Can AI Read the Minds of Corporate Executives? *

Nicolas Chapados[3], Zhenzhen Fan[4,2], Russ Goyenko[1,2], Issam Hadj Laradji[3], Fred Liu[5,2], and Chengyu Zhang[1,2]

[1] *Desautels Faculty of Management, McGill University*
[2] *Financial Innovations and Risk Management Labs, FIRM*
[3] *ServiceNow Research*
[4] *University of Manitoba*
[5] *University of Guelph*

This Version: April 5, 2024

Abstract

It can. Using textual information from a complete history of regular quarterly and annual filings by U.S. corporations, we train classic machine learning algorithms and large language models, LLMs, to predict future earnings surprises. We first find that the length of MD&A section on its own is negatively associated with future earnings surprises and firm returns in the cross-section. Second, neither sentiment-based nor bag-of-words classic machine learning regression-based approaches are able to "learn" from the past managerial discussions to forecast future earnings. Third, only *finance-objective trained* LLMs have the capacity to "understand" the contexts of previous 10-Q (10-K) releases to predict both positive and negative earnings surprises, and subsequent future firm returns. We find significant, and often hidden in the complexity of presentations, positive and negative informational content of publicly disclosed corporate filings, and superior (to human and classic NLP approaches) abilities of more recent AI models to identify it.

*FIRM Labs: https://firmlabs.ca/.

Corresponding author, Russ Goyenko: ruslan.goyenko@mcgill.ca

1 Introduction

How do economic agents process information, especially when this information is abundant? Textual data have become widely popular in finance (Goldstein et al., 2021), and extracting and processing these data have seen a substantial reduction in costs over past decade. The abundance of signals and complexity in disclosed information leads to investors inattention to subtle but important signals even in the most foundational to the corporate reporting process items such as quarterly and annual, 10-Q (10-K), reports (Cohen et al., 2020).

To cope with the increasing length and excessive complexity of corporate filings (Loughran and McDonald, 2014), and management's incentives to obfuscate negative information by providing irrelevant or immaterial details (Li, 2008), the literature introduced several measures of content analysis. The most prominent one is based on a word list in which each word is categorized as positive or negative, i.e. a manually-built lexicon approach. Early papers in the literature use the word classification in the Harvard IV-4 Psycho-sociological Dictionary to identify positive versus negative financial news content (Tetlock, 2007). Loughran and McDonald (2011) (hereafter, LM 2011) however argue that the Harvard list might not be suitable for finance applications as these words have different connotations in a finance context. LM 2011 create a comprehensive list of positive versus negative words based on 10-K reports, and argue that their negative word list captures the tone of 10-K reports better than the Harvard list.

Another approach, which improves on LM 2011 classification, is bag-of-words type models, and is based on how each word in the lexicon is weighted. It can either be achieved with a linear regression (Jegadeesh and Wu, 2013), or classical machine learning techniques such as support vector machines (Manela and Moreira, 2017).

More recently, however, Cao et al. (2023) find that soon after LM 2011 publication, firms which expect high machine downloads of their 10-K statements from EDGAR, i.e. firms with higher chances of machine algorithmic classification of their reports, start avoiding using LM-negative words. This not only weakens dictionary-based but also bag-of-words approach which heavily relies on word counts in a document. Therefore, as the length of corporate reports has been exponentially increasing over the past decade (Cohen et al., 2020), their classification and identification of positive versus negative information content about future financial performance remains a challenging task.

Economic text is continually generated by human writers in their quest to comprehend and forecast economic phenomena. Over the past few decades, the finance literature has started to harness information from specific text sources such as financial news outlets, regulatory filings, and social media. However, the research agenda aimed at enhancing economic models through

text mining is still in its nascent stage. To date, investigations have primarily delved into a limited subset of text data relevant to the market, frequently concentrating on individual specialized sources (e.g., the front page of The Wall Street Journal or the "risk factor" section of 10-K filings). Additionally, the representation of text information from these sources often remains rudimentary, typically relying on dictionary-based sentiment scores or a basic "bag of words" approach. The limited utilization of text data thus far can be attributed to its inherent lack of regular structure, rendering it considerably more challenging to work with compared to standard numeric datasets. Language, being an incredibly nuanced information encoding system, demands highly intricate models to effectively extract the wealth of information concealed within text. However, the complexity of these models poses a barrier for many researchers, both in terms of technological expertise and computational resources. This implies that the recent foray into textual analysis within finance and economics is merely scratching the surface. Text remains an underutilized data source for comprehending asset markets. Yet, the present challenges in textual analysis foreshadow an intriguing research agenda for the future, wherein economists gradually expand their text corpora and refine their capabilities to extract valuable insights from this resource.

Advanced large language models, LLMs, where ChatGPT after its public release since November 2022 is the best-known one, are different from classic NLP approaches. They do not rely solely on words or their counts, but are able to identify relationships between words, sentences, and paragraphs in a document. These LLMs should theoretically be better at capturing information from financial textual data because of the highly contextualized nature of finance written text, which cannot easily be captured with lexicon-based models. Not just a failing of classic NLP models, market participants themselves cannot grasp subtle management's messages, which are well-hidden in large volumes of text (Cohen et al., 2020).

BERT, (Bidirectional Encoder Representations from Transformers) developed by Google (Devlin et al., 2018), holds a historical significance in the annals of LLMs and NLP as it marked a crucial shift in the creation and application of language models. Before BERT, models predominantly relied on unidirectional or superficially bidirectional understanding of text. BERT brought about a revolution with its deeply bidirectional model, allowing a contextual understanding from both preceding and succeeding words for prediction. This change sparked an influx of research and development in NLP, yielding more advanced models like GPT-2, GPT-3, and RoBERTa. As such, we adopt BERT as our LLM benchmark model.[1]

[1]GPT-3, the foundation of ChatGPT, and all subsequent versions of GPT, unlike BERT, are not available for commercial-free download. Moreover, re-training or fine-tuning these LLMs are extremely computationally expensive. There are of course other models like Robustly Optimized BERT Pre-training Approach (RoBERTa) (Liu et al., 2019), and Open Pre-trained Transformers (OPT) by (Zhang et al., 2022), or T5 (Colin, 2020). For our purpose, they are not significantly different from BERT, which we discuss more in methodology section.

BERT is pre-trained on a large corpus of text which covers a collection of internet-available content without sole focus on the financial context of corporate reports. It can therefore be a noisy model to apply to the classification of corporate filings.

Huang et al. (2022) fine-tune BERT, naming it FinBERT, on 10,000 sentences from financial analyst reports classified into positive, negative or neutral by a human and argue superior classification accuracy of financial reports by FinBERT compared to other lexicon/dictionary based approaches. Thus, FinBERT (Huang et al., 2022) emerges as a viable alternative. FinBERT's strength lies in its fine-tuning process, which hones its ability to discern positive, negative, and neutral sentiments. Importantly, this fine-tuning relies on human-labeled text data, which remains static and reflects a human perspective rather than capturing the broader dynamics of the financial market. Our analysis reveals a compelling insight: the portfolio of stocks identified as having positive sentiment by FinBERT consistently underperforms the portfolio associated with negative sentiment as determined by FinBERT by significant 31 bps per month (or 3.72% per year) in our out-of-sample tests spanning from 01/2003 to 12/2021. Thus, for investment purposes identification, FinBERT provides completely opposite predictions – purchase the negative FinBERT sentiment score stocks and short-sell the positive ones. This opposite to FinBERT sentiment performance difference disappears after risk-adjusting returns with (Fama and French, 2015) and momentum factors. It becomes insignificant. This finding underscores the nuances and challenges of sentiment-based investment strategies in the financial domain.

To date, there has not been a comprehensive study about: (i) how one approach compares to another in identifying positive versus negative information in 10-K (10-Q) reports about the future cash flows and overall firm's financial performance; (ii) whether an accurate classification of corporate reports is possible at all given that management adapts the language when machines are listening (Cao et al., 2023); (iii) whether corporate insiders are able to communicate and market participants are able to grasp and extract the right signals behind the complexity of reports (Cohen et al., 2020). The latter is essentially an indirect test of market efficiency—is there any hidden information remaining undetected, not yet perceived and incorporated into prices by market participants, in public 10-Q and 10-K corporate filings that are scrutinized by humans and machines?

To answer these questions, using the whole history of 10-Q and 10-K reports for the US companies, we run a horse race between three approaches: (i) sentiment group: keyword lexicon sentiment (LM 2011), LLM sentiment classification (FinBERT) measures or simply the length of managerial discussions (MD&A, or Risk Factors sections), (ii) bag-of-words group: a classification, regression-based approach similar to those of Jegadeesh and Wu (2013), or Manela and Moreira (2017) but using broader spectre of classic ML algorithms,

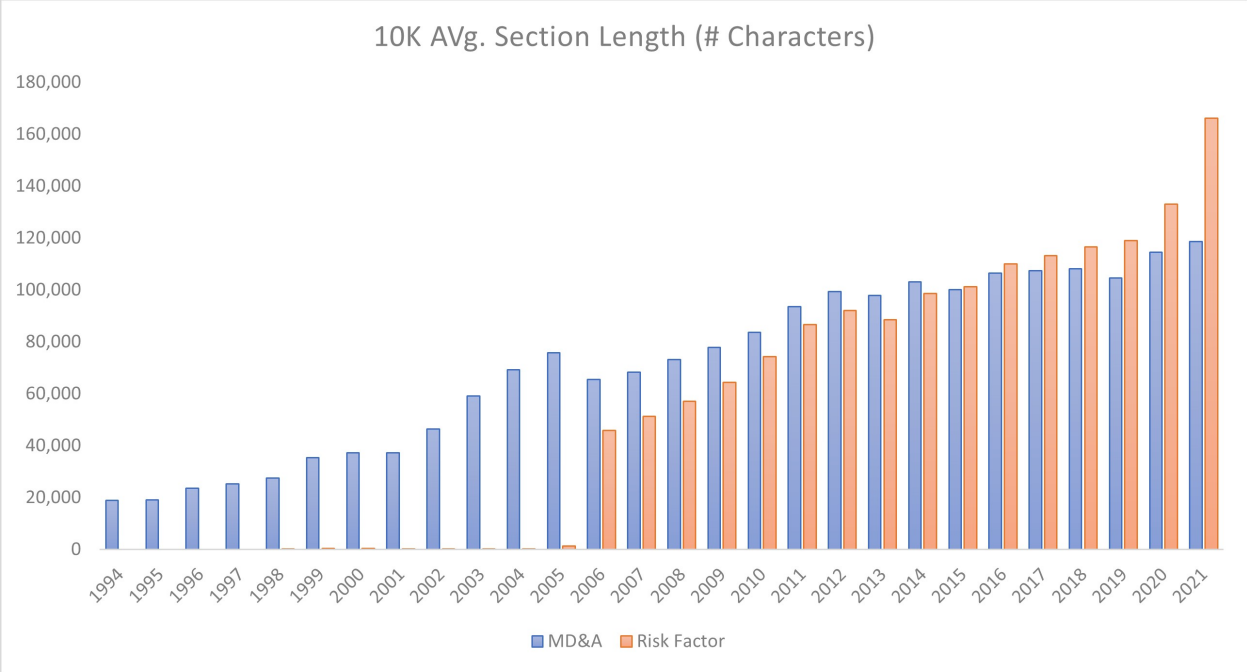

Figure 1: Character Length of MD&A, and Risk Factor sections of 10-K reports over years

and (iii) novel LLMs approaches that we introduce in this paper.

Our analysis focuses, similar to Cohen et al. (2020), on MD&A and Risk Factor sections of reports. Figure 1 shows the average length of MD&A and Risk Factor, RF, sections of 10-K reports over years. From 1994 to 2021, the average length of MD&A section increased six times, while the length of RF section increased four times from 2006 to 2021. This constantly increasing length and complexity of reports makes investors neglect important fundamental information about future firms' performance (Cohen et al., 2020).

What are the rules of the race? Unlike traditional NLP approaches of a sentiment scoring of the documents, in finance we are accustomed to see economic value-added provided by these identifications. In other words, any proposed improvement should materialize into significant future price differences between positively vs. negatively rated firms' reports.

We thus proceed as follows. First, we contribute on the methodology side where unlike dictionary/keyword, LM 2011, or human labelled financial text training approaches (ex.: FinBERT), we train algorithms on financial targets. Most of literature uses earnings announcement day returns, or abnormal returns around earnings announcement windows as financial targets (see among others LM 2011, Jegadeesh and Wu (2013)). These financial targets are based on the assumption of market efficiency. Yet, Cohen et al. (2020) clearly show that there are no market reactions at all to the changes in the format of financial documents which subsequently predict significant financial losses and negative returns within months in

a quarter after the reports become publicly available. Moreover, the earnings announcement returns are also known for the investors' under- or over-reaction to negative and positive surprises respectively (Atmaz and Basak (2018), Golez and Goyenko (2022)). Put together, these make earnings announcement returns a noisy target. Instead, we use next quarter earnings announcement surprise as a financial target, as this fundamental information is audited and does not depend on market's interpretations/reactions. Cohen et al. (2020) also show that the subsequent announcement does indeed reflect information which the market neglected to react to in the previous quarter's announcement. Moreover, earnings surprise as a financial target, which we measure by the deviation of the realized earnings versus the analysts consensus forecast, allows to speak directly to the informational content neglected by the market. While this target is a high bar to pass not treated by previous studies (LM 2011, Jegadeesh and Wu (2013)), this is the only one free of noise. To measure the economic gains from these forecasts, we follow Cohen et al. (2020), and sort all firms into quintile portfolios based on future earnings surprise prediction, and then measure the performance of the strategy which goes long future winners (positive surprise) and short future losers (negative surprise) quintiles.

LLMs like BERT and RoBERTa can process input sequences of up to 512 tokens, translating these tokens into a 1024-dimensional vector representation. In contrast, OPT can manage sequences as long as 1,024 tokens and embed each token into 2,560-dimensional space. Figure 1 above demonstrates that the length of MD&A or RF sections dramatically exceed the lengths that BERT or OPT can read, as it can easily exceed 50K tokens. A common approach in the literature is to use the first, or the last 512 tokens of MD&A sections (see FinBERT, (Huang et al., 2022)). While it can work for analyzing news articles where the summary of the article is normally in the beginning, it is less likely to work for MD&A section where the forward-looking discussions are not uniformly situated in any specific part of the text. Moreover, it is imperative to read the whole text of MD&A and Risk factor sections.

To do so, as our second contribution, we are the first to propose a hierarchical LLM architecture that can process financial disclosure statements of arbitrary lengths, and train these models on a financial target. We first train a model derived from the original, off-the-shelf, BERT (Devlin et al., 2018) to predict earnings surprises, which we name FrozenBERT; i.e. we do not change or fine-tune any parameters in the original BERT model, only train a predictive network, a transformer layer, that builds on the pre-trained BERT representations. Second, we fine-tune the original BERT while training to predict earnings surprises; we name this model FtBERT ("fine-tuned BERT"). To the best of our knowledge, this has never been done for LLMs in the finance literature. FtBERT overcomes the problem of LM 2011 or FinBERT when companies dynamically change and adapt the language to machine readings

(Cao et al., 2023), as we dynamically re-train FtBERT on the newly released fundamentals and the managerial discussions that accompany them. This dynamic re-training is intended to learn, capture, and adjust identifications to possible language adjustments, both in time series and the cross section. FtBERT also overcomes the problem of the original BERT which was trained on a large corpus of general text, and hence can be a noisy representation when it comes to capturing the content of finance-specific documents.

Our results could be of concern for LLMs/ChatGPT-like approaches for purely finance sentiment scoring. While FinBERT (Huang et al., 2022) has been shown to outperform LM 2011 dictionary and other bag-of-words approaches in more precise sentiment identification, it underperforms the most in our portfolio sorting analysis. Similar to Cao et al. (2023), we compute FinBERT negative sentiment score as the number of negatively ranked by FinBERT sentences divided by the total number of sentences in M&A and RF sections. We find that in value-weighted portfolios, the quintile portfolio with the highest negative FinBERT score not only generates future positive returns, but these returns surpass those from the lowest quintile by a significant magnitude, i.e. it has a wrong sign. However, this result does not survive size and book-to-market controls in the regressions and becomes insignificant. Further, neither LM 2011, nor bag-of-words approaches, or even more sophisticated bag-of-words models based on popular feed-forward neural networks, do not provide significant High minus Low quintile portfolio spreads. That is, these approaches, while useful for identifying sentiment in reports, are not useful in predicting future financial performance given information provided therein.

Our surprising result is that a very simple measure, the length of MD&A section itself, is a better predictor of future performance, compared to much more sophisticated sentiment and bag-of-words identifications. Companies with cross-sectionally lower MD&A length, which we simply measure with the total number of characters significantly outperform those with higher MD&A length. For example, in the value-weighted portfolios, the lowest MD&A length quintile has CAPM alpha of 3.8% per annum ($t = 3.35$). After controlling for Fama-French five factors (Fama and French, 2015) and momentum (Carhart, 1997), this number drops to 2.3% ($t = 2.13$). The high minus low strategy, i.e. buying low and selling high MD&A length quintiles, produces CAPM alpha of 4.13% per year ($t = 2.48$) in value weighted portfolios. This alpha becomes insignificant though after controlling for Fama-French five factors. Thus, the markets seem to not fully incorporate into prices the positive information of the short versus less favorable content of long corporate reports related to such basic fundamentals as the size, book-to-market, profitability and investment. This however is consistent with the previous results in the literature where the increasing length of the reports is not associated with the positive information about companies' performance (Li (2008),

Loughran and McDonald (2014), Cohen et al. (2020)). This measure however is not robust in the regressions using various firm and time fixed effects or firm characteristics as control variables.

The best performance, which passes all robustness tests, is observed for the model we introduce in this paper - fine-tuned BERT, namely, FtBERT. Here the quintile with the most positive earnings surprise predictions outperforms the most negative predictions' quintile by 0.56% per month ($t = 2.94$), or 6.74% per year in raw, unadjusted for risk returns. The CAPM risk-adjusted returns of this High minus Low strategy is very similar, 0.5% per month ($t = 2.57$) or 6.01% per year. It is therefore not driven at all by the market trends. The economic significance of this strategy starts decreasing while adding extra factor adjustments from 4% per year with Fama-French five factors to 3.71% per year with all six factors including momentum. While still economically meaningful, these numbers also remain statistically significant at the conventional levels.

We also find that FrozenBERT performs very similarly in portfolio sorts to FtBERT after FF6 factor adjustment, yet it underperforms FtBERT predictions sorted high-minus-low portfolios by approximately 2% per year in raw/unadjusted or only CAPM risk-adjusted returns. However, the performance of FrozenBERT in identifying positive surprises, the long side of high-minus-low strategy, is almost identical to the one of the lowest MD&A length quintile. Therefore, on the long, the most positive earnings surprise prediction portfolio, it fails to outperform a simple letter-counting approach. Yet, it does a superior job in identifying negative/short portfolio selection. In contrast, FtBERT dominates all other approaches in identifying both positive and negative future financial performance.

The latter results can talk to a bigger picture. Pre-trained LLMs are very good in summarizing large texts and we see their applications coming to financial text analysis via prompt summaries. These models can definitely be a superior sentiment-score identification mechanism (Huang et al., 2022). Yet, we find that as they are not specifically trained on financial objectives or specific targets (e.g. earnings surprises' predictions), they do not outperform pre-existing very simple methods in identifying future positive performance. Only and only after being fine-tuned and trained on a specific financial task do they dominate all other approaches, and especially in identifying positive future performance which has been a challenge in the current literature (Loughran and McDonald (2011), Cohen et al. (2020)).

Do we find the evidence that the whole market fails to capture the information we are able to identify with FtBERT? Not at all. Around 10-Q, 10-K reports filing dates, we find that FtBERT is able to correctly predict future price impacts, cumulative post-filing returns, caused by institutional trades who react to the news fast (Huang et al., 2020). However, our results also indicate that for the rest of the market, it takes a few months to fully incorporate

this information into the prices.

Why is this the case? We compute the analyst disagreement measured by standard deviation of analysts forecasts, and find that portfolios with the most negative and the most positive FtBERT prediction signals are also the portfolios where analyst disagreement is the highest. This is not due to small size stocks, as all the stocks in our cross-section are above the average market size. Moreover, the stocks in the highest, the most positive FtBERT predictive signal portfolio, are the largest ones, with the average market cap of approximately $12 billion. This is also the lowest book-to-market value cross-section of stocks which has the highest growth option. Therefore, the market-wide disagreement about future growth options reflected in a generally higher analyst disagreement can be a source of under-reaction, or slow price adjustments.

The attention mechanism that FtBERT uses allows us identifying the paragraphs of MD&A and Risk Factor sections that are the most influential for the future performance forecasts. In general, the future positive performance is related to managerial discussions about future outlook, and immediate short-term measures the management promises to implement to increase future revenues. The high stock performance within the subsequent quarter is associated with the management promises coming through via efficient execution.

In contrast, future negative performance is associated with management discussions either being excessively focused on past performance, or, consistent with Cohen et al. (2020), the Risk Factors section overemphasizing various sector-specific risks. For example the mandatory risk disclosure on clinical trials risks and FDA interference in healthcare sector allows FtBERT to predict future earnings very well, and especially the negative earnings surprises.

The rest of the paper is organized as follows. Section 2 describes the main data in the analysis. Section 3 describes all NLP methodologies that we use in this paper. Section 4 presents the main empirical analysis, and identifies the winner of the race. Section 5 discusses economic channels and sources of superior predictive performance.

2 Data

We obtain the data from several sources. To start our analysis, we retrieve all the 10-K, 10-K405, 10-KSB, and 10-Q filings submitted between 1993 and 2021 from the SEC's EDGAR website. Following Loughran and McDonald (2011), we parse each filing document by removing markup tags, ASCII-encoded graphics, tables, and other non-textual artifacts. We include only one filing per firm in each quarter. In most cases, we use 10-Qs in the first three quarters and 10-Ks (or 10-K405, 10-KSB, whichever applies) in the last quarter of each firm's fiscal year.

We focus on two primary types of corporate disclosures: Management Discussion and Analysis (MD&A), and, similar to Cohen et al. (2020), on the Risk Factor (RF) Discussions subsections of the 10-Q and 10-K files.

Since the MD&A section is unaudited, management has the most discretion in terms of creating its content. Typically, this section provides commentary on financial statements, controls, compliance with laws and regulations, financial activities, actions that have been planned or taken to address any challenges the company is facing. Importantly, in this section management also discusses the firm's outlook by analyzing industry trends, competitive environment, economic conditions, and risks in the financial market.

The Risk Factors section is where a company outlines the potential risks that could negatively impact its business, operations, financial condition, or stock price. According to Regulation SK (Item 305(c), SEC 2005), firms are legally obliged to disclose "the most significant factors that make the company speculative or risky". The typical risk factors discussions include local economic, financial, and political conditions, government regulation, business licensing or certification requirements, limitations on the repatriation and investment of funds and foreign currency exchange restrictions, varying payable and longer receivable cycles and the resulting negative impact on cash flows. Companies may get sued if they do not warn investors about potential risk. Therefore, it is in their interests to include all risk discussions that could remotely or immediately be relevant. The sufficient text data covering RF section begin after 2005.

We identify the textual content of the subsections by capturing regular expressions that contain the word "item" and the subsection name (e.g., "Item 2. Management's Discussion and Analysis" or "Item 1A. Risk Factor"). The subsection titles are very inconsistent across filings, and we make sure the regular expression is flexible enough to capture all possible occurrences of these two subsections.[2] Similar to Loughran and McDonald (2011), we require at least 250 words to appear in the MD&A section, because in many cases this information is "incorporated by reference" (typically deferring to the shareholders annual report). We eliminate the observations if neither RF or MD&A are available in the filing.

We obtain monthly stock returns from the Center for Research in Security Prices (CRSP). We apply the usual filter on individual stocks and require stocks to be common shares listed on the NYSE, Amex, or NASDAQ. We adjust for delisting returns and eliminate low-priced firms with stock prices smaller than \$5.

We calculate the earnings surprise using data from the Institutional Brokers Estimate System (I/B/E/S). In particular, we obtain quarterly analysts' forecasts and actual earnings

[2]Extracting the relevant subsections is non trivial, as their appearances are highly inconsistent across the filings. Please refer to Appendix IA3 for more details.

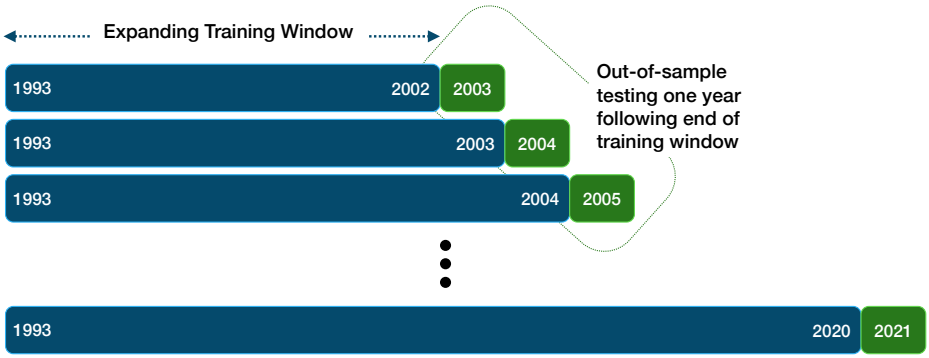

Figure 2: Illustration of the sequential validation procedure, wherein we repeatedly simulate an out-of-sample evaluation for one year following the end of the training period, recursively expanding our training set by one year from an initial ending in 2003 until 2021.

from the IBES unadjusted files from 1993 to 2021. I/B/E/S collects the forecast data at different dates for the upcoming quarter. To obtain the most up-to-date estimate prior to the earnings release, we rely on the consensus forecast generated in the final month of the fiscal quarter for which the earnings are being projected. We define Standardized Earnings Surprises (SUE) based on analysts' forecasts as actual earnings per share minus the average of analysts' forecasts, divided by share price 20 days prior to the earnings announcement:

$$SUE_{i,t} = \frac{E_{i,t} - E_{i,t}^{\text{Analyst}}}{P_{i,t}}, \tag{1}$$

where $E_{i,t}$ is actual quarterly earnings per share for firm i announced at month t, $E_{i,t}^{\text{Analyst}}$ is the corresponding mean analyst forecast, and $P_{i,t}$ is the share price 20 days prior to the earnings announcement.

For each month t, we collect all firms with an eligible earnings surprise $SUE_{i,t}$ and rank these firms from the lowest to the highest based on their $SUE_{i,t}$. We normalize the ranks by the number of firms in month t to create a normalized ranking score $r_{i,t} \in [0, 1]$, where the firm with the lowest (highest) $SUE_{i,t}$ at time t gets the score of 0 (1). We use the normalized ranking score as the target variable.

The I/B/E/S, CRSP, and EDGAR data come with different stock identifiers. First, we match the IBES ticker, the I/B/E/S identifier, with the permanent identification number ("permno"), the stock identifier in CRSP, using the IBES-CRSP Linking Table provided by the Wharton Research Data Services (WRDS). Second, we match the SEC-assigned Central Index Key (CIK), the 10-K and 10-Q filing identifier, with permno using the CRSP linking table.

For training large language models, LLMs, and the baseline machine learning models, we

use the following approach to the historical sample. The first 10 years of the sample, from 01/1993 to 12/2002 is the initial training sample. The last 6 months of the training sample are always retained for the validation. Therefore, for the very first training sample, we have 9.5 years of the actual training and the last half a year is the validation sample. We then make the first four sequential quarterly predictions for one year out, i.e. out-of-sample, for the year 2003. That is we keep the model parameters constant for 2003, i.e. we retrain the model only once a year, while quarterly predictions change from one quarter to another once information set is updated. This procedure is similar to those use by (Gu et al., 2020). After that we evaluate the models' performance results for 2003. Then we add to the training sample all available new data for the year 2003, retrain the model while keeping the last 6 months of 2003 for the validation, and make predictions for 2004. After that we add the data for 2004 to the training and similarly continue onwards till 2021. Figure 2 graphically presents this expanding training window procedure.

3 NLP Methodologies

3.1 Lexicon (Sentiment-score) approach

Following Loughran and McDonald (2011), we parse the text document by employing standard natural language processing techniques. First, all words in the document are converted to lowercase. Second, contractions, such as "haven't," are expanded to "have not." Third, document is stripped of numbers, punctuation marks, and special symbols. Fourth, we remove all the "stop words" which do not carry much meaning on their own, including articles ("the", "a"), conjunctions ("and", "or"), prepositions ("in", "on"), etc. Fifth, we employ the "lemmatization" procedure from WordNet, which involves replacing words with their root form by conducting a detailed morphological analysis, for example, the lemma of "was" is "be" and the lemma of "mice" is "mouse". Furthermore, lemmatization looks beyond genetic word conversion and considers the context of the words. For instance, the lemma of "meeting" might be "meet" or "meeting", depending on its actual use in a sentence. These processing techniques aim to retain meaningful and relevant information while reducing the dimensionality of the inputs for bag of words approaches.

Subsequently, we break the document into a list of word-like building blocks, known as tokens. We use tokens of single words and pairs of words, i.e., unigrams and bigrams. These tokens are converted into a vector of word counts.

To measure each document's sentiment, we use the well-known Loughran and McDonald (2011) finance context dictionary. Words featured in the Fin-Neg word list are classified as

negative sentiment words. The negative sentiment measure, *LM negative sentiment*, of each document equals the sum of negative words, divided by the total number of words in the document.

Unlike a simple word count, the BERT model can provide an identification for whole sentences that takes into account the meaning, order, and interactions of words. Similar to Cao et al. (2023), we use FinBERT (Huang et al., 2022), a version of BERT trained with financial disclosure data (including 10-K, conference call transcripts, and analyst reports), to classify the sentiment of individual sentences to be positive or negative. We thus construct the FinBERT negative sentiment measure as the ratio of the number of FinBERT-negative sentences to the total number of sentences in the document.

Our final measure in this group is motivated by the literature arguing that the increasing length and complexity of the reports is not necessarily accompanied by the dominance of positive informational content, but rather the opposite – to obfuscate and dilute the negative news (Li (2008), Loughran and McDonald (2014), Cohen et al. (2020)). Further, Loughran and McDonald (2014) argue that conventional measures of readability, like Fox-index for example, are not well suited for financial documents. Instead the authors propose using the file size of the 10-K complete submission text file as a readability measure. As we only use the information from MD&A and Risk Factor sections, we define the measure of size in our setting as the total number of characters (letters) in these sections. We thus introduce two new measures: *MD&A length* and *RF length*.

3.2 Bag of words approach

Following Gu et al. (2020), we consider a variety of linear and nonlinear machine learning methods as our baseline models. In its most general form, we describe the normalized ranking score of firm i at time t, $r_{i,t}$, as:

$$r_{i,t} = E_{t-3}[r_{i,t}] + \epsilon_{i,t}, \tag{2}$$

where

$$E_{t-3}[r_{i,t}] = g(\mathbf{z}_{i,t-3}), \tag{3}$$

is the time $t-3$ expected normalized ranking score and $g(\cdot)$ is a flexible function of firm i's P-dimensional vector of word counts, i.e., $\mathbf{z}_{i,t-3} = (z_{i,1,t-3}, \ldots, z_{i,P,t-3})$, corresponding to the document's unigram and bigram terms. To maintain relevance, only the 7,000 most frequent

unigram and bigram terms from the training sample are retained in $\mathbf{z}_{i,t-3}$ (i.e., $P = 7000$), while common terms appearing in more than 99.9% of documents are eliminated.

Linear methods We include ordinary least squares (OLS) with all covariates as a widely-used benchmark. OLS is simple, but can overfit on high-dimensional textual data, prompting us to adopt two regularization techniques to address this concern. First, inspired by Jegadeesh and Wu (2013), we apply OLS with the LM negative sentiment score as the sole covariate, which can be considered as a form of dimension reduction. Second, we employ penalized linear models, Lasso and Elastic Net (EN), which impose sparsity by setting certain coefficients to exactly zero.

Nonlinear methods To model nonlinearities, we consider tree-based methods, including random forest (RF) and XGBoost (XGB). We also include kernel-based method, Support Vector Regression (SVR), which has been shown to perform well for high-dimensional textual data (Manela and Moreira, 2017). Furthermore, we include a traditional feed-forward neural network (NN) as a straight-forward nonlinear benchmark. Section IA1 in the Internet Appendix provides a detailed description of the baseline machine learning methods, and Section IA2 in the Internet Appendix details the hyperparamter tuning procedure for these methods.

3.3 Large Language Models and Transformer Approach to Process Text of Any Length

LLMs like BERT, RoBERTa, or T5 can process input sequences of up to 512 tokens, translating these tokens into a 1024-dimensional vector representation. In contrast, OPT can manage sequences as long as 1,024 tokens and embed each token into 2,560-dimensional space. Even that would not be enough to process full text of MD&A or RF sections which easily exceeds 50K tokens.

The GPT-3 model (which powers ChatGPT) has a maximum token limit of 4096 tokens per input sequence. This means that any input text longer than 4096 tokens would need to be split into multiple segments or processed in chunks to be handled by the model effectively. Even then, GPT-3, like other language models, is a general-purpose model trained on diverse text from the internet. While GPT-3 can generate text related to various topics, including finance, it is not specifically trained to predict financial performance. Yet, researchers can fine-tune or adapt GPT-3 using domain-specific data to perform tasks related to financial analysis or prediction, such as sentiment analysis, text generation for financial reports, or

language-based financial modeling. However, training and fine-tuning a model as large as GPT-3, which has 175 billion parameters, is an extremely computationally intensive task. The original training of GPT-3 by OpenAI required substantial resources, including specialized hardware like GPUs (Graphics Processing Units) and TPUs (Tensor Processing Units). The exact number of GPUs or TPUs needed for training GPT-3 was not publicly disclosed by OpenAI.

In contrast, BERT-Large that we use in this paper, has 336 million parameters which makes it computationally feasible to train and fine tune. Similar to ChaGPT which would require splitting the text into chunks to handle long documents, we split (MD&A) and risk factors (RF) in chunks of 512 tokens to be handled consecutively and which we accomplish with Transformer architecture.

The Transformer (Vaswani et al., 2017; Lin et al., 2022) is a machine learning architecture that has become the *de facto* standard across natural language processing (NLP) tasks, such as language translation and text classification (Yvon, 2023). One of the key advantages of the Transformer is its ability to handle variable-sized sequence inputs, such as language, without the need for recurrent neural networks (RNNs) or convolutional neural networks (CNNs). This is achieved through the use of a self-attention mechanism, which allows the model to selectively attend to different parts of the input sequence. This makes the Transformer able to capture relatively long-range dependencies between elements in the input sequence, something difficult to achieve with RNNs and CNNs. Separately, the Transformer has been shown to have excellent scaling properties for language, wherein so-called *scaling laws* describe how the model's performance improves as the number of parameters, training data, and available computational resources respectively increase (Kaplan et al., 2020; Bahri et al., 2021).

However, one of the main limitations of the Transformer is its $O(N^2)$ computational scaling in the input size, which makes it unable to directly process very long sequences. This is because the self-attention mechanism requires pairwise comparisons between all elements in the input sequence, resulting in a quadratic increase in computation time as the sequence length increases. As a result, the Transformer is typically used for tasks that involve sequence lengths in the hundreds—or at best thousands—of elements.

In this paper, we propose a recursive application of the Transformer architecture to obtain a global understanding of the relevant sections of interest, obviating the input context limitations of off-the-shelf pretrained Transformers. We illustrate our approach in Figure 3. At the core, the proposed model makes repeated use of a pretrained BERT encoder-only large language model (Devlin et al., 2018), which we fine-tune on our 10-X corpus.[3]At a high-level,

[3]We make use of the `bert-large-uncased-whole-word-masking` variant, available at https://huggingface.co/bert-large-uncased-whole-word-masking. This model consists of 24 Transformer lay-

from the bottom-up, the model processes input reports as follows, which we detail in the sections below:

1. With the BERT tokenizer, we convert the text of the MD&A and RF sections of a report into tokens. We split this tokenized representation into groups of 511 tokens, and prepend each group with the special BERT `[CLS]` token, yielding chunks of 512 tokens, which is the BERT input context limit.

2. We feed each chunk into the pretrained BERT encoder-only large language model, and extract the output of the `[CLS]` token, which we refer to as the *chunk embedding*, with chunk k denoted $\texttt{[CLS]}_k$ in the figure.[4]

3. We coalesce the `[CLS]` tokens from all chunks into a single Transformer layer, with a hidden dimensionality of 64 and a single attention head.

4. These tokens are then reduced into a single vector of dimensionality 1024 using an average-pooling operation.

5. This vector is then passed to a linear predictor which outputs two parameters $\hat{\alpha}$ and $\hat{\beta}$ of a Beta distribution, which we use to model the predicted normalized rank $r_{i,t}$ of the report. The Beta distribution is a good choice for this task, since it has support between 0 and 1.

3.3.1 Input Encoding

The MD&A and RF sections are first divided using the Sentecizer from Spacy,[5] a popular library for segmenting a document into meaningful sentences. These sentences are then concatenated together with a `[SEP]` token in between and then tokenized using the BERT tokenizer; token embeddings have dimensionality 1024. The resulting tokenized text is then split into chunks of 511+1 tokens, as explained above. We assume that there is a total of n chunks. Should the last chunk count fewer than 511 tokens, it is padded with `[PAD]` tokens.

ers, an embedding dimensionality of 1024, and 16 attention heads. It counts 336M parameters. We use the `transformers` library (Wolf et al., 2020) to load the pretrained model and fine-tune it on our 10-X corpus.

[4] In BERT, the `[CLS]` token, which stands for "classification," has a specific purpose. Typically placed at the beginning of the input text, its main purpose is to act as an aggregate representation for downstream prediction tasks. After the BERT model has processed the input data, the final hidden state (the output of the transformer) corresponding to this `[CLS]` token is used as the overall chunk embedding. In our case, this token represents a semantic embedding of a particular chunk of text, and its meaning is refined during the fine-tuning process, as explained later.

[5] https://spacy.io/api/sentencizer

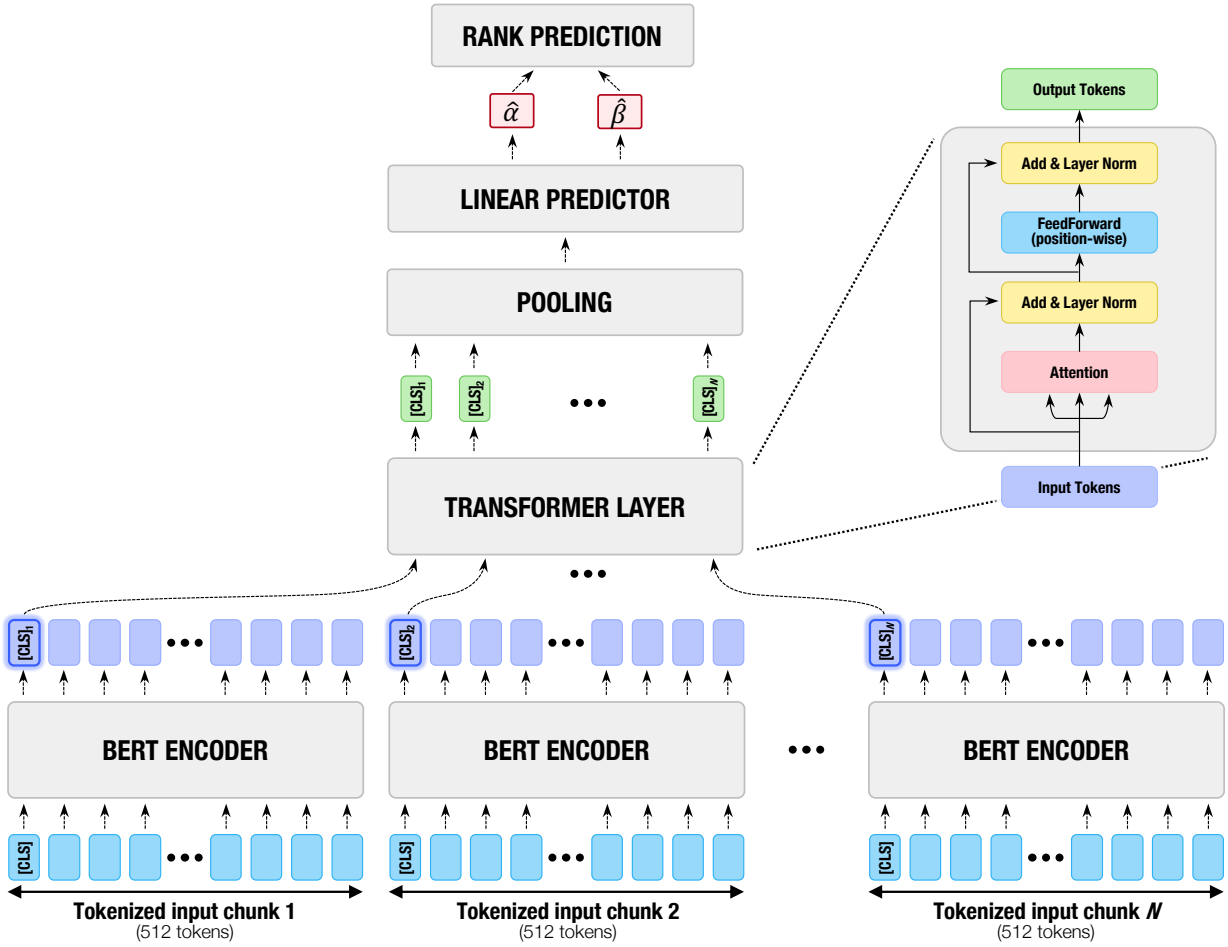

Figure 3: Hierarchical Transformer Architecture.

3.3.2 BERT Layer

We then pass each chunk through a pre-trained BERT encoder, of which we study two versions: (i) a first version uses a pre-trained BERT model which we never fine-tune, called the FrozenBERT variant of our model, and (ii) a second version fine-tunes BERT using a curriculum learning procedure described below, called the FtBERT variant of our model. At the encoder output for each chunk, we extract the initial `[CLS]` token, which we refer to as the *chunk embedding*, with the output for chunk k denoted $\texttt{[CLS]}_k$ in fig. 3.

3.3.3 Transformer Layer

These $\texttt{[CLS]}_k$ vectors are brought together into a Transformer layer (denoted as such in fig. 3). This consists in a single layer of the Transformer architecture with a single attention head, which mathematically performs the following computations.

Let $\mathbf{X}_k$ be the k-th chunk of tokenized MD&A or RF section, and $\mathbf{h}_k \equiv \texttt{[CLS]}_k$ be the output of the BERT encoder for $\mathbf{X}_k$. Let $\mathbf{H} = \text{Concat}(\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n)$ be the columnwise concatenation of all $\mathbf{h}$'s, where n is the number of chunks. Then, the output of the Transformer layer, can be expressed as,

$$\mathbf{H}^{\text{out}} = \text{Transformer}(\mathbf{H}),$$

consisting of the following sequence of steps:

$$\begin{aligned}
\mathbf{H}' &= \text{Attention}(\mathbf{H}\mathbf{W}^Q, \mathbf{H}\mathbf{W}^K, \mathbf{H}\mathbf{W}^V)\mathbf{W}^O, \\
\mathbf{H}'' &= \text{LayerNorm}(\mathbf{H} + \mathbf{H}'), \\
\mathbf{H}''' &= \text{FeedForward}(\mathbf{H}''), \\
\mathbf{H}^{\text{out}} &= \text{LayerNorm}(\mathbf{H}'' + \mathbf{H}'''),
\end{aligned} \tag{4}$$

where we respectively define Attention to be our (self-)attention mechanism, FeedForward to be a two-layer feedforward network with ReLU activations with 64 hidden units, and LayerNorm to be a layer normalization operation (Ba et al., 2016):

$$\begin{aligned}
\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) &= \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{D_k}}\right)\mathbf{V}, \\
\text{FeedForward}(\mathbf{X}) &= \text{ReLU}(\mathbf{X}\mathbf{W}^1 + \mathbf{b}^1)\mathbf{W}^2 + \mathbf{b}^2, \\
\text{LayerNorm}(\mathbf{x}) &= \frac{\mathbf{x} - \text{E}[\mathbf{x}]}{\sqrt{\text{Var}[\mathbf{x}] + \epsilon}} \odot \gamma + \beta, \\
\text{softmax}(\mathbf{x})_i &= \frac{\exp(\mathbf{x}_i)}{\sum_j \exp(\mathbf{x}_j)}.
\end{aligned} \tag{5}$$

with $\mathbf{Q} \in \mathbb{R}^{n \times D_k}$, $\mathbf{K} \in \mathbb{R}^{n \times D_k}$ and $\mathbf{V} \in \mathbb{R}^{n \times D_v}$ respectively representing the queries, keys and values, and where D_k and D_v are respectively the dimensionality of the keys and values (both are 1024 in our BERT set-up). The $\mathbf{W}$'s are trainable parameter matrices, the $\mathbf{b}$'s are trainable bias vectors for the feed-forward layer, and $\gamma, \lambda \in \mathbb{R}^+$ are trainable parameters for the affine transform in layer normalization; ϵ is a small constant to avoid division by zero.

3.3.4 Pooling Layer

Next, a pooling layer averages all tokens produced by the Transformer layer in order to produce a single vector, which captures the most salient features of the MD&A and RF sections. With $\mathbf{H}^{\text{out}} = \text{Concat}(\mathbf{h}_1^{\text{out}}, \ldots, \mathbf{h}_n^{\text{out}})$ denoting the output of the Transformer layer, the pooling operation is computed as

$$\mathbf{p} = \frac{1}{n} \sum_{k=1}^{n} \mathbf{h}_k^{\text{out}}.$$

3.3.5 Linear Predictor

We then use a linear predictor to map $\mathbf{p}$ to the alpha and beta parameters of a predictive Beta distribution; since the support of the Beta is strictly between zero and one, this is a natural representation for $r_{i,t}$, the normalized rank as outlined in section 2,

$$\mathbf{z} = \mathbf{W}^{\text{lin}} \mathbf{p} + \mathbf{b}^{\text{lin}}, \tag{6}$$

with $\mathbf{W}^{\text{lin}}$ and $\mathbf{b}^{\text{lin}}$ being trainable parameters, from which we obtain the predicted parameters $\hat{\alpha}$ and $\hat{\beta}$ of the Beta distribution as

$$\begin{aligned} \hat{\alpha} &= \text{softplus}(\mathbf{z}_1), \\ \hat{\beta} &= \text{softplus}(\mathbf{z}_2), \end{aligned} \tag{7}$$

where $\text{softplus}(x) \equiv \log(1 + \exp x)$ is used to ensure that $\hat{\alpha}$ and $\hat{\beta}$ are positive, and $\mathbf{z}_k$ denotes the k-th element of $\mathbf{z}$. We obtain the expected predicted normalized rank $\hat{r}_{i,t}$ of the company as

$$\text{E}\left[\hat{r}_{i,t}\right] = \frac{\hat{\alpha}}{\hat{\alpha} + \hat{\beta}}.$$

3.3.6 Prediction Target and Loss Function

We train the model to predict the normalized rank $r_{i,t}$ of company i at time t by minimizing the negative log-likelihood (NLL) of the predictive Beta distribution as our loss function:

$$\mathcal{L} = -\sum_{i,t} \log\left(\frac{\Gamma(\hat{\alpha}_{i,t} + \hat{\beta}_{i,t})}{\Gamma(\hat{\alpha}_{i,t})\Gamma(\hat{\beta}_{i,t})}\right) + (\hat{\alpha}_{i,t} - 1)\log r_{i,t} + (\hat{\beta}_{i,t} - 1)\log(1 - r_{i,t}), \tag{8}$$

where the summation is taken over all companies i and all time periods t in the training set, with $\hat{\alpha}_{i,t}$ and $\hat{\beta}_{i,t}$ obtained from eq. (7), and $\Gamma(\cdot)$ is the gamma function. The NLL is minimized by stochastic gradient descent on all trainable parameters of the model using the Lion (EvoLved Sign Momentum) optimizer (Chen et al., 2023).

We recursively alternate model training and out-of-sample evaluation following the procedure illustrated in fig. 2. The distinctions between FrozenBERT and FtBERT are as follows:

- **FrozenBERT**: we minimize the training loss (8) by optimizing all trainable parameters listed in eqs. (4) to (6). In this version, we solely rely on the pretrained BERT model, and we leave the BERT parameters untouched during training.
- **FtBERT**: for the first five training epochs, we train exactly as with FrozenBERT, modifying only the parameters listed in eqs. (4) to (6). After these five epochs, we "unfreeze" the pretrained BERT model and add its parameters to trainable parameters in addition to the previously-listed ones. The same loss function is used. Following common fine-tuning practice, we use two different learning rates for this purpose: the BERT model is finetuned with a learning rate of 3×10^{-8}, whereas the Transformer layer is trained with a learning rate of 3×10^{-6}; we found that these rates allowed for stable training.

These hyperparameter values were determined using a validation set for which we also performed early stopping. For each year that was selected as the test set, the validation set consisted of 6 months of company reports that precedes the test set. While the maximum number of training epochs is 200, we save and use the model that performed best on the validation set in terms of mean squared error (MSE). The number of training epochs might not reach 200 as we stop the training when MSE does not improve over 25 consecutive epochs (also known as patience).

4 Empirical Analysis: Models' Performance Comparison

We group all models in three categories. #1. Lexicon- (keyword) based sentiment identification: LM negative sentiment, FinBERT classified fraction of negative sentences, MD&A length, and RF section length. #2. Bag-of-words models relying on the following regression approaches: OLS; LM negative sentiment score with OLS weights (similar to Jegadeesh and Wu (2013)), LM OLS; EN; Lasso; and SVR (similar to Manela and Moreira (2017)). #3 LLMs trained on financial objectives: FrozenBERT and FtBERT.

We compare the models prediction performances on several dimensions. Kelly et al. (2023) argue that when it comes, for example, to stock return predictability, due to the variance of the forecast, which can be quite high in financial data, the traditional statistical evaluation techniques such as out-of-sample (OOS) R^2 or MSE are not very informative. When the variance of the forecast is large, OOS R^2 can easily be negative, while the Sharpe ratio of the strategy that goes long in the stocks that are predicted to have the highest returns, and short-sells those with the lowest predicted returns can achieve quite high positive economic magnitudes. In other words, the performance of finance-targets' predictability should be evaluated based on the performance of certain investment strategies. Unexpected earnings surprises are as challenging to predict due to their high variability as stock returns.

Following Cohen et al. (2020), we adapt portfolio management strategies where we rank stocks based on their earnings surprise forecasts into quintiles, and then we evaluate the OOS performance of the High-minus-Low strategy that buys the highest and sells the lowest earnings forecasts' quintiles. We evaluate the performance of both equal- and value-weighted portfolio returns, as unexpected earning positive or negative surprises should follow by higher or lower stock returns respectively.

For those measures that pass the portfolio performance tests, we further verify the robustness of the results in the stock returns predictive cross-sectional and time series regressions with the time and firm fixed effects, and adding various firm characteristics.

The final robustness test is conducted via event study panel regressions to predict the earnings surprise itself, or cumulative abnormal returns five days after the announcement.

Table 1 presents diagnostic statistics, MSE and OOS R^2 for groups 2 and 3 which are based on regression approach. MSE's might not seem very large as we predict the rank of earnings surprises with all cross-sectional observations ranked within the interval [0,1], and hence with the cross-sectional mean of 0.5. Even then, the OLS approach stands out with the highest MSE which is more than the mean. As per OOS R^2, OLS has very high negative value which indicates overfitting to the the training sample. Other methods with negative OOS R^2 are EN, SVR, and NN. Note, that as long as OOS R^2 is not too negative (Kelly

et al., 2023), we cannot reject the model based on these statistics. All other models have positive *OOS* R^2 among which RF and Xgboost have the highest values, 2.7% and 1.4% respectively.

4.1 Portfolio sorts

Here we estimate the performance of stock portfolios sorted on the predicted earnings surprises for the forthcoming next quarter. Each month, we collect all the stocks which have public releases of 10-Q or 10-K reports. We then rank them based on the predicted next-quarter earnings surprise into quintile portfolios at the end of the month, with the highest (lowest) portfolio containing stocks with the highest (lowest) predicted next-quarter earnings surprises. Following Cohen et al. (2020), once placed in a quintile portfolio, a stock will be held in this portfolio for three months until a new 10-Q or 10-K is released. The portfolios are however rebalanced monthly as different companies file their reports in different months.

Table 2 reports equally-weighted, EW, and value-weighted, VW, portfolio performance for the measures in group 1. For each portfolio we report raw excess returns, one factor, CAPM, alpha, as well as risk-adjusted returns with Fama-French five factors, FF5, and six factors, FF6, including momentum. The Newey and West (1986) 3-lag adjusted t-statistics are reported in parentheses. Further, similar to Cohen et al. (2020), we also report the performance of High-minus-Low strategy, which goes long high quintile, Q5, portfolio and short-sells low, Q1, quintile portfolio. The *OOS* performance of the investment strategy based on the forecasts is the most suitable evaluation of the precision of forecasts in the context of financial data (Kelly et al., 2023).

Panel A of Table 2 reports portfolio performance based on LM 2011 negative sentiment score. The High-minus-Low, H–L, portfolio has very small and statistically insignificant from zero EW or VW returns. This suggests that the current negative lexicon-based sentiment does not have strong predictive effect for the future performance. A similar evidence is observed for the negative sentence-based FinBERT classification and EW portfolios, Panel B. In the VW portfolios and FinBERT classification, the results are opposite from the expectations. Here, the portfolio with the highest negativity score, Q5, has the highest returns, i.e. a wrong sign, and so is H–L strategy with the positive, 31 bps per month, raw return. The CAPM alpha is positive as well, 25 bps per month ($t = 1.76$). It does not change even after FF6 factor adjustment, 26 bps per month ($t = 1.83$). Therefore in VW portfolios, FinBERT classification performs the worst as it gives opposite from the negative sentiment performance prediction.

Panel C reports portfolio sorting results for MD&A length. Here, the portfolio of firms with

the longest length has the lowest returns, for both EW and VW portfolios, and H–L portfolio has -0.189% ($t = 1.71$) and -0.264% ($t = 1.92$) returns respectively. After adjustment for the market movements, CAPM alphas of these portfolios are -0.23% ($t = 2.05$) and -0.344% ($t = 2.48$) respectively. Annualized, it translates into 3.8% and 4.13% underperformance respectively of companies with the highest MD&A length. This results is quite intriguing as a simple measure like this outperforms more sophisticated sentiment measures, such as FinBERT for instance. Moreover, a portfolio with the shortest length, Q1, has a positive and statistically significant alpha. Thus, unlike other measures in the literature which mostly succeed to capture negative content (Loughran and McDonald (2011), Cohen et al. (2020)), this measure is able to identify positive information. The positive alphas for Q1 remain statistically significant even after FF6 factors' adjustment. It is 37 bps per month ($t = 5.88$), or 4.4% annual benchmark out-performance in EW portfolios, or 19 bps ($t = 2.13$), or 2.3% annually in VW portfolios. Further, the portfolios' alphas are almost monotonically decreasing from Q1 to Q5 quintiles for both EW and VW portfolios. The H–L strategy, however, becomes statistically insignificant after FF5 or FF6 risk-adjustment. Therefore, this measure is not particularly robust in identifying negative information.

Finally, Panel D reports the performance of portfolios sorted on the length of RF section. We find no evidence of future performance difference between high and low RF length portfolios, and H–L spreads are mostly insignificant, except for EW portfolios and FF5 and FF6 factor adjustments. Here the raw excess returns and CAPM alphas are insignificant. However, FF5 and FF6 alphas suddenly become higher in magnitude than CAPM alphas alone and so do H–L portfolio alphas. The reason is that the betas of value, profitability and investment FF5 factors are negative for these portfolios which explains an increase of alphas compared to CAPM adjustment. Note that FF6 factors are intended to explain the premiums of the related five to six characteristic-sorted portfolios, and not necessarily the right fit for the document length-sorted portfolio premium characterization. Further, this result is different from those in (Cohen et al., 2020) who present the time series evidence related to changes in the format of RF section from one report to another, while we present the cross-sectional evidence related to the total length of the report.

In summary, the best performer of this group is a simple measure, the length of MD&A section, and mostly in identifying positive information associated with the shortest length.

Table 3 reports similar quintile portfolio performance returns for group 2, the bag-of-words based approaches. The main criteria in identifying the economic and statistical significance is the consistency of performance among EW and VW portfolios, and significance of risk-adjusted returns. None of the reported models in this table, group 2 models, passes the consistency and robustness to factor risk-adjustment tests.

For example, XGBoost (Panel F), Random forest (Panel G), and Neural network (Panel H) generate significant H–L excess returns for EW portfolios. However, the excess returns generated by XGBoost in the EW portfolio do not survive a simple, one-factor market adjustment, implying that it largely captures market trends. Similarly, the risk-adjusted returns for Random forest and Neural network are only marginally significant after accounting for major risk factors. Furthermore, the raw excess returns of all these models become insignificant in value-weighted portfolios, suggesting that the excess returns in the EW portfolios may be driven by small firms and are not robust across the sample. We conclude that the bag-of-words approach fails to identify future positive or negative earnings surprises.

Finally, Table 4 reports portfolio performance results for LLMs. Unlike all previous tables, here all H–L portfolio strategies have economically and statistically significant values, and the correct signs.

Consider Panel A, the portfolio sorts based on FrozenBERT predictions. Raw excess returns and risk-adjusted portfolio alphas are almost monotonically increasing from Low to High quintile portfolios. For VW portfolios, H–L raw excess return is 43 bps per month ($t = 2.51$), or 5.16% per year. After one factor, CAPM, risk adjustment it drops to 37 bps per month ($t = 2.12$) or 4.44% per year. Finally, after FF6 factor adjustment, it becomes 32 bps per month ($t = 2.01$) or 3.8% per year – still an economically high number. Most of these abnormal returns are driven by High/Q5 quintile, the long position. The results are very similar for EW portfolios.

Panel B presents even more consistent and economically meaningful results for portfolios based on FtBERT predictions. Here, raw excess returns and risk-adjusted alphas increase strictly monotonically with the portfolio quintile. The results are economically higher for VW portfolios. H–L raw excess return is 56 bps per month ($t = 2.94$), or 6.74% per year. This number almost does not change after CAPM risk adjustment, 50 bps per month ($t = 2.57$), or 6.01% per year. This is impressive as none of FtBERT identification is related to the general market movement. Further risk adjustment leads to the lower H–L abnormal performance, with 33 bps per month ($t = 1.88$), or 4% per year for FF5 factor adjustment, or 31 bps per month ($t = 1.77$), or 3.71% per year for FF6 factor risk-adjustment. Similar to the previous discussions, FF5 or FF6 are not the best benchmark models for risk adjustment for these portfolio sorts. For example, H–L spread decreases with FF5 factor adjustment largely because the alphas of Low quintile portfolio, the short position, start increasing almost 10 times compared to respective CAPM alpha in economic magnitudes. For example, the un-adjusted raw excess return of Low value-weighted quintile, L, is 77 bps per month ($t = 2.67$). After CAPM risk adjsutment it drops to -0.106% per month and insignificant at the conventional levels. However, after FF5 adjustment it increases to -0.0002%, or to the

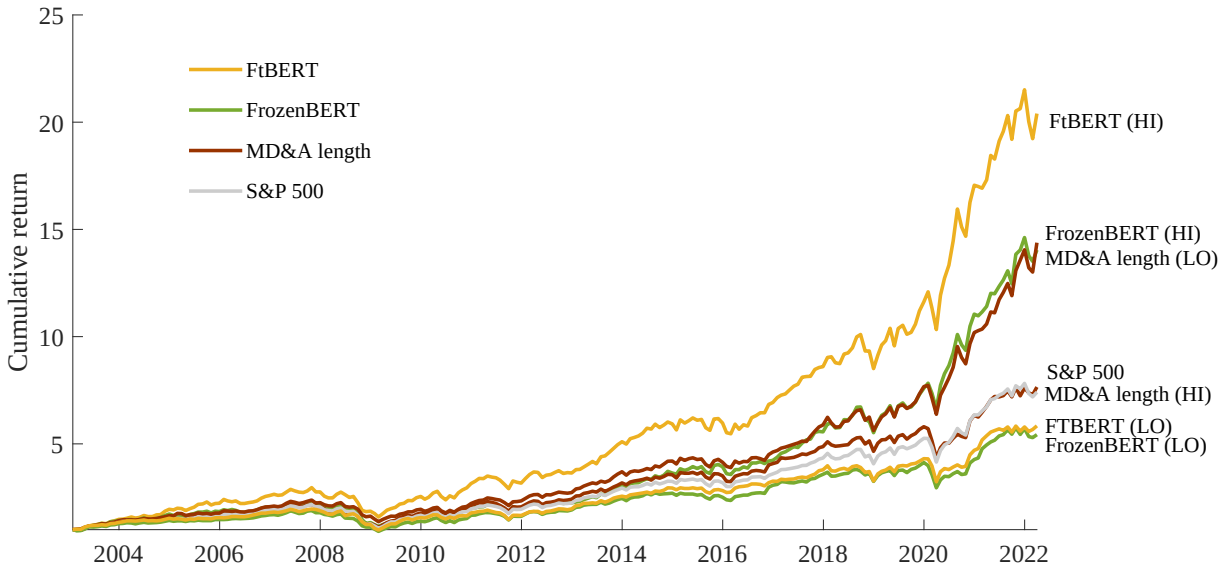

Figure 4: *OOS* Cumulative High versus Low Portfolios Value-weighted Returns, and S&P 500

positive 0.008% after FF6 adjustment.

This is even more pronounced for EW portfolios, where from statistically insignificant CAPM portfolio alpha for the Low quintile, 0.057% per month, this number jumps to 0.22% per month and becomes significant ($t = 2.30$) after FF5, or to 0.234% per month ($t = 2.45$) after FF6 factor adjustment. It happens due to the negative loadings on the FF5/6 factors, except the market factor. Even though all H–L abnormal reruns still remain significant at the conventional statistical level, we suggest that a simple CAPM alpha is a better measure to gauge an abnormal performance of these portfolios.

Overall, the only clear and consistent winner in the *OOS* portfolio strategy assessment, which is the main accuracy test in financial data (Kelly et al. (2023)), is group 3, LLMs. In this group, FtBERT marginally outperforms FrozenBERT in economic magnitudes. Group 2 fails the proper identification on all dimensions, and the only candidate from group 1 which marginally passes most of criteria is MD&A section length.

4.1.1 Dynamic Inter-temporal Portfolio Performance

How do these portfolios perform over time? Figure 4 presents cumulative, based on $1 dollar initial investment in the beginning of 01/2003, *OOS* portfolio performance of High/Q5 vs. Low/Q1 quintiles of FtBERT, FrozenBERT and MD&A length classifications, and further compares it to S&P 500 cumulative return (including distributions) for the similar period as a benchmark.

The High portfolio identified by FtBERT is the top performer, where the initial $1 dollar investment appreciates more than 20 times during *OOS* period from 01/2003 to 03/2022. The High portfolio identified by FrozenBERT and the Low portfolio identified by the low MD&A length quintile trace each other. That is a sophisticated, pre-trained LLM, before fine-tuning, does very similar job in identifying positive information as a very simple measure of characters' count. FrozenBERT was trained on the text corpus of all of the internet before 2018, and even it had never seen EDGAR filings per se, financial news are by far not the dominant part of the training text. As such, FrozenBERT comes with the "noise" of non-financial text, which performs no differently than a simple length of MD&A section in identifying positive information.

Further, the long length of MD&A section does not necessarily mean bad performance, as this portfolio mostly traces the S&P 500 index. In other words, the length of MD&A section does not help in identifying negative information.

On the contrary, both FtBERT and FrozenBERT are very similar in identifying negative information or under-performing firms, as the Low quintile portfolios of these measures under-perform the market significantly. Therefore, FrozenBERT is very similar to other measures in the literature (e.g., Cohen et al. (2020)) which succeed in identifying negative information. Unlike all other measures in the literature, FtBERT succeeds in identifying both positive and negative information about future performance.

4.2 Predicting Stock Returns: A Linear Regression Approach

In this section we evaluate the robustness of our main results in the panel, time-series and cross-sectional, regressions while controlling for other well-known firm-specific return predictors.

Table 5 presents the first set of results for predicting next month excess firm-returns in a simple univarite regression, Panel A, or adding main firm-characteristics such as size, book-to-market, momentum and return reversals, Panel B. Each panel has the time (month) and firm fixed effects and the errors are clustered on the firm level. The main predictors we evaluate are MD&A and RF length, which we divide by 10^6 to make them on the same scale as stock returns, the fraction of negative sentences identified by FinBERT, as well as FrozenBERT and FtBERT predictive earnings surprises ranks.

In the univariate regressions, Panel A, similar to the portfolio sorts, FinBERT has positive, rather than expected negative, coefficient and marginally significant. MD&A length is not significant after controlling for the firm and time effects. FrozenBERT and FtBERT have positive and highly significant coefficients. Their coefficients remain very similar and highly

statistically significant after adding other firm-specific return predictors in Panel B.

After adding size, BM, momentum and reversals control variables, the coefficiet of FinBERT becomes negative, as expected, but statistically insignificant from zero.

Table 6 presents similar results as in Table 5, with extra control variables, including the most recent earnings surprise itself, SUE_t, accruals, cash flows, profitability, and investment. Interestingly, after adding these many controls, SUE_t has no predictability for the future returns, while the coefficients and their significance of FrozenBERT and FtBERT remain almost unchanged compared to Panel B, Table 5. Thus the predictions based on these LLM measures are the most robust overall so far.

4.3 Event Study Regressions: Earnings Announcements

FrozenBERT and FtBERT were explicitly trained to predict unexpected earnings surprises. All the tests above are focused on monthly returns predictability with the implicit assumption that positive (negative) earnings surprise will lead to subsequent positive (negative) returns. Here we focus on the earnings announcements themselves, and the post-announcement 5-day cumulative abnormal returns.

The underlying question is that, given that the current 10-Q, 10-K release is informative about future earnings, and this information is public, we should see some market reactions to the current releases. For example, Huang et al. (2020) argue that institutions react quickly to the most recent news and their trades contribute to price discovery, i.e. predict direction of stock returns, up to one week after the news releases. Thus, if our measures identify the information correctly, they too should predict the direction of price discovery, i.e. post-filing returns, within the event, the filing day, window.

10-Q, 10-K reports are available for free to download from EDGAR website two days after the filing. Paying a special charge, one can obtain them instantly, on the filing day. Therefore, the cumulative abnormal post-filing day returns, CAR[1,5] can be real trading profits based on the fast reaction to the released news (Huang et al., 2020). The final robustness test for our measures is thus to predict this price discovery, and its direction.

We calculate CAR abnormal returns as CAPM-adjusted returns. We use the value-weighted CRSP market index as a proxy for the market portfolio. At the end of every month, we estimate the market beta of each stock by regressing a stock's daily excess returns in the past year on the contemporaneous excess market return as well as five lags of the market return to account for the illiquidity of small stocks (Dimson, 1979). We calculate the market beta as the sum of the six OLS regression coefficients. The expected return is then estimated by multiplying the market excess return over the event period and the

pre-estimated market beta for the month prior to the event. Post-filing cumulative 5-day abnormal return, CAR[1,5], is then calculated as the difference between the 5-day cumulative realized and the corresponding expected returns.

Table 7 presents panel regression results around the filing day for FrozenBERT, Panel A, and FtBERT, Panel B. First, both measures positively and statistically significantly predict future earnings surprises, SUE_{t+3}. This is expected as they are trained to perform these forecasts. Yet, the first column in each panel provides further *OOS* robustness validation.

The final robustness validation is whether these measures properly predict the direction of returns, i.e. the price discovery and price impacts associated with informed institutional trading around these releases (Huang et al., 2020). FrozenBERT fails this test as the coefficients for CAR[1,5] predictability, although with the right positive sign, is insignificant. In contrast, FtBERT is able to predict the direction of institutional trading with positive and statistically significant coefficient for CAR[1,5] predictability. We conclude that only FtBERT can capture information identified by professional institutional investors. However, given our previous results, we also suggest that it takes months after earnings announcements before this information is fully incorporated into the prices. The latter observation is fully consistent with the post-earnings announcements' trends widely reported by the previous literature.

5 What Does FtBERT Pay Attention to?

The winner of the race is FtBERT. It passes all robustness tests. It is able to identify similar information as the professional institutional investors (Huang et al., 2020). Our results point to market inefficiency as while some market participants react to this information instantly, it still takes months before this information is fully reflected in firms' valuations.

What are the sources of the market inattention and what does FtBERT pay attention to to be able to identify important signals? To answer these questions we first revert back to the portfolio sorts reported in the Table 4, Panel B. Recall that these portfolios are formed at the end of the month of 10-Q or 10-K releases. That is, if the earnings announcement is in the middle of the month, for example, we would skip approximately two weeks before we rank this stock in a quintile portfolio. Yet, even with the 2-week or more portfolio formation delay, the profitability of H–L strategies are economically and statistically significant.

Here we first present the characteristics of stocks in each of FtBERT quintile portfolio. Table 8 presents a summary of several stock characteristics on a quintile portfolio level. Notably, the average size of each quintile is above the market average, with the L portfolio having the average size of USD $7.7 bln. The size monotonically increases with FtBERT

quintile portfolio, with the highest quintile average market cap of USD $11 bln.

In contrast, BM is monotonically decreasing with FtBERT quintile portfolio, where the highest quintile has the lowest average BM. Thus, the highest quintile, which produces the highest positive future abnormal returns is also the largest stocks with the most growth embedded options. Interestingly, this is also the quintile with the higest analyst forecast dispersion measured by the standard deviation of analysts' consensus forecasts. This can explain the market inefficiency argument we laid down earlier, as the most profitable quintile is almost the most controversial one, at least from the sell-side analyst point of view.

We can also see that the predictions for the next quarter, SUE_{t+3}, standardized earnings surprises are quite accurate as SUE_{t+3} is monotonically increasing from the most negative to the most positive with FtBERT quintile. There is no monotonicity at all in the current, most recent SUE_t, as Q2 to Q4 quintile portfolios rather experience reversals in their earnings surprises from one quarter to another.

Finally, for each quintile portfolio we estimate the *hit ratio*. It is computed as the fraction of ex-ante earnings surprise quintile assignment, and the ex-post correct validation of this portfolio assignment. For example, if FtBERT predicts at time t that a stock i will have one of the highest earnings surprises at $t+3$, this places this stock in Q5. At $t+3$, when the new earnings announcement is made, and our prediction turns out to be correct, we assign the dummy variable value of 1 to this case-observation, and zero otherwise. Because we have 5 portfolios, the benchmark of a pure random chance for a stock to be assigned to each portfolio is 0.2. Anything better than 0.2 is considered better than chance.

The last row of Table 8 reports *Hit Ratio* for each quintile. As one can see, the extreme quintiles, L and H, have the highest values of 0.226 and 0.242 respectively, and these numbers are statistically significantly different from 0.2, pure chance. Moreover, the accuracy rate of FtBERT is higher where the dispersion of analyst forecasts is the highest, i.e. for quintile H.

We next want to shed light at what FtBERT pays the most attention to. Tables 9 to 12 provide several examples of positive and negative FtBERT future performance identifications. The general format of these tables is as follows. We report *Text Highlights*, or the parts of MD&A or RF sections where FtBERT places the highest weight, i.e. the most attention, for either positive or negative future financial performance identifications. These are snapshots from either MD&A or RF sections with the highest attention weights from FtBERT, within the model's upper Transformer layer (section 3.3.3). While short, these paragraphs are still condensed with information which often carries a similar massage across multiple sentences. To ease the exhibition, in yellow, we highlight the sentences which we ex post manually identify as the most important ones. They reflect the nature of information of at least half of the presented text, and we are able to relate this information to the future, post

announcement news, *Post-Announcement News Sources*, explaining either high or low stock returns. The panel *Explanations* summarizes these news, and the very first, the top panel provides the summary of the announcement itself, and 3 monthly post-announcement stock level returns.

Consider for example Table 9, Zoom Video Communications Inc. In the end of March 2020, FtBERT places it in the highest quintile portfolio to predict next quarter, June 2020 announcement. The prediction is based on March 20, 2020 filing. The forward-looking statements from the management's MD&A were to offer more services to attract more users, and hire more marketing and sales personnel to expand the user base. The management overall is bullish towards investing into expansion among the existing and new customers. Obviously it coincides with the onset of Covid-19, and the demand for these services appeared out of necessity. Yet, it was not immediately obvious, as the stock's return in April is negative, -7%. Only in May does the stock finally jump with a monthly return of 33%. The high June return, 44% is then attributed to the new filing in early June and forecasts of longer-term demand for Zoom services. Further, Zoom still had to compete for their customers over Microsoft Teams services, or to bring the awareness of its services over Skype for example. Ex-post, marketing and sales expenses were well justified.

Another example, Zillow, Table 10, where management is putting emphasis on engaging customers to increase their use of their online platform. Here, for the November 05, 2020 announcement, revenues declined due to decline of home-buying and other effects of Covid-19. Yet management is putting a lot of efforts into their discussions to highlight the increasing user engagement on their platform which should result in higher next-quarter revenues from other business segments. We sort Zillow into High FtBERT quintile portfolio in the end of November, i.e. almost a month after the announcement. The return of the stock for December 2020 is impressive 23%. It looks like it took the market participants almost two months to believe MD&A statements. There is no further valuation improvement in January 2021, and the next highest return, 22%, for February, 2021, is due to the new, February 10, 2021 earnings announcement – Zillow beats Q4 expectations as revenues did not fall as much as expected due to a surge in revenues from brokerage segments facilitated by the platform services.

On the negative side, Table 11 and Table 12 provide extracts from RF sections for healthcare companies, where FtBERT puts the most of its weight to place these companies in the lowest performance quintile. The discussions here are focused on clinical study risks which eventually do not deliver positive outcomes ex-post. Cohen et al. (2020) make extensive use of content from RF sections; however, the authors provide time series evidence, i.e. changes from one report to another, whereas we report cross-sectional results.

Consider for example Table 11, a small-cap pharmaceutical company, Harpoon Therapeutics Inc. The extract of the RF section, to which FtBERT assigns the highest attention weight, talks excessively over the possible failures in clinical trials and mentions all possible reasons of what can go wrong. This is based on May 06, 2021 report, and we place this stock in the Low FtBERT quintile portfolio only in the end of May, i.e. almost one month after. The stock collapses first in June 2021, -33%, and further on in July 2021, -29%. The August returns is very small, -3% given that all the negative news already had been reported before the official filing release on August 05, 2021.

Another example is the large cap pharmaceutical, Allogene Therapeutics Inc, Table 12. FtBERT places more weight here again on RF section of November 04 2021 10-Q report, and predicts negative earnings surprise next quarter. This section is relatively long, and the risks associated with clinical trials are very similar to those in Table 11, i.e. they are standard for this industry, except it talks more about the risks of FDA approvals. What happens ex-post is indeed FDA imposing a clinical hold on the trials. The news sources demonstrate investors' disappointment of management being not straightforward in their communications, and even after the lift of clinical trials hold, the negative sentiment remains. The stock declines for three consecutive months after being placed into the Low performance quintile by FtBERT.

6 Conclusion

This is the first paper to: (i) provide a comprehensive analysis across different NLP approaches about their efficiency in identifying positive and negative informational content in the most publicly scrutinized corporate 10-Q and 10-K filings; (ii) introduce new applications of LLMs in the finance context, in particular hierarchical aggregation of information to process corporate disclosures of arbitrary length; (iii) being able to identify not only negative, similar to the previous literature, but also positive informational content in the 10-Q and 10-K filings, and subsequent negative and positive stock abnormal returns respectively.

We have a few important conclusions. First, none of traditional NLP approaches are able to robustly identify future positive or negative firms' valuation changes. This however does not mean that 10-Q or 10-K reports are not useful to communicate new forward-looking information to market participants. They are simply too long and too complex, and traditional methodologies fail in part due to the reports' complexity.

Second, the off-shelf LLMs, even after training on financial targets, might not be worth the efforts as they are as good as more simple character-count approaches.

Third, fine-tuning and training LLMs on financial targets, or "making LLMs first learn finance" before using their predictions, is a solution and a fruitful path for the future research.

FtBERT which we introduce in this paper provides unparalleled results in identifying future positive and negative performances. Further to it, we can also identify what parts of financial text receive the highest weights for FtBERT to make positive vs negative forecasts. We cross-check ex-post, manually, whether the ex-post performance is indeed attributed to the reports' text that FtBERT relies the most, and confirm its high accuracy.

Finally, we bring attention to the valuable information content of 10-Q and 10-K reports. We also find that market participants react very slowly to this information, largely due to high disagreement about its interpretation.

References

Atmaz, A. and Basak, S. (2018). Belief dispersion in the stock market. The Journal of Finance, 73(3):1225–1279.

Ba, J., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. ArXiv, abs/1607.06450.

Bahri, Y., Dyer, E., Kaplan, J., Lee, J., and Sharma, U. (2021). Explaining neural scaling laws.

Cao, S., Jiang, W., Yang, B., and Zhang, A. L. (2023). How to talk when a machine is listening: Corporate disclosure in the age of ai. Review of Financial Studies, Forthcoming.

Carhart, M. M. (1997). On persistence in mutual fund performance. The Journal of finance, 52(1):57–82.

Chen, T. and Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pages 785–794, New York, NY, USA. ACM.

Chen, X., Liang, C., Huang, D., Real, E., Wang, K., Liu, Y., Pham, H., Dong, X., Luong, T., Hsieh, C.-J., Lu, Y., and Le, Q. V. (2023). Symbolic discovery of optimization algorithms.

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

Colin, R. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR, 21(140):1.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Dimson, E. (1979). Risk measurement when shares are subject to infrequent trading. Journal of Financial Economics, 7(2):197–226.

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

Goldstein, I., Spatt, C. S., and Ye, M. (2021). Big data in finance. The Review of Financial Studies, 34(7):3213–3225.

Golez, B. and Goyenko, R. (2022). Disagreement in the equity options market and stock returns. The Review of Financial Studies, 35(3):1443–1479.

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

Huang, A. G., Tan, H., and Wermers, R. (2020). Institutional trading around corporate news: Evidence from textual analysis. The Review of Financial Studies, 33(10):4627–4675.

Huang, A. H., Wang, H., and Yang, Y. (2022). Finbert: A large language model for extracting information from financial text. Contemporary Accounting Research.

Jegadeesh, N. and Wu, D. (2013). Word power: A new approach for content analysis. Journal of financial economics, 110(3):712–729.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. (2020). Scaling laws for neural language models.

Kelly, B. T., Malamud, S., and Zhou, K. (2023). The virtue of complexity in return prediction. The Journal of Finance, Forthcoming.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. Journal of Accounting and economics, 45(2-3):221–247.

Lin, T., Wang, Y., Liu, X., and Qiu, X. (2022). A survey of transformers. AI Open, 3:111–132.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

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. (2014). Measuring readability in financial disclosures. the Journal of Finance, 69(4):1643–1671.

Manela, A. and Moreira, A. (2017). News implied volatility and disaster concerns. Journal of Financial Economics, 123(1):137–162.

Newey, W. K. and West, K. D. (1986). A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Yvon, F. (2023). Transformers in natural language processing. In Chetouani, M., Dignum, V., Lukowicz, P., and Sierra, C., editors, Human-Centered Artificial Intelligence: Advanced Lectures, pages 81–105, Cham. Springer International Publishing.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. (2022). Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068.

	OLS	EN	Lasso	XGB	RF	SVR	NN	LM	FrozenBERT	FtBERT
MSE	0.64	0.09	0.08	0.08	0.08	0.08	0.25	0.08	0.08	0.08
OOS R^2	-6.670	-0.027	0.000	0.014	0.027	-0.014	-2.040	0.001	0.004	0.001

Notes: The table reports out-of-sample mean squared error (MSE) and R^2 for different regression-based model predictions. Mean squared error is calculated using $\sum_t \sum_i \sqrt{(r_{i,t} - E[\hat{r}_{i,t}])^2}$, where $E[\hat{r}_{i,t}]$ is the out-of-sample predicted rank for stock i at time t. Out-of-sample R^2 is calculated as $1 - \text{MSE}/\text{MSE}_b$, with MSE_b the MSE of a benchmark model which uses 0.5 as the predicted ranks. The out-of-sample period is from 01/2003 to 12/2021.

Table 1: Out-of-sample MSE and R^2

Table 2: Lexicon & Sentiment based models: portfolio performance

	EW Portfolios						VW Portfolios					
	L	Q2	Q3	Q4	H	H-L	L	Q2	Q3	Q4	H	H-L
	Panel A. LM negative sentiment											
Raw Returns	1.246	1.323	1.282	1.424	1.225	-0.021	0.921	1.218	0.971	0.951	0.966	0.045
	(3.43)	(3.66)	(3.53)	(4.02)	(3.41)	-(0.36)	(3.20)	(4.31)	(3.37)	(3.32)	(3.40)	(0.40)
CAPM	0.132	0.222	0.177	0.347	0.124	-0.008	0.013	0.354	0.065	0.049	0.065	0.053
	(1.02)	(1.64)	(1.31)	(2.59)	(0.96)	-(0.14)	(0.16)	(3.44)	(0.81)	(0.63)	(0.90)	(0.46)
FF5	0.246	0.329	0.278	0.431	0.226	-0.021	0.012	0.283	0.050	0.053	0.060	0.048
	(4.71)	(5.73)	(4.74)	(6.83)	(3.72)	-(0.33)	(0.15)	(2.72)	(0.61)	(0.65)	(0.80)	(0.40)
FF6	0.249	0.342	0.286	0.443	0.237	-0.012	-0.006	0.289	0.050	0.058	0.067	0.074
	(4.77)	(6.21)	(4.95)	(7.21)	(4.04)	-(0.19)	-(0.08)	(2.77)	(0.60)	(0.72)	(0.90)	(0.64)
	Panel B. FinBERT											
Raw Returns	1.233	1.254	1.267	1.346	1.383	0.150	0.835	0.924	0.965	1.022	1.147	0.313
	(3.60)	(3.57)	(3.61)	(3.62)	(3.46)	(1.11)	(2.97)	(3.26)	(3.48)	(3.49)	(3.75)	(2.23)
CAPM	0.197	0.176	0.199	0.221	0.188	-0.010	-0.044	0.032	0.093	0.099	0.206	0.250
	(1.48)	(1.41)	(1.51)	(1.53)	(1.13)	-(0.08)	-(0.52)	(0.41)	(1.19)	(1.22)	(1.93)	(1.76)
FF5	0.284	0.260	0.268	0.324	0.356	0.072	-0.056	0.008	0.096	0.077	0.195	0.251
	(3.89)	(4.89)	(4.29)	(4.47)	(4.20)	(0.58)	-(0.64)	(0.10)	(1.24)	(0.92)	(1.85)	(1.73)
FF6	0.274	0.262	0.279	0.352	0.372	0.098	-0.073	0.008	0.101	0.090	0.191	0.264
	(3.81)	(4.91)	(4.55)	(5.74)	(4.52)	(0.82)	-(0.86)	(0.09)	(1.30)	(1.09)	(1.81)	(1.83)
	Panel C. MD&A length											
Raw Returns	1.361	1.306	1.394	1.247	1.173	-0.189	1.157	0.993	1.047	0.951	0.893	-0.264
	(3.90)	(3.53)	(3.81)	(3.42)	(3.22)	-(1.71)	(4.23)	(3.37)	(3.67)	(3.29)	(3.06)	-(1.92)
CAPM	0.302	0.193	0.279	0.130	0.073	-0.230	0.316	0.074	0.158	0.044	-0.028	-0.344
	(2.25)	(1.31)	(2.05)	(0.99)	(0.51)	-(2.05)	(3.35)	(0.83)	(1.81)	(0.53)	-(0.36)	-(2.48)
FF5	0.367	0.290	0.370	0.236	0.219	-0.147	0.204	0.095	0.122	0.021	0.027	-0.177
	(5.84)	(4.39)	(6.08)	(3.77)	(2.83)	-(1.53)	(2.23)	(1.09)	(1.44)	(0.25)	(0.38)	-(1.38)
FF6	0.370	0.295	0.384	0.251	0.232	-0.138	0.192	0.092	0.117	0.021	0.036	-0.156
	(5.88)	(4.48)	(6.58)	(4.19)	(3.05)	-(1.43)	(2.13)	(1.05)	(1.38)	(0.24)	(0.52)	-(1.25)
	Panel D. RF section length											
Raw Returns	1.112	1.128	1.098	1.252	1.297	0.186	0.951	0.934	0.930	1.230	0.862	-0.089
	(2.79)	(2.88)	(2.63)	(2.99)	(2.98)	(1.16)	(3.18)	(2.99)	(2.69)	(3.45)	(2.41)	-(0.56)
CAPM	0.128	0.146	0.075	0.208	0.243	0.114	0.184	0.137	0.042	0.353	-0.051	-0.235
	(0.82)	(1.03)	(0.44)	(1.34)	(1.28)	(0.71)	(2.14)	(1.39)	(0.42)	(2.48)	-(0.47)	-(1.53)
FF5	0.229	0.307	0.215	0.385	0.512	0.282	0.039	0.149	0.048	0.320	0.098	0.059
	(3.14)	(4.90)	(2.40)	(4.53)	(4.53)	(2.15)	(0.50)	(1.45)	(0.47)	(2.36)	(0.93)	(0.42)
FF6	0.236	0.314	0.223	0.391	0.518	0.282	0.036	0.146	0.049	0.309	0.106	0.070
	(3.29)	(5.13)	(2.53)	(4.64)	(4.60)	(2.14)	(0.46)	(1.42)	(0.48)	(2.32)	(1.01)	(0.51)

Notes: The table reports monthly percentage quintile portfolio excess returns sorted on lexicon- and sentiment-based models' predictions. The protfolio returns are either equally-weighted, EW, or value-weighted, VW. The models are: LM negative sentiment score, LM 2011, Panel A, FinBERT negative sentence fraction, Panel B, Managerial Discussion and Analysis section length, MD&A length, Panel C, and Risk Factor, RF, section length, Panel D. Stocks are sorted into the quintile portfolios in the end of the month in which one of their 10-K or 10-Q reports is publicly released based on their predicted ranks, with the "L (H)" portfolio containing stocks with the lowest (highest) predicted ranks. Stocks are held in the portfolio for three months, while the portfolios are rebalanced monthly. The statistics reported are: raw excess returns (Raw Returns), CAPM alpha (CAPM), Fama-French five-factor alpha (FF5), and Fama-French five-factor plus momentum alpha (FF6). t-Statistics are reported below the estimates in parentheses. The *OOS* test period is from 01/2003 to 12/2021 for all measures, except RF, Panel D, from 01/2006 to 12/2021.

Table 3: Bag-of-Words Models: portfolio performance

	EW Portfolios						VW Portfolios					
	L	Q2	Q3	Q4	H	H-L	L	Q2	Q3	Q4	H	H-L
						Panel A. OLS						
Raw Returns	1.112	1.126	1.279	1.493	1.210	0.099	1.115	0.929	0.972	1.051	0.854	-0.260
	(3.11)	(3.24)	(3.70)	(3.52)	(3.15)	(0.77)	(3.87)	(2.89)	(3.09)	(3.39)	(2.65)	(-1.46)
CAPM	0.048	0.093	0.268	0.439	0.074	0.027	0.268	0.026	0.094	0.154	-0.104	-0.372
	(0.32)	(0.64)	(1.70)	(1.56)	(0.45)	(0.21)	(2.08)	(0.15)	(0.57)	(1.05)	(-0.75)	(-2.07)
FF5	0.104	0.181	0.345	0.547	0.162	0.058	0.249	0.002	0.108	0.184	-0.056	-0.306
	(0.99)	(1.90)	(3.60)	(2.45)	(1.35)	(0.43)	(1.89)	(0.01)	(0.64)	(1.25)	(-0.41)	(-1.73)
FF6	0.109	0.172	0.352	0.545	0.170	0.061	0.244	0.004	0.116	0.192	-0.061	-0.305
	(1.03)	(1.82)	(3.69)	(2.43)	(1.41)	(0.45)	(1.85)	(0.02)	(0.68)	(1.30)	(-0.44)	(-1.72)
						Panel B. LM OLS						
Raw Returns	1.244	1.215	1.254	1.341	1.431	0.187	0.990	0.877	0.972	0.930	1.209	0.219
	(3.71)	(3.40)	(3.42)	(3.64)	(3.64)	(1.42)	(3.66)	(3.12)	(3.58)	(3.22)	(3.65)	(1.41)
CAPM	0.207	0.130	0.143	0.234	0.267	0.060	0.148	0.001	0.125	0.026	0.189	0.040
	(1.85)	(0.96)	(1.00)	(1.58)	(1.56)	(0.47)	(1.76)	(0.02)	(1.48)	(0.31)	(1.67)	(0.27)
FF5	0.249	0.201	0.231	0.350	0.463	0.214	0.089	-0.030	0.087	0.020	0.249	0.160
	(4.24)	(3.03)	(3.49)	(4.93)	(5.56)	(2.02)	(1.09)	-(0.33)	(1.06)	(0.23)	(2.30)	(1.10)
FF6	0.250	0.207	0.242	0.369	0.470	0.220	0.073	-0.027	0.081	0.033	0.249	0.176
	(4.25)	(3.14)	(3.76)	(5.59)	(5.66)	(2.07)	(0.92)	-(0.31)	(1.00)	(0.39)	(2.29)	(1.22)
						Panel C. EN						
Raw Returns	0.998	1.203	1.314	1.520	1.182	0.184	1.021	1.060	0.836	1.000	1.016	-0.005
	(2.90)	(3.51)	(3.70)	(3.61)	(2.94)	(1.24)	(3.57)	(3.57)	(2.76)	(3.08)	(2.99)	(-0.03)
CAPM	-0.011	0.195	0.278	0.489	-0.028	-0.017	0.216	0.217	-0.032	0.086	0.005	-0.211
	(-0.07)	(1.28)	(1.70)	(1.72)	(-0.17)	(-0.12)	(1.45)	(1.44)	(-0.22)	(0.51)	(0.04)	(-1.15)
FF5	0.064	0.278	0.362	0.578	0.045	-0.019	0.215	0.193	-0.074	0.068	0.042	-0.172
	(0.59)	(3.01)	(3.55)	(2.42)	(0.40)	(-0.13)	(1.39)	(1.24)	(-0.48)	(0.41)	(0.29)	(-0.92)
FF6	0.074	0.284	0.362	0.572	0.046	-0.028	0.222	0.191	-0.080	0.061	0.037	-0.186
	(0.68)	(3.09)	(3.53)	(2.39)	(0.40)	(-0.20)	(1.44)	(1.23)	(-0.52)	(0.37)	(0.25)	(-0.99)
						Panel D. Lasso						
Raw Returns	1.049	1.064	1.488	1.335	1.289	0.240	0.906	0.727	0.916	1.116	1.238	0.332
	(3.04)	(3.03)	(3.52)	(3.70)	(3.27)	(1.56)	(3.10)	(2.44)	(3.06)	(3.37)	(3.48)	(1.51)
CAPM	0.042	0.037	0.470	0.276	0.106	0.063	0.076	-0.139	0.106	0.164	0.211	0.135
	(0.27)	(0.23)	(1.60)	(1.71)	(0.66)	(0.43)	(0.51)	(-1.00)	(0.62)	(1.02)	(1.23)	(0.62)
FF5	0.110	0.121	0.528	0.371	0.199	0.089	0.037	-0.150	0.101	0.134	0.287	0.250
	(0.98)	(1.29)	(2.15)	(3.31)	(1.82)	(0.63)	(0.25)	(-1.05)	(0.57)	(0.82)	(1.70)	(1.18)
FF6	0.115	0.133	0.513	0.375	0.202	0.087	0.052	-0.148	0.090	0.129	0.271	0.219
	(1.02)	(1.45)	(2.09)	(3.34)	(1.84)	(0.61)	(0.35)	(-1.03)	(0.50)	(0.79)	(1.62)	(1.06)
						Panel E. SVR						
Raw Returns	1.474	1.067	1.108	1.282	1.287	-0.187	1.069	0.875	0.950	1.097	0.934	-0.135
	(3.71)	(3.06)	(2.99)	(3.47)	(3.30)	(-0.75)	(3.43)	(2.78)	(2.99)	(3.32)	(2.89)	(-0.57)
CAPM	0.560	0.060	0.013	0.180	0.130	-0.430	0.296	0.011	0.018	0.115	-0.021	-0.316
	(1.94)	(0.36)	(0.08)	(1.16)	(0.78)	(-1.75)	(1.42)	(0.06)	(0.12)	(0.82)	(-0.15)	(-1.35)
FF5	0.650	0.093	0.071	0.300	0.235	-0.415	0.288	-0.044	0.012	0.136	0.042	-0.246
	(2.70)	(0.79)	(0.71)	(2.83)	(1.96)	(-1.70)	(1.37)	(-0.25)	(0.08)	(0.95)	(0.29)	(-1.03)
FF6	0.639	0.098	0.071	0.305	0.244	-0.395	0.277	-0.046	0.003	0.139	0.055	-0.222

Continued on next page

Table 3 – continued from previous page

	EW Portfolios						VW Portfolios					
	L	Q2	Q3	Q4	H	H-L	L	Q2	Q3	Q4	H	H-L
	(2.65)	(0.83)	(0.72)	(2.88)	(2.04)	(-1.62)	(1.32)	(-0.26)	(0.02)	(0.96)	(0.39)	(-0.94)
	Panel F. XGB											
Raw Returns	0.940	1.113	1.186	1.605	1.341	0.401	0.920	0.799	0.831	1.294	1.033	0.113
	(2.86)	(3.28)	(3.20)	(3.62)	(3.34)	(1.91)	(3.13)	(2.76)	(2.58)	(3.97)	(2.89)	(0.46)
CAPM	0.018	0.128	0.099	0.482	0.161	0.143	0.134	-0.018	-0.080	0.347	-0.012	-0.146
	(0.10)	(0.81)	(0.59)	(1.68)	(0.90)	(0.72)	(0.78)	(-0.12)	(-0.49)	(2.27)	(-0.08)	(-0.61)
FF5	0.141	0.129	0.140	0.556	0.308	0.168	0.187	-0.023	-0.200	0.326	0.075	-0.113
	(1.18)	(1.19)	(1.24)	(2.28)	(2.53)	(0.96)	(1.08)	(-0.15)	(-1.20)	(2.07)	(0.48)	(-0.49)
FF6	0.145	0.141	0.148	0.543	0.306	0.162	0.189	-0.021	-0.184	0.313	0.070	-0.119
	(1.21)	(1.31)	(1.31)	(2.23)	(2.50)	(0.93)	(1.08)	(-0.14)	(-1.11)	(2.00)	(0.45)	(-0.52)
	Panel G. Random forest											
Raw Returns	0.899	1.159	1.443	1.333	1.386	0.487	0.727	1.122	1.031	1.087	0.843	0.116
	(2.62)	(3.24)	(3.44)	(3.55)	(3.56)	(2.63)	(2.31)	(3.76)	(3.41)	(3.23)	(2.44)	(0.47)
CAPM	-0.073	0.109	0.424	0.208	0.252	0.325	-0.142	0.257	0.178	0.096	-0.149	-0.006
	(-0.42)	(0.68)	(1.47)	(1.36)	(1.40)	(1.78)	(-0.83)	(1.82)	(1.14)	(0.64)	(-0.88)	(-0.03)
FF5	0.049	0.160	0.491	0.267	0.367	0.318	-0.098	0.146	0.169	0.091	-0.026	0.072
	(0.38)	(1.45)	(2.02)	(2.81)	(2.72)	(1.71)	(-0.56)	(1.02)	(1.05)	(0.59)	(-0.15)	(0.29)
FF6	0.061	0.160	0.482	0.272	0.365	0.304	-0.082	0.146	0.169	0.084	-0.029	0.053
	(0.48)	(1.45)	(1.98)	(2.88)	(2.70)	(1.64)	(-0.47)	(1.01)	(1.05)	(0.55)	(-0.18)	(0.21)
	Panel H. NN											
Raw Returns	1.004	1.229	1.210	1.258	1.533	0.529	0.869	0.959	0.985	0.999	0.994	0.125
	(2.74)	(3.51)	(3.45)	(3.56)	(3.42)	(2.05)	(2.90)	(3.19)	(3.22)	(3.14)	(2.99)	(0.65)
CAPM	-0.069	0.188	0.166	0.225	0.418	0.487	-0.004	0.117	0.093	0.069	0.044	0.048
	(-0.41)	(1.26)	(1.13)	(1.40)	(1.40)	(1.84)	(-0.03)	(0.74)	(0.66)	(0.48)	(0.27)	(0.24)
FF5	0.009	0.281	0.233	0.287	0.520	0.512	-0.059	0.078	0.100	0.081	0.080	0.140
	(0.08)	(2.86)	(2.21)	(2.55)	(2.15)	(1.93)	(-0.42)	(0.48)	(0.69)	(0.55)	(0.49)	(0.70)
FF6	0.020	0.283	0.239	0.283	0.513	0.494	-0.060	0.084	0.102	0.065	0.080	0.140
	(0.18)	(2.88)	(2.26)	(2.51)	(2.11)	(1.86)	(-0.42)	(0.52)	(0.71)	(0.45)	(0.49)	(0.70)

Notes: The table reports monthly percentage quintile portfolio excess returns sorted on bag-of-word models' predictions. The protfolio returns are either equally-weighted, EW, or value-weighted, VW. The models are: OLS (OLS) Panel A, OLS model with LM negative sentiment as regresses (LM OLS) Panel B, elastic nets (EN) Panel C, Lasso Panel D, support vector regression (SVR) Panel E, XGBoost Panel F, random forest Panel G, and feed-forward neural networks, NN, Panel H. Stocks are sorted into the quintile portfolios in the end of the month in which one of their 10-K or 10-Q reports is publicly released based on their predicted ranks, with the "L (H)" portfolio containing stocks with the lowest (highest) predicted ranks. Stocks are held in the portfolio for three months, while the portfolios are rebalanced monthly. The statistics reported are: raw excess returns (Raw Returns), CAPM alpha (CAPM), Fama-French five-factor alpha (FF5), and Fama-French five-factor plus momentum alpha (FF6). t-Statistics are reported below the estimates in parentheses. The *OOS* test period is from 01/2003 to 12/2021

	EW Portfolios						VW Portfolios					
	L	Q2	Q3	Q4	H	H-L	L	Q2	Q3	Q4	H	H-L
	Panel A. FrozenBERT											
Raw Returns	1.081	1.210	1.320	1.391	1.488	0.407	0.745	0.813	0.998	1.211	1.175	0.430
	(3.09)	(3.39)	(3.61)	(3.79)	(3.91)	(2.77)	(2.52)	(2.98)	(3.51)	(4.24)	(3.73)	(2.51)
CAPM	0.039	0.125	0.204	0.282	0.338	0.299	-0.164	-0.037	0.104	0.325	0.206	0.370
	(0.26)	(0.92)	(1.52)	(1.96)	(2.29)	(2.03)	-(1.59)	-(0.44)	(1.32)	(3.52)	(1.89)	(2.12)
FF5	0.168	0.202	0.275	0.370	0.480	0.312	-0.109	-0.070	0.095	0.290	0.226	0.334
	(2.07)	(3.11)	(4.30)	(5.44)	(5.98)	(2.50)	-(1.17)	-(0.82)	(1.19)	(3.07)	(2.12)	(2.11)
FF6	0.178	0.215	0.285	0.377	0.487	0.309	-0.106	-0.067	0.091	0.291	0.210	0.316
	(2.21)	(3.41)	(4.55)	(5.56)	(6.07)	(2.46)	-(1.13)	-(0.78)	(1.14)	(3.08)	(2.00)	(2.01)
	Panel B. FtBERT											
Raw Returns	1.092	1.206	1.292	1.362	1.531	0.439	0.773	0.846	0.998	1.095	1.334	0.561
	(3.08)	(3.41)	(3.58)	(3.65)	(4.01)	(2.70)	(2.67)	(2.95)	(3.57)	(3.65)	(4.33)	(2.94)
CAPM	0.057	0.129	0.188	0.228	0.380	0.323	-0.106	-0.042	0.128	0.153	0.395	0.501
	(0.35)	(0.98)	(1.43)	(1.61)	(2.49)	(1.99)	-(0.97)	-(0.45)	(1.44)	(1.77)	(3.44)	(2.57)
FF5	0.222	0.204	0.264	0.321	0.480	0.257	-0.002	-0.004	0.152	0.147	0.331	0.333
	(2.30)	(3.12)	(4.38)	(4.49)	(5.73)	(1.82)	-(0.02)	-(0.04)	(1.70)	(1.66)	(2.96)	(1.88)
FF6	0.234	0.218	0.273	0.328	0.484	0.250	0.008	0.002	0.141	0.140	0.317	0.309
	(2.45)	(3.46)	(4.61)	(4.61)	(5.78)	(1.76)	(0.08)	(0.02)	(1.60)	(1.59)	(2.86)	(1.77)

Notes: The table reports monthly percentage quintile portfolio excess returns sorted on Large Language Models, LLMs, predictions. The protfolio returns are either equally-weighted, EW, or value-weighted, VW. The models are: Frozen BERT, Panel A, an off-shelf BERT trained on predicting earnings surprises; and Fine-tuned BERT, FtBERT, Panel B, an off-shelf BERT trained & fine-tuned to predict earnings surprises. Stocks are sorted into the quintile portfolios in the end of the month in which one of their 10-K or 10-Q reports is publicly released based on their predicted ranks, with the "L (H)" portfolio containing stocks with the lowest (highest) predicted ranks. Stocks are held in the portfolio for three months, while the portfolios are rebalanced monthly. The statistics reported are: raw excess returns (Raw Returns), CAPM alpha (CAPM), Fama-French five-factor alpha (FF5), and Fama-French five-factor plus momentum alpha (FF6). t-Statistics are reported below the estimates in parentheses. The *OOS* test period is from 01/2003 to 12/2021.

Table 4: Large Language Models: portfolio performance

Panel A: Univariate regressions					
MD&A length	0.0002				
	(0.07)				
RF section length		0.006			
		(1.25)			
FinBERT			0.008		
			(1.73)		
FrozenBERT				0.106	
				(5.03)	
FtBERT					0.049
					(5.27)
Time Fixed Effect	Yes	Yes	Yes	Yes	Yes
Firm Fixed Effect	Yes	Yes	Yes	Yes	Yes
R^2	0.0000	0.0000	0.0000	0.0002	0.0002
Panel B: With controls					
MD&A length	-0.0048				
	-(1.12)				
RF section length		-0.001			
		-(0.11)			
FinBERT			-0.002		
			-(0.44)		
FrozenBERT				0.106	
				(4.97)	
FtBERT					0.058
					(5.93)
Size	0.000	0.000	0.000	0.000	0.000
	-(2.21)	-(2.13)	-(2.16)	-(2.17)	-(2.17)
BM	0.014	0.017	0.015	0.015	0.015
	(7.70)	(6.18)	(8.04)	(8.15)	(8.17)
Ret(-1,0)	0.003	0.005	0.002	0.002	0.002
	(0.84)	(1.01)	(0.65)	(0.62)	(0.58)
Ret(-12,-1)	-0.005	-0.006	-0.005	-0.005	-0.005
	-(4.63)	-(4.05)	-(4.73)	-(4.77)	-(4.83)
Time Fixed Effect	Yes	Yes	Yes	Yes	Yes
Firm Fixed Effect	Yes	Yes	Yes	Yes	Yes
R^2	0.0023	0.0028	0.0024	0.0026	0.0027

Notes: The table presents the results of predictive panel regressions for monthly individual firm-level stock returns on selected model predicted ranks and a number of known firm-specific return predictors. The dependent variable is firm-level returns, in percentage points, following the month of 10-K or 10-Q filing releases. MD&A length is the number of characters in the Managerial Discussion and Analysis section; RF section length is the number of characters in the Risk Factor section; FinBERT is the negative sentiment score identified by FinBERT; FrozenBERT and FtBERT are predicted ranks from frozen BERT model and fine-tuned BERT model, respectively. Size is market value of equity in billions of dollars, BM is the book value of equity over market value of equity, Ret(-1,0) is the previous month's return, and Ret(–12,–1) is the cumulative return from month –12 to month -1. The right-hand side firm-specific return predictors are winsorized at the 1% level. t-Statistics reported in parentheses below the estimates are based on clustered standard errors at the firm level. The results are reported for *OOS* period, 01/2003 to 12/2021.

Table 5: Predicting Stock Returns: a Linear Regression Approach

MD&A length	-0.005				
	-(0.91)				
RF Section Length		-0.006			
		-(0.85)			
FinBERT			-0.007		
			-(1.38)		
FrozenBERT				0.122	
				(5.10)	
FtBERT					0.059
					(5.19)
Size	0.000	0.000	0.000	0.000	0.000
	-(1.96)	-(2.00)	-(1.92)	-(1.93)	-(1.93)
BM	0.018	0.022	0.018	0.018	0.018
	(8.60)	(7.09)	(8.90)	(9.01)	(9.03)
Ret(-1,0)	0.000	-0.001	-0.002	-0.002	-0.002
	-(0.07)	-(0.13)	-(0.42)	-(0.43)	-(0.45)
Ret(-12,-1)	-0.007	-0.007	-0.007	-0.007	-0.007
	-(2.48)	-(2.13)	-(2.50)	-(2.50)	-(2.53)
Accruals	-0.003	-0.007	-0.004	-0.004	-0.004
	-(0.21)	-(0.31)	-(0.31)	-(0.31)	-(0.31)
Cash flow	-0.001	-0.006	-0.002	-0.002	-0.002
	-(0.06)	-(0.47)	-(0.23)	-(0.25)	-(0.27)
Profit	-0.023	-0.036	-0.024	-0.024	-0.024
	-(3.68)	-(4.31)	-(3.87)	-(3.90)	-(3.90)
Investment	0.000	0.000	0.000	0.000	0.000
	-(0.46)	-(0.37)	-(0.30)	-(0.32)	-(0.36)
Ret(-6,-1)	0.006	0.003	0.006	0.006	0.006
	(0.89)	(0.34)	(0.82)	(0.81)	(0.81)
SUE_t	0.064	0.047	0.070	0.070	0.069
	(1.07)	(0.77)	(1.19)	(1.19)	(1.17)
Time Fixed Effect	Yes	Yes	Yes	Yes	Yes
Firm Fixed Effect	Yes	Yes	Yes	Yes	Yes
R^2	0.0035	0.0042	0.0035	0.0038	0.0038

Notes: The table presents the results of predictive panel regressions for individual monthly firm-level stock returns on selected model predicted ranks and a number of known firm-specific return predictors. The dependent variable is firm-level returns, in percentage points, following the month of 10-K or 10-Q filing releases. MD&A length is the number of characters in the Managerial Discussion and Analysis section; RF section length is the number of characters in the Risk Factor section; FinBERT is the negative sentiment score given by FinBERT; FrozenBERT and FTBERT are predicted ranks from frozen BERT model and fine-tuned BERT model, respectively. Size is market value of equity in billions of dollars; BM is the book value of equity over market value of equity; Ret(–1,0) is the previous month's return; Ret(-3,-1), Ret(–6,–1), Ret(–9,–1), and Ret(–12,–1) are the cumulative return from month -3 to month -1, month -6 to month -1, month -9 to month -1, and month -12 to month -1, respectively. Invest is the ratio of capital investment (capx) to revenue (revt) divided by the firm-specific 36-month rolling mean of that ratio. Profit is Revenue (sale) - cost of goods sold (cogs), divided by total assets (at). Cash Flow is net income (ib) plus depreciation (dp) divided by market equity. Accrual is $(\Delta\text{act} - \text{chech} - \Delta\text{lct} + \Delta\text{dct} + \Delta\text{txp} - \text{dp})$ scaled by average assets $(\text{at}/2 + \text{lag(at)}/2)$. SUE_t is the most recent earnings surprise preceding the month we measure returns. The right-hand side firm-specific return predictors are winsorized at the 1% level. t-Statistics reported in parentheses below the estimates are based on clustered standard errors at the firm level. The results are reported for *OOS* period, 01/2003 to 12/2021.

Table 6: Predicting Stock Returns: a Linear Regression Approach and Further Firm-Specific Controls

	SUE_{t+3}	CAR[1,5]
	Panel A. FrozenBERT	
FrozenBERT	1.198	1.251
	(2.14)	(1.00)
Time Fixed Effect	Yes	Yes
Firm Fixed Effect	Yes	Yes
R2	0.0001	0.0000
	Panel B. FtBERT	
FtBERT	0.963	1.188
	(3.62)	(1.99)
Time Fixed Effect	Yes	Yes
Firm Fixed Effect	Yes	Yes
R2	0.0002	0.0000

Notes: The table reports panel regressions around earnings announcements and releases of 10-Q, 10-K reports. The most recent filing of 10-Q, 10-K reports allows to update FrozenBERT, Panel A, and FtBERT, Panel B, predictions about next period earnings surprise, SUE_{t+3}. It then further allows to identify the direction of price discovery caused by institutional trading around the filing day, i.e. the post filing 5 day cumulative abnormal return, CAR[1,5], at the current, month t, announcement. t-statistics reported below the estimates are based on standard errors clustered by firm.

Table 7: Event Study: Earnings Announcements and Filing Day Returns

	L	Q2	Q3	Q4	H
Size	7.726	7.962	7.935	8.460	11.738
BM	0.640	0.570	0.534	0.513	0.484
# analysts	8.648	9.215	9.212	9.453	10.009
Stdev analysts	0.143	0.134	0.129	0.131	0.141
SUE_{t+3}	-0.058	0.020	0.059	0.086	0.078
SUE_t	-0.061	-0.180	0.079	-0.025	0.070
Hit ratio	0.226	0.217	0.207	0.218	0.242

Notes: The table reports stock characteristics of FtBERT quintile portfolios reported in Table 4, Panel B. Size is the average market capitalization of equity in billions of dollars; BM is the average book-to-market value ; # analysts is the average number of analysts making their eanigns forecasts; Stdev analysts is the standard deviations of analyst forecasts; SUE_{t+3} is the next-quarter earnings surprise standardized by the stock price; SUE_t is the current quarter earnings surprise standardized by the stock price; Hit ratio is the proportion of the quintile portfolio forecasts which ex-post turn out to be correct. The benchmark for Hit-ratio is 0.2 for each portfolio (i.e. =1/5 portfolios)

Table 8: Characteristics of FtBERT quintile portfolios

ZOOM VIDEO COMMUNICATIONS INC								
File Date	Sort Date	Assigned Quintile by FtBERT	Market Cap ($Mln)	GIC	Next Announcement Date	$R_{April,2020}$	$R_{May,2020}$	$R_{June,2020}$
2020-03-20	2020-03-31	High	24242.04	IT	2020-06-02	-0.07	0.33	0.41

Explanation

The fear of another wave of COVID-19 infections increased the demand for remote communication tools like Zoom.

Text Highlights

Our paid offerings include our pro, business, and enterprise plans, which provide incremental features and functionality, such as different participant limits, administrative controls, and reporting. our revenue was 622.7 million, 330.5 million, and 151.5 million for the fiscal years ended January 31, 2020, 2019, and 2018, respectively, representing period-over-period growth rate of 88 and 118 for fiscal year 2020 and fiscal year 2019, respectively. we had net income of 25.3 million and 7.6 million for the fiscal years ended January 31, 2020 and 2019, respectively, and a net loss of 3.8 million for the fiscal year ended January 31, 2018. net cash provided by operating activities was 151.9 million, 51.3 million, and 19.4 million for the fiscal years ended January 31, 2020, 2019, and 2018, respectively. key factors affecting our performance acquiring new customers we are focused on continuing to grow the number of customers that use our platform. our operating results and growth prospects will depend, in part, on our ability to attract new customers. while we believe we have a significant market opportunity that our platform addresses, it is difficult to predict customer adoption rates or the future growth rate and size of the market for our platform. we will need to continue to invest in sales and marketing in order to address this opportunity by hiring, developing, and retaining talented sales personnel who are able to achieve desired productivity levels in a reasonable period of time. expansion of zoom across existing customers we believe that there is a large opportunity for growth with many of our existing customers. many customers have increased the size of their subscriptions as they have expanded their use of our platform across their operations. some of our larger enterprise customers start with a single deployment of zoom meetings with one team, location, or geography before rolling out our platform throughout their organization. several of our largest customers have deployed our platform globally to their entire workforce following smaller initial deployments. this expansion in the use of our platform also provides us with opportunities to market and sell additional products to our customers, such as zoom rooms, at each office location and enablement of zoom video webinars. in order for us to address this opportunity to expand the use of our products with our existing customers, we will need to maintain the reliability of our platform and produce new features and functionality that are responsive to our customers requirements for enterprise-grade solutions.

Post-Announcement News Sources

https://seekingalpha.com/news/3585127-stay-home-tech-stocks-gain-virus-surges
https://seekingalpha.com/news/3583234-zoom-among-gainers-on-new-coronavirus-concerns

Table 9: Zoom Example

ZILLOW GROUP INC								
File Date	Sort Date	Assigned Quintile by FtBERT	Market Cap ($Mln)	GIC	Next Announcement Date	$R_{December,2020}$	$R_{January,2021}$	$R_{February,2021}$
2020-11-05	2020-11-30	High	8306.07	Real Estate	2021-02-10	0.23	0.02	0.22
Explanation								
Revenues dropped but not as much as expected. Online traffic has surged.								
Text Highlights								
while we have resumed home buying in all zillow offers markets, a decline in home buying and other potential effects of covid-19 on residential real estate transactions may adversely impact the number of homes sold in future periods, which could result in a decline in revenue in future periods.the number of visits is an important metric because it is an indicator of consumers level of engagement with our mobile applications, websites and other services. we believe highly engaged consumers are more likely to participate in our zillow offers program, use zillow homes loans or be transaction-ready real estate market participants and therefore are more sought-after by our premier agent and premier broker real estate partners. we define a visit as a group of interactions by users with the zillow, trulia and streeteasy mobile applications and websites. a single visit can contain multiple page views and actions, and a single user can open multiple visits across domains, web browsers, desktop or mobile devices. visits can occur on the same day, or over several days, weeks or months. zillow and streeteasy measure visits with google analytics, and trulia measures visits with adobe analytics. visits to trulia end after thirty minutes of user inactivity. visits to zillow and streeteasy end either: (i) after thirty minutes of user inactivity or at midnight ; or (ii) through a campaign change. a visit ends through a campaign change if a visitor arrives via one campaign or source (for example, via a search engine or referring link on a third - party website), leaves the mobile application or website, and then returns via another campaign or source. the following table presents the number of visits to our mobile applications and websites for the periods presented (in millions) : three months ended September 30, 2019 to 2020 change 2020 2019 visits 2,786.2 2,104.932 unique users measuring unique users is important to us because much of our revenue depends in part on our ability to connect home buyers and sellers, renters and individuals with or looking for a mortgage to real estate, rental and mortgage professionals, products and services. growth in consumer traffic to our mobile applications and websites increases the number of impressions, clicks, connections, leads and other events we can monetize to generate revenue.								
Post-Announcement News Sources								
https://seekingalpha.com/news/3660960-zillow-jumps-as-drop-in-q4-offers-revenue-is-softened-by-other-double-digit-gains								

Table 10: Zillow Example

HARPOON THERAPEUTICS INC								
File Date	Sort Date	Assigned Quintile by FtBERT	Market Cap ($Mln)	GIC	Next Announcement Date	$R_{June,2021}$	$R_{July,2021}$	$R_{August,2021}$
2021-05-06	2020-05-31	Low	452.70	Health Care	2021-08-05	-0.33	-0.29	-0.03

Explanation

Interim data released from Harpoon's ongoing prostate cancer study were not impressive.

Text Highlights

because the number of qualified clinical investigators and clinical trial sites is limited, we may conduct some of our clinical trials at the same clinical trial sites that some of our competitors use, which could reduce patient enrollment depends on many factors, including the size and nature of the patient population, the severity of the disease under investigation, eligibility criteria for the trial, the proximity of patients to clinical sites, the design of the clinical protocol, the ability to obtain and maintain patient consents, the ability to recruit clinical trial investigators with the appropriate competencies and experience, the risk that patients enrolled in clinical trials will drop out of the trials before the administration of our product candidates or trial completion, the availability of competing clinical trials, the availability of new drugs approved for the indication the clinical trial is investigating, and clinicians and patients perceptions as to the potential advantages of the drug being studied in relation to other available therapies. these factors may make it difficult for us to enroll enough patients to complete our clinical trials in a timely and cost-effective manner. delays in the completion of any clinical trial of our product candidates will increase our costs, slow down our product candidate development and approval process and delay or potentially jeopardize our ability to commence product sales and generate revenue. in addition, some of the factors that cause, or lead to, a delay in the commencement or completion of clinical trials may also ultimately lead to the denial of regulatory approval of our product candidates. our product candidates may have serious adverse, undesirable or unacceptable side effects or other properties which may delay or prevent marketing approval. if such side effects are identified during the development of our product candidates or following approval, if any, we may need to abandon our development of such product candidates, the commercial profile of any approved label may be limited, or we may be subject to other significant negative consequences following marketing approval, if any. undesirable side effects that may be caused by our product candidates could cause us or regulatory authorities to interrupt, delay or halt clinical trials and could result in a more restrictive label or the delay or denial of regulatory approval by the fda or other comparable foreign authorities. our product candidates target protein expression on tumor cells, which expression may also be present on healthy cells. accordingly, our product candidates may result in high or unacceptable levels of toxicity when tested in humans.

Post-Announcement News Sources

https://seekingalpha.com/news/3703658-merus-upgraded-on-asco-update-citi-favors-harpoon-after-selloff-and-more-in-todays-analyst-action
https://seekingalpha.com/news/3703354-harpoon-shares-slide-after-early-stage-prostate-cancer-study-data-fails-to-impress

Table 11: Harpoon Example

ALLOGENE THERAPEUTICS INC								
File Date	Sort Date	Assigned Quintile by FtBERT	Market Cap ($Mln)	GIC	Next Announcement Date	$R_{December,2021}$	$R_{January,2022}$	$R_{February,2022}$
2021-11-04	2021-11-30	Low	2127.94	Health Care	2022-02-23	-0.19	-0.23	-0.20

Explanation

Investors were disappointed to see that the management team has been circumspect about the details of the clinical hold decision by FDA.

Text Highlights

patients may also undergo plasmapheresis to remove rituximab prior to infusion of allo-501, which may cause separate adverse effects. we have removed the rituximab recognition domains in the second generation of all allo-501, known as allo-501a, which we believe will potentially facilitate treatment of patients who were recently treated with rituximab. however, allo-501a may not behave as expected and may be challenging to develop or manufacture. our clinical trials will also compete with other clinical trials for product candidates that are in the same therapeutic areas as our product candidates, and this competition will reduce the number and types of patients available to us because some patients who might have opted to enroll in our trials may instead opt to enroll in a trial being conducted by one of our competitors. since the number of qualified clinical investigators is limited, some of our clinical trial sites are also being used by some of our competitors, which may reduce the number of patients who are available for our clinical trials in that clinical trial site. moreover, because our product candidates represent a departure from more commonly used methods for cancer treatment, potential patients and their doctors may be inclined to use conventional therapies, such as chemotherapy and hematopoietic cell transplantation or autologous car t cell therapies, rather than enroll patients in our clinical trial, including if our product candidates have or are perceived to have additional safety or efficacy risks or if using our product candidates may affect insurance coverage of conventional therapies. patients eligible for allogeneic car t cell therapies but ineligible for autologous car t cell therapies due to aggressive cancer and inability to wait for autologous car t cell therapies may be at greater risk for complications and death from therapy. delays in patient enrollment may result in increased costs or may affect the timing or outcome of our clinical trials, which could prevent completion of these trials and adversely affect our ability to advance the development of our product candidates. the market opportunities for our product candidates may be limited to those patients who are ineligible for or have failed prior treatments and may be small. the FDA often approves new therapies initially only for use in patients with r/r metastatic disease. we expect to initially seek approval of our product candidates in this setting. subsequently, for those products that prove to be sufficiently beneficial, if any, we would expect to seek approval in earlier lines of treatment. there is no guarantee that our product candidates, even if approved, would be approved for earlier lines of therapy, and, prior to any such approvals, we will have to conduct additional clinical trials, including potentially comparative trials against approved therapies.

Post-Announcement News Sources

https://seekingalpha.com/article/4478079-allogene-lifting-of-the-clinical-hold-will-be-a-major-binary
https://seekingalpha.com/news/3786894-allogene-closes-down-9-despite-lifting-of-clinical-hold-other-gene-editing-names-mixed

Table 12: Allogene Example

Can AI Read the Minds of Corporate Executives?

Internet Appendix

Table of Contents:

- Section IA1 provides details on the **baseline machine learning models**.
- Section IA2 provides details on the **setup of the hyperparameter tuning** for the baseline machine learning models.
- Table IA1 presents the **set of hyperparameters** used in the baseline machine learning models.

IA1 Baseline machine learning methods

Linear regression Linear regression is a simple and widely used method, which assumes that $g(\cdot)$ can be approximated by a linear function of the features and the parameter vector, $\boldsymbol{\theta}$,

$$g(\mathbf{z}_{i,t-3}) = \mathbf{z}_{i,t-3}^{\top}\boldsymbol{\theta}. \tag{9}$$

Our baseline linear regression model is estimated by ordinary least squares (OLS) by minimizing

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(r_{i,t} - g(\mathbf{z}_{i,t-3}))^2, \tag{10}$$

which yields the pooled OLS estimator. The estimate of θ in (10) is unbiased and efficient provided that the number of predictors P is small relative to T. In the bag-of-words setting, the document-term matrix is high-dimensional and sparse, resulting in the number of predictors P being similar to or larger than T, and leading to overfitting. To tackle the high-dimensional and sparse nature of the document-term matrix, we consider a variety machine learning methods.

LM OLS Motivated by Jegadeesh and Wu (2013), we regress the normalized ranking score on the LM negative sentiment measure discussed in Section 3.1. This supervised method reduces overfitting by preselecting a single covariate for the regression.

Penalized linear A prevalent method for mitigating overfitting in Equation (10) involves incorporating a penalty term into the objective function. This regularization approach intentionally degrades a model's in-sample performance to enhance its out-of-sample stability. Rather than (10), penalized linear models estimate $\boldsymbol{\theta}$ by minimizing

$$\mathcal{L}(\boldsymbol{\theta};\cdot) = \mathcal{L}(\boldsymbol{\theta}) + \phi(\boldsymbol{\theta};\cdot), \tag{11}$$

where $\phi(\boldsymbol{\theta};\cdot)$ is the penalty function. We use the popular elastic net (EN) penalty, which takes the form

$$\phi(\boldsymbol{\theta};\lambda,\rho) = \lambda(1-\rho)\sum_{j=1}^{P}|\theta_j| + \frac{1}{2}\lambda\rho\sum_{j=1}^{P}\theta_j^2. \tag{12}$$

The elastic net involves two nonnegative hyperparameters, λ and ρ, and includes two well-known regularization methods as special cases. When $\rho = 0$, eq. (11) corresponds to the

Lasso, which can set a subset of $\boldsymbol{\theta}$ to exactly zero, imposing sparsity on the specification, and can thus be thought of as a variable selection method. When $\rho = 1$, eq. (11) corresponds to ridge regression, which shrinks coefficient estimates close to zero, but does not impose exact zeros anywhere. For intermediate values of ρ, the elastic net compromises between shrinkage and sparsity. Tuning parameter λ controls the amount of shrinkage, where larger values of λ correspond to greater amounts of shrinkage.

Boosted regression trees and random forests Regression trees are nonparametric methods that effectively model nonlinearities and interactions among predictors. These trees are constructed by recursively partitioning the input predictor space into a series of distinct regions, and predicting the average value of the response within each partition. The tree growth occurs through a sequence of steps, wherein at each step, a new branch divides the data based on the predictor and split value that minimizes the squared error.

Regression trees, while flexible, are particularly prone to overfitting, which necessitates regularization to improve their predictive performance. In this study, we examine two tree ensemble methods that achieve regularization by combining forecasts from multiple trees to produce a single prediction.

The first ensemble method, random forest (RF), builds a collection of decorrelated trees and averages their predictions. Each individual tree is trained on a bootstrap sample of the data, and at each branch, only a random subset of predictors is considered for splitting. This process results in a set of uncorrelated trees, each with high variance. However, by averaging the predictions across multiple trees, the variance is reduced, yielding a more stable algorithm.

The second ensemble method, gradient boosted regression trees, constructs a series of decision trees sequentially, with each tree learning from the residuals of its predecessor. Boosting recursively combines the forecasts from numerous shallow trees, which individually function as weak learners with limited predictive power. However, when combined sequentially, they form a more stable and accurate model. In this study, we employ the XGBoost (XGB) implementation of gradient boosting (Chen and Guestrin, 2016), which integrates a more efficient optimization algorithm and additional regularization techniques to prevent overfitting.

Support Vector Regression Support Vector Regression (SVR) is a technique that performs well for high-dimensional data, such as document-term matrices (Manela and Moreira, 2017). Unlike OLS, which minimizes the mean squared error, SVR minimizes the following objective function:

$$H(\boldsymbol{\omega}) = \frac{1}{2}\|\boldsymbol{\omega}\|^2 + \frac{C}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} h_\epsilon(r_{i,t} - g(\mathbf{z}_{i,t-3})),$$

where $h_\epsilon(e) = \max\{0, |e| - \epsilon\}$ is an ϵ-insensitivity margin. SVR fits the best hyperplane within the ϵ-insensitivity margin, which is a buffer zone around the predicted output line, where errors of size less than ϵ are ignored. C is a hyperparameter that helps to regularize the estimated weights and avoid overfitting.

To manage the non-linear transformation of data, SVR actively employs kernels, such as the radial basis function (RBF). These kernels facilitate the transformation of data into higher dimensions, which in turn enables the algorithm to find a fitting hyperplane in this newly transformed space. The "kernel trick" allows the algorithm to operate within this transformed space without the need for explicit computation of the data coordinates, making the problem computationally tractable even for high-dimensional data.

Feed-Forward Neural Networks We include a traditional feed-forward neural network (NN) as a straightforward machine learning benchmark. Feed-forward networks comprise an input layer with raw features, one or more hidden layers that interact with and nonlinearly transform predictors, and an output layer that consolidates hidden layers into a prediction. We utilize a shallow neural network architecture featuring a single hidden layer containing 32 units. The rectified linear unit, defined as $\text{ReLU}(z) \equiv \max(z, 0)$, serves as the nonlinear activation function. To prevent overfitting due to the high parameterization of neural networks, we apply regularization techniques, including a ℓ^1 penalty on weights and batch normalization.

IA2 Fine-tuning

For the linear Lasso and EN regressions, tuning parameter λ is the main hyperparameter that determines the level of shrinkage. This hyperparameter controls the trade-off between model complexity and the degree of regularization applied to the model coefficients. By adjusting the value of λ, we can balance the bias-variance trade-off and minimize overfitting.

For the XGB and RF models, the main hyperparameters include the number of trees, the maximum depth of each tree, and the shrinkage parameter λ (only for XGB). Increasing the number of trees can improve the model's predictive performance but may also increase the risk of overfitting. Deeper trees can capture more complex interactions between features, but they can also lead to overfitting. A smaller value of λ results in a more conservative model with a lower risk of overfitting but may require more iterations to converge.

In SVR, the main hyperparameter C regularizes the weights. A large value of C will lead

Lasso	$\lambda \in (10^{-5}, 10^{-2})$, $\alpha = 0$
EN	$\lambda \in (10^{-5}, 10^{-2})$ $\alpha = 0.5$
XGB	#trees = $100 \sim 300$, tree depth = $1 \sim 4$ learning rate = {0.001, 0.01} sample rate = 0.5
RF	#trees = $100 \sim 300$ tree depth = {20, 50, 100}
SVR	Kernel = {Linear, RBF} $C = \{0.01, 0.1, 1, 10\}$
NN	Activation Functions = Relu batch size = 32 epochs = 100 learning rate = {0.001,0.01} $\ell^1 = (10^{-9}, 10^{-7})$

This table describes the set of hyperparameters that are tuned for each baseline machine learning model considered in the paper.

Table IA1: Baseline Machine Learning Model Hyperparameters

the optimization to prioritize fitting the training data, while a smaller C value allows for more errors but prevents overfitting by maintaining smaller weights.

Neural networks require choosing many hyperparameters to find the best model flexibility. We guard against model overfitting by adding an ℓ^1 penalty to the loss function with various degrees of shrinkage. This penalty encourages the model to learn a sparse representation of the features, reducing the complexity of the learned relationships and minimizing overfitting. We also choose the initial learning rate for the Adam stochastic gradient descent optimizer (Kingma and Ba, 2015). A smaller learning rate results in a slower but potentially more accurate convergence, while a larger learning rate can speed up convergence but may overshoot the optimal solution.

Table IA1 provides a summary of the hyperparameters tuned for each model considered in this paper. In order to select the optimal combination of hyperparameters at a given time, the data sample is partitioned into training, validation, and testing sets as described in Section 2. The set of hyperparameters with the lowest MSE in the validation set is selected as the best model. In line with conventional machine learning practices, we standardize each variable in the training, validation, and test sets using its mean and variance from the training set.

IA3 Extracting subsections

The Risk Factor section typically begins with the header "Item 1A. Risk Factor". However, there are many non-standard occurrences in the documents. To identify exceptional cases, we first browse through the non-standard cases to understand what constitutes an exceptional case. We then use a regular expression that is flexible enough to match most section starts. Specifically, we look for expressions of the form (ignoring cases) "item" + numbers and/or letters + "risk"+ "factor". We allow spaces, punctuation (e.g., period and hyphen), and line breaks to appear between words. This allows us to match any of the following non-standard occurrences: "Item 1 Risk Factor", "item 1. Risk \n factors", "item riskfactor", "item 1a-risk factor", etc.

The MD&A section typically starts with the header "Item 2. Managements' Discussion and Analysis". We look for expressions of the form "item" + numbers + "management"+ "s"+"discussion and analysis". Similar to the RF section, we ignore cases and allow spaces, punctuation, and line breaks to appear between the words.

We also ensure that the matched cases are stand-alone phrases, i.e., not part of a sentence, since sections may be referred to in the text of other sections. For example, "please refer to Item 1A. Risk Factor for more details" could be a potential match if such conditions are not imposed.

To identify section ends, we look for the next section header, which is the first appearance of a stand-alone phrase that starts with "item"+$*$ in the remaining of the file. We require that the matched section ends cannot be a complete sentence to avoid mismatches of in-text references.

Once we have identified both section header and section end, we extract the text between the header and the ends as the section main text. We discard matched cases if the main text is 250 words or less, as it could be from the table of contents. If we still have more than one matched case for a given section, we join them together. This is necessary since sections may span several pages, and the same section header may appear on every page that it spans. By combining the matches cases, we maximize our chance of extracting the entire section.