one token for each return horizon. This sequence is then passed through a single self-attention layer. A shared MLP head uses the output CLS tokens to predict binary market sentiment at each horizon.

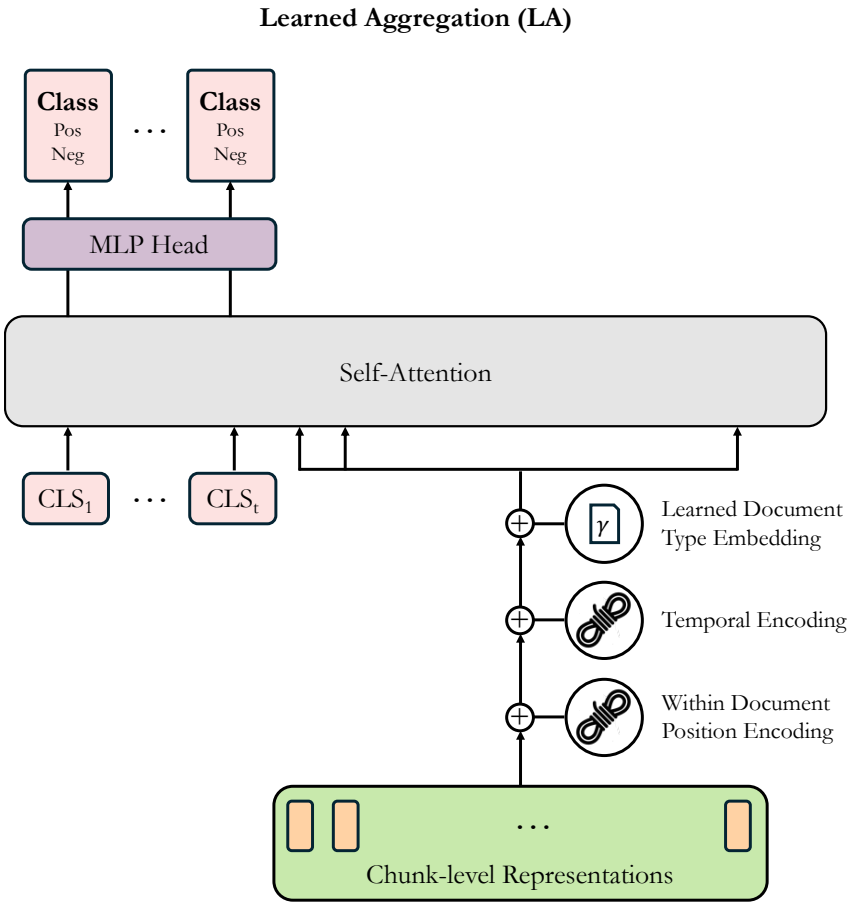


Figure 10. Value-relevant Information

This figure plots the volume of content in the highest quartile of *Info* (blue filled bar) and the portion of that content which is also in the highest quartile of value-relevance (blue hatched bar). The volume of content is separated according to the topics identified by Dyer et al. (2017). Value-relevance is measured using the weight placed on each piece of content when creating a representation whose mutual information with returns is maximized–see Section 4.7 for details. Sample of 18,265 Form 10-K/Q filings and 49,452 Form 8-K filings.

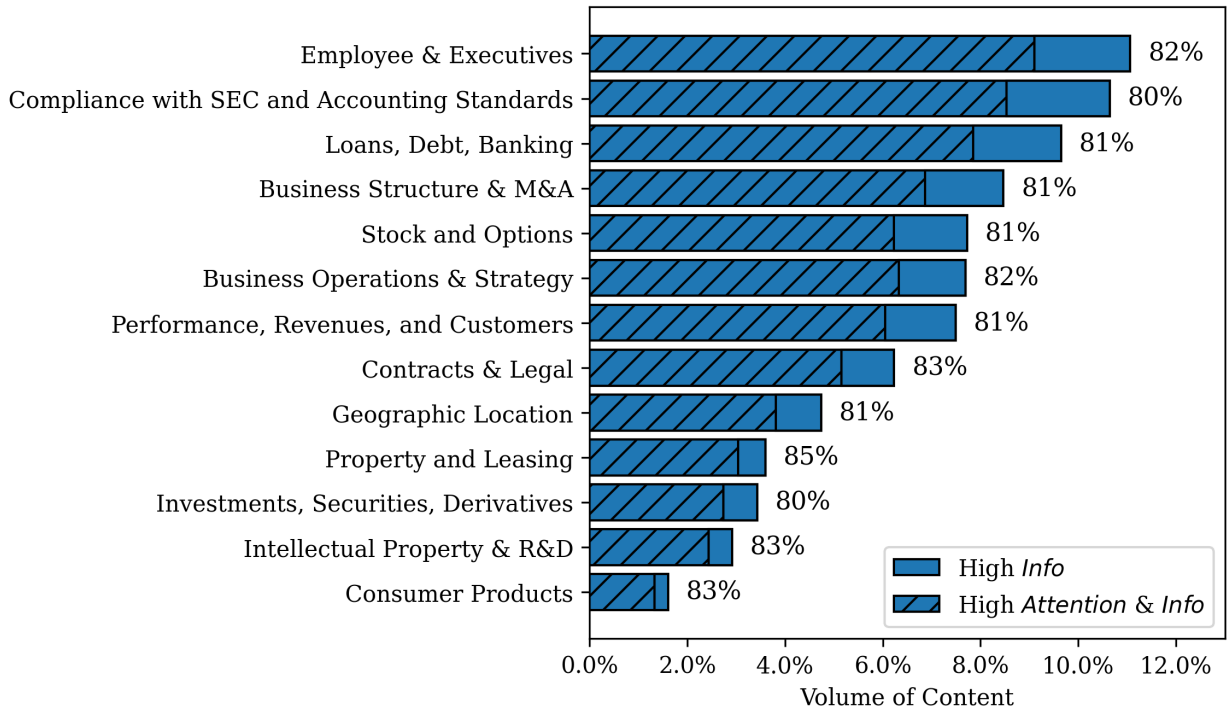


Table 1. Volume of Narrative Disclosure on SEC EDGAR

This table presents the volume and sources of narrative disclosure used to train the models in this paper. Panels A, B, and C are based on samples of 2,478,642, 60,304, and 138,737 filings, respectively.

Panel A. Data for pretraining cross-sectional model (all filings prior to 2007)

Form	Characters (M)	Tokens (M)	Tokens/Filing	Sentences (M)	% of All	Main %	Exhibits %
All	191,251	35,461	14,306	894	100%	66%	34%
8-K	21,146	3,940	7,571	74	11%	1%	10%
485BPOS	21,233	3,919	65,357	115	11%	10%	1%
10-Q	19,308	3,581	13,724	92	10%	7%	4%
10-K	16,315	3,021	42,009	82	9%	5%	3%
S-4	7,016	1,287	144,677	27	4%	2%	1%

Panel B. Data for continued pretraining of firm-specific models (sample firms' filings prior to 2007)

Form	Characters (M)	Tokens (M)	Tokens/Filing	Sentences (M)	% of All	Main %	Exhibits %
All	4,086	755	12,512	19	100%	57%	43%
10-Q	826	152	17,637	4	20%	13%	7%
10-K	681	125	55,101	3	17%	10%	6%
8-K	634	118	6,164	2	16%	2%	14%
DEF 14A	230	43	17,089	1	6%	6%	0%
S-4	184	34	134,105	1	4%	3%	1%

Panel C. Data for rolling forward firm-specific models (sample firms' filings from 2007 onwards)

Form	Characters (M)	Tokens (M)	Tokens/Filing	Sentences (M)	% of All	Main %	Exhibits %
All	9,252	1,696	12,221	40	100%	61%	39%
8-K	2,361	437	6,948	8	26%	3%	23%
10-Q	2,220	404	27,585	10	24%	18%	6%
10-K	1,932	348	70,205	10	21%	16%	4%
DEF 14A	812	148	34,637	4	9%	9%	0%
S-4	170	32	145,044	1	2%	2%	0%

Table 2. **Sample Construction**

This table presents the impact of various filters and random sampling methods on our sample construction. Final sample consists of 500 firms stratified by firm-size decile.

Criteria	**N-firms**
Valid link between Compustat and CRSP as of 2006-12-31	6,312
Financial statement criteria: $atq > 0$, $saleq > 0$, $csohq > 0$, $prccq > 0$, ibq.notnull()	5,643
Stock criteria: $shrcd <= 11$, $primexch \in \{A, N, Q\}$	4,548
Non-financial firms	3,558
Random sample	500
Decile 0: ($0, $39]	50
Decile 1: ($39, $85]	50
Decile 2: ($85, $152]	50
Decile 3: ($152, $263]	50
Decile 4: ($263, $426]	50
Decile 5: ($426, $713]	50
Decile 6: ($713, $1,194]	50
Decile 7: ($1,194, $2,212]	50
Decile 8: ($2,212, $5,928]	50
Decile 9: ($5,928, >$5,928)	50

Table 3. **Measure of Information**

This table presents descriptive statistics for our measure of information. For each firm we train a firm-specific large language model (LLM) using filings from 1996 through 2006. We then apply the firm-specific models out-of-sample, i.e., on filings from 2007 onwards, to measure the information in new filings. Specifically, for each new filing we iteratively measure the information token-by-token according to Eq. 6 and then update the respective firm-specific LLM. We define the average information in a filing as the mean of the set of token-by-token measurements. Sample of 4,656 Form 10-K, 13,609 Form 10-Q, 49,452 Form 8-K, and 49,857 other reports.

	Annual Reports (Forms 10-K)		Quarterly Reports (Forms 10-Q)		Current Reports (Forms 8-K)		Other Reports	
	Main	Exhibits	Main	Exhibits	Main	Exhibits	Main	Exhibits
Mean	1.244	1.430***	1.046	0.770***	1.533	2.854***	1.898	1.143***
Percentile								
5-th	0.632	0.086	0.424	0.000	0.067	0.241	0.001	0.001
25-th	0.906	0.361	0.703	0.022	0.354	1.451	0.348	0.004
50-th	1.153	0.905	0.967	0.121	1.126	2.553	1.162	0.180
75-th	1.472	1.957	1.300	0.821	2.433	3.980	2.983	1.190
95-th	2.147	4.779	1.933	4.127	4.258	6.416	5.976	5.954

Table 4. Daily Event Study Descriptive Statistics

This table presents descriptive statistics for the variables used in our daily event study partitioned by inclusion in our random sample. Panel A presents statistics for the year of our sampling and Panel B for the years following our sampling. *AbnormalReturn* is the daily residual from the expected returns model of Fama and French (2015) and Carhart (1997) estimated over the [-70,-11] window relative to the event, expressed as a percent. *AbsoluteReturn* is the absolute value of the daily residual from the expected returns model of Fama and French (2015) and Carhart (1997) estimated over the [-70,-11] window relative to the event, expressed as a percent. *Volume* is the ratio of daily volume of shares traded to shares outstanding, expressed as a percent. *AdExp* is the ratio of advertising expense to sales. *CapEx* is the ratio of capital expenditures to total assets. *CapIntens* is the ratio of property, plant, and equipment to total assets. *CorpAcq* is an indicator variable equal to one if an acquisition accounts for at least 20% of sales. *Financing* is of the sum of equity and debt issuances over the fiscal quarter scaled by total assets. *IdioVol* is the standard deviation of the residual from a Fama and French (2015) and Carhart (1997) six-factor expected returns model using the daily returns over the prior fiscal quarter expressed as a percent. *Leverage* is the natural logarithm of the sum of long-term debt and debt in current liabilities scaled by total assets. *MTB* is the ratio of market to book value of equity. *QtrlyReturn* is the buy and hold return over the fiscal quarter expressed as a percent. *R&D* is the ratio of research and development expense to sales. *ShortTermReturn* is the buy and hold return during the window [t-5, t-1] for each date t expressed as a percent. *ShortTermVolat* is the daily stock return volatility during the window [t-5, t-1] for each date t expressed as a percent. *Size* is the natural logarithm of total assets. *SpecItems* is special items scaled market value of equity. ***, **, and * denote statistical significance at the 0.01, 0.05, and 0.10 levels (two–tail) based on standard errors clustered by firm and date, respectively. Panel A (Panel B) based on a sample of 4,558 (198,422) observations.

Panel A. Year Sampled (2006)

	Sample		All Firms	
Variable	**Mean**	**Median**	**Mean**	**Median**
AbnormalReturn	0.006	0.000	0.006	0.000
AbsoluteReturn	1.826	1.247	1.799	1.223*
Volume	0.815	0.505	0.818	0.508
AdExp	0.056	0.000	0.056	0.002
CapEx	0.050	0.028	0.047*	0.026
CapIntens	0.236	0.147	0.236	0.152
CorpAcq	0.134	0.000	0.130	0.000
Financing	0.166	0.033	0.151**	0.028
IdioVol	0.023	0.019	0.023	0.019
Leverage	0.486	0.461	0.484	0.462
MTB	2.329	1.786	2.242**	1.737
QtrlyReturn	0.085	0.077	0.094	0.078
R&D	0.812	0.000	0.733	0.000
ShortTermReturn	0.298	0.083	0.292	0.077
ShortTermVolat	0.084	0.036	0.083	0.035*
Size	5.873	5.846	5.915	5.846
SpecItems	-0.005	0.000	-0.005	0.000

Table 4 Daily Event Study Descriptive Statistics (Cont'd)

Panel B. Years After Sampling (2007 onward)

	Sample		All Firms	
Variable	**Mean**	**Median**	**Mean**	**Median**
AbnormalReturn	0.003	0.000	0.003	0.000
AbsoluteReturn	2.070	1.333	2.055	1.323*
Volume	0.892	0.581	0.901	0.583
AdExp	0.054	0.000	0.053	0.002
CapEx	0.028	0.028	0.027**	0.026
CapIntens	0.270	0.147	0.261	0.152
CorpAcq	0.211	0.000	0.219	0.000
Financing	0.086	0.033	0.086	0.028
IdioVol	0.024	0.019	0.024	0.019
Leverage	0.535	0.461	0.532	0.462
MTB	1.925	1.786	1.964	1.737
QtrlyReturn	0.023	0.077	0.024	0.078
R&D	0.174	0.000	0.177	0.000
ShortTermReturn	0.155	0.066	0.165	0.081
ShortTermVolat	0.121	0.040	0.120	0.039
Size	6.785	5.846	6.780	5.846
SpecItems	-0.007	0.000	-0.007	0.000

Table 5. Market Reaction

This table presents the market reaction as a function of measures of information. Panel A presents results for annual and quarterly reports on Forms 10-K/Q. Panel B presents results for current reports on Form 8-K. Events are defined as days where a filing is posted to EDGAR. *Event*[0] is first day that the market can react to a filing posted to EDGAR. *Length* is the length of the filing deciled and scaled to range from zero to one. *Info* is the information in the filing. Specifically, for each token we use a firm and time-specific LLM to measure the token's information according to Eq. 6. We take the mean over the set of each token's information then decile and scale the result to range from zero to one. Controls includes *AdExp*, *CapEx*, *CapIntens*, *CorpAcq*, *Financing*, *IdioVol*, *Leverage*, *MTB*, *QtrlyReturn*, *R&D*, *ShortTermReturn*, *ShortTermVolat*, *Size*, and *SpecItems*. Standard errors clustered by firm and date appear in parentheses. ***, **, and * denote statistical significance at the 0.01, 0.05, and 0.10 levels (two–tail), respectively. Panels A, B, and C are based on samples of 376,921, 1,037,697, and 348,071 firm-days in the [–10,+10] day window around 18,258, 49,452, and 16,861 Form 10-K/Q filings, Form 8-K filings, and MD&A items, respectively.

Panel A. Annual and Quarterly Reports (Forms 10-K/Q)

	AbsoluteReturn			*Volume*		
	(1)	(2)	(3)	(4)	(5)	(6)
Event[0]	1.14***	0.30	0.71***	0.60***	-0.40	0.15
	(0.08)	(0.33)	(0.11)	(0.06)	(0.31)	(0.12)
Event[0] ×*Length*		0.01**			0.01***	
		(0.00)			(0.00)	
Event[0] ×*Info*			0.86***			0.88***
			(0.18)			(0.28)
Controls	Y	Y	Y	Y	Y	Y
Date FEs	Y	Y	Y	Y	Y	Y
Firm FEs	Y	Y	Y	Y	Y	Y
N-obs	376,921	376,921	376,921	376,921	376,921	376,921
R^2	6.5%	6.7%	6.9%	9.0%	9.0%	9.1%

Panel B. Current Reports (Form 8-K)

	AbsoluteReturn			*Volume*		
	(1)	(2)	(3)	(4)	(5)	(6)
Event[0]	1.39***	1.00***	0.37***	0.99***	0.55***	0.29
	(0.06)	(0.06)	(0.06)	(0.13)	(0.08)	(0.20)
Event[0] ×*Length*		0.09***			0.10***	
		(0.01)			(0.03)	
Event[0] ×*Info*			2.04***			1.41***
			(0.13)			(0.33)
Controls	Y	Y	Y	Y	Y	Y
Date FEs	Y	Y	Y	Y	Y	Y
Firm-Yr-Qtr FEs	Y	Y	Y	Y	Y	Y
N-obs	1,037,697	1,037,697	1,037,697	1,037,697	1,037,697	1,037,697
R^2	0.30%	0.42%	1.2%	0.11%	0.13%	0.15%

Table 5. Market Reaction (Cont'd)

Panel C. Management's Discussion and Analysis (Extracted from Forms 10-K/Q)

	AbsoluteReturn			*Volume*		
	(1)	(2)	(3)	(4)	(5)	(6)
Event[0]	0.81***	0.47***	0.13	0.44***	0.23***	-0.06
	(0.09)	(0.13)	(0.14)	(0.07)	(0.09)	(0.10)
Event[0] ×*Length*		0.08***			0.05**	
		(0.03)			(0.02)	
Event[0] ×*Info*			1.37***			1.00***
			(0.28)			(0.25)
Controls	Y	Y	Y	Y	Y	Y
Date FEs	Y	Y	Y	Y	Y	Y
Firm FEs	Y	Y	Y	Y	Y	Y
N-obs	348,071	348,071	348,071	348,071	348,071	348,071
R^2	8.1%	8.4%	9.3%	2.3%	2.4%	2.9%

Table 6. Cross-sectional Variation in Market Reaction

This table presents tests for cross-sectional variation in the market reaction. Panel A presents results for annual and quarterly reports on Forms 10-K/Q. Panel B presents results for current reports on Form 8-K. Events are defined as days where a filing is posted to EDGAR. *FinNeg* is the count of the number of negative words in the filing scaled by the length of the filing (Loughran and McDonald, 2011) multiplied by negative one, then deciled and scaled to range from negative one to one. *IWS* is the *token-by-token* interaction between information and sentiment. Specifically, for each token we use a firm and time-specific LLM to measure the token's information according to Eq. 6 and sign the information based on the Fin-Neg dictionary (Loughran and McDonald, 2011). We take the mean over the set of signed information then decile and scale the result to range from negative one to one. All other variables previously defined. Standard errors clustered by firm and date appear in parentheses. ***, **, and * denote statistical significance at the 0.01, 0.05, and 0.10 levels (two–tail), respectively. Panel A (Panel B) is based on a sample of 376,921 (1,037,697) firm-days in the [–10,+10] day window around 18,258 (49,452) Form 10-K/Q (Form 8-K) filings.

Panel A. Annual and Quarterly Reports (Forms 10-K/Q)

	AbnormalReturn			
	(1)	(2)	(3)	(4)
Event[0]	-0.18***	-0.22**	-2.59***	-2.62***
	(0.05)	(0.09)	(0.16)	(0.18)
Event[0] ×*Sentiment*		0.09		0.05
		(0.15)		(0.16)
Event[0] ×*IWS*			4.84***	4.84***
			(0.30)	(0.30)
Controls	Y	Y	Y	Y
Date FEs	Y	Y	Y	Y
Firm FEs	Y	Y	Y	Y
N-obs	376,921	376,921	376,921	376,921
R^2	0.00%	0.01%	5.08%	5.08%

Panel B. Current Reports (Form 8-K)

	AbnormalReturn			
	(1)	(2)	(3)	(4)
Event[0]	0.05	-0.28***	-1.88***	-2.21***
	(0.03)	(0.08)	(0.10)	(0.13)
Event[0] ×*Sentiment*		0.66***		0.67***
		(0.11)		(0.11)
Event[0] ×*IWS*			3.87***	3.87***
			(0.20)	(0.20)
Controls	Y	Y	Y	Y
Date FEs	Y	Y	Y	Y
Firm-Yr-Qtr FEs	Y	Y	Y	Y
N-obs	1,037,697	1,037,697	1,037,697	1,037,697
R^2	0.11%	0.19%	2.56%	2.64%

Table 7. Content Subsets

This table presents results from re-estimating the specifications in Table 5 (columns 1 through 6 here) and Table 6 (columns 7 through 8 here) using LLMs trained on particular subsets of content rather than *all* content. The subsets considered are: only annual reports (10-K), annual and quarterly reports (10-K/Q), and only current reports (8-K). All other variables previously defined. Standard errors clustered by firm and date appear in parentheses. ***, **, and * denote statistical significance at the 0.01, 0.05, and 0.10 levels (two–tail), respectively. Panel A (Panel B) is based on a sample of 376,921 (1,037,697) firm-days in the [–10,+10] day window around 18,258 (49,452) Form 10-K/Q (Form 8-K) filings.

Panel A. Annual and Quarterly Reports (Forms 10-K/Q)

	AbsoluteReturn			*Volume*			*AbnormalReturn*		
Content Subset	10-K	10-K/Q	8-K	10-K	10-K/Q	8-K	10-K	10-K/Q	8-K
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Event[0]	1.21***	1.32***	1.04***	0.61***	0.66***	0.51***	-0.37**	-0.30***	-0.39***
	(0.09)	(0.12)	(0.07)	(0.08)	(0.14)	(0.06)	(0.18)	(0.09)	(0.10)
Event[0] ×*Info*	-0.34**	-0.42**	3.41***	-0.07	-0.14	2.77***			
	(0.16)	(0.18)	(0.59)	(0.19)	(0.24)	(1.03)			
Event[0] ×*IWS*							0.24	0.25*	0.46***
							(0.20)	(0.15)	(0.16)
Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y
Date FEs	Y	Y	Y	Y	Y	Y	Y	Y	Y
Firm FEs	Y	Y	Y	Y	Y	Y	Y	Y	Y
N-obs	376,921	376,921	376,921	376,921	376,921	376,921	376,921	376,921	376,921
R^2	6.57%	6.56%	7.27%	8.98%	8.98%	9.13%	0.49%	0.49%	0.53%

Table 7. Content Subsets (Cont'd)

Panel B. Current Reports (Forms 8-K)

	AbsoluteReturn			*Volume*			*AbnormalReturn*		
Content Subset	10-K	10-K/Q	8-K	10-K	10-K/Q	8-K	10-K	10-K/Q	8-K
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Event[0]	1.51***	1.67***	0.56***	1.16***	1.16***	0.54**	-0.07	-0.07	-1.78***
	(0.10)	(0.10)	(0.06)	(0.33)	(0.29)	(0.22)	(0.07)	(0.07)	(0.10)
Event[0] ×*Info*	-0.24*	-0.56***	1.69***	-0.34	-0.35	0.91***			
	(0.14)	(0.14)	(0.13)	(0.49)	(0.41)	(0.31)			
Event[0] ×*IWS*							0.26**	0.26**	3.68***
							(0.10)	(0.11)	(0.18)
Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y
Date FEs	Y	Y	Y	Y	Y	Y	Y	Y	Y
Firm-Yr-Qtr FEs	Y	Y	Y	Y	Y	Y	Y	Y	Y
N-obs	1,037,697	1,037,697	1,037,697	1,037,697	1,037,697	1,037,697	1,037,697	1,037,697	1,037,697
R^2	0.28%	0.30%	0.98%	0.07%	0.07%	0.08%	0.14%	0.13%	2.70%

Table 8. Out-of-Sample Prediction

This table presents results for the five methods of predicting binary market sentiment over a variety of horizons as detailed in Section 4.7. Market sentiment is measured as the sign of the daily residuals cumulated over a given horizon based on the expected returns model of Fama and French (2015) and Carhart (1997). Expected returns models are estimated over the [-70,-11] window relative to each disclosure. *Llama* refers to the methods which rely on text representations generated using the Llama-3.1-8B model. The subscripts *MP*, *IP*, and *LA* refer to three methods for pooling representations from lengthy texts: mean or equal-weighted, information weighted, and learned aggregation. ***, **, and * denote statistical significance at the 0.01, 0.05, and 0.10 levels (two–tail), respectively, based on boot-strapped standard errors clustered by firm and time. Sample of 17,111 Form 10-K/Q and 46,321 Form 8-K filings from 2008 through 2023.

	Sentiment		*IWS*		$Llama_{MP}$		$Llama_{IP}$		$Llama_{LA}$	
Form Type	10-K/Q	8-K	10-K/Q	8-K	10-K/Q	8-K	10-K/Q	8-K	10-K/Q	8-K
Horizon	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1 day	52.20%***	51.79%***	52.42%	52.69%***	51.88%	51.78%	52.70%	52.32%**	54.73%***	54.45%***
1 week	51.92%***	51.46%***	52.15%	51.62%	51.96%	51.66%	53.61%***	52.14%***	53.85%***	54.31%***
1 month	52.06%***	51.49%***	52.85%**	51.93%**	51.37%*	51.30%	53.53%***	51.99%**	53.14%***	53.38%***
3 months	51.17%***	50.74%***	52.01%**	51.06%	50.10%***	50.65%	51.52%	51.59%***	53.89%***	52.47%***
6 months	51.18%***	50.39%*	52.33%***	50.90%**	50.03%***	50.30%	52.91%***	50.87%**	52.71%***	52.02%***
12 months	51.01%***	50.05%	52.01%**	50.73%***	52.83%***	51.95%***	52.46%***	51.85%***	52.63%***	52.05%***