Measuring Readability with Language Predictability: A Large Language Model Approach

Abstract

We introduce a new readability measure, the language predictability score (LPS), to assess the processing costs of corporate disclosure texts. Grounded in psycholinguistics and cognitive science, LPS differs from conventional measures, such as the Fog index and file size, by considering word context and target audience. We employ large language models to estimate LPS for management discussion and analysis sections from annual reports. Validity tests show that LPS effectively identifies incoherent text as less readable. Main tests show that LPS correlates significantly with measures of processing costs for market participants, including financial analysts.

Measuring Readability with Language Predictability: A Large Language Model Approach

1. Introduction

Despite the emphasis on quantitative information in accounting and finance research, most companies and market participants communicate using natural language through channels such as company filings, conference calls, news releases, analyst reports, social media, and regulatory announcements. Market participants, ranging from small investors to sophisticated intermediaries, likely incur significant costs in extracting useful information from text-based sources and integrating it into asset valuations and other decision-making. These costs potentially have important implications for a wide range of capital market outcomes and corporate behaviors. The advancement of technologies in recent years has resulted in an explosion of studies involving textual analyses. However, as discussed by Loughran and McDonald (2020) and Blankespoor et al. (2020), the current understanding of how various market participants process qualitative disclosure, as well as their processing costs, remains rudimentary and lacks theoretical guidance.

One of the most important yet controversial components of the processing cost of qualitative information (i.e., text) is its "readability." Intuitively, readability gauges the amount of effort required for a reader to comprehend written text. However, when it comes to financial documents, measuring readability is "a challenging endeavor that regulators and academics have been struggling with for decades," making it the "greatest concern in the evolution of textual methods in the finance and accounting disciplines" (Loughran and McDonald, 2020, pp. 365–366; hereafter, these scholars are referred to as LM).

In this study, we propose and validate a new readability measure based on the concept of language predictability that has strong theoretical roots and experimental support in

psycholinguistics and cognitive science. According to surprisal theory (Hale, 2001; Levy, 2008), predictable words are easier to process because readers actively anticipate what comes next and integrate them into the preceding context more easily than unexpected words. Thanks to the advancement of large language models (LLMs), measuring language predictability for a large sample of text has become technologically implementable. Using a large sample of Management Discussion and Analysis (MD&As), one of the most important sections in annual reports, wherein firms summarize their current performance and discuss their strategies and outlook, we demonstrate that language predictability captures the processing costs of market participants and is distinct from the existing readability measures.

Two conceptually different readability constructs are in common use in the extant literature. The first type are readability formulas that involve slightly different combinations of word length and sentence length, such as the Fog index and the Flesch-Kincaid Index.[1] Many linguistic studies point out that these formulas developed to determine the grade level of texts for elementary and secondary graders, are methodologically inaccurate for assessing readability because longer sentences are not necessarily more complex syntactically or more difficult to comprehend (e.g., "I think, therefore I am."), and longer words are not necessarily semantically more difficult or uncommon (e.g., "corporations," and "telecommunication") (e.g., Davison and Kantor, 1982). Applying these formulas on financial documents, the target audience of which consists mostly of professionals with specialized domain knowledge is even more problematic (LM, 2014, 2020). In fact, in its guidance on how to create readable corporate disclosures, the Securities and Exchange Commission (SEC) recommends against using readability formulas,

[1] LM (2020) show extremely high correlations (between 0.84 and 0.99) among the following readability formulas, the Fog index, Flesch–Kincaid Index, LIX (Läsbarhetsindex), RIX (Rate Index), ARI (Automated Readability Index), and SMOG (Simple Measure of Gobbledygook).

pointing out that "Einstein's theory of relativity reads at a 5th grade level" (p. 57).[2] The second type of measures are based on the total quantity of information, such as file size and word counts (e.g., Li, 2008; You and Zhang, 2009; LM, 2014). They could be conceptually inconsistent with readability because a longer document can be written clearly and be easy to read, and because document length likely reflects firm size and reporting rules (Blankespoor et al., 2020).

"What is readability?" LM (2020) pose the question after carefully evaluating the existing readability measures and call for a better readability concept suitable for financial documents. We believe that the key to answering this question is recognizing that the reading process is an interaction between readers and the text. Hence, readability—the ease with which a reader comprehends the meaning of a text—depends on the reader's language ability, background knowledge, and purpose in reading the text as well as the author's language ability, background knowledge, and the purpose for writing the text. We argue that by not considering the reader-text interaction, the existing measures mistakenly assume that comprehension of the text depends solely on the text itself. Therefore, we agree with the conclusion of LM (2020) that a conceptually appropriate measure of "readability must be designed in the context of the targeted audience." In contrast with researchers' neglect, the SEC (1998) advises firms prominently in *A Plain English Handbook* that "knowing your audience is the most important step in assuring that your document is understandable to your current or prospective investors" (p. 9).

Therefore, we propose a new readability measure based on the concept of language predictability, namely, the degree of alignment between the language being read and the

[2] In other attempts to improve the measurement of readability, Bonsall et al. (2017) suggest using the Bog index, a readability formula based on a proprietary software, *Stylewriter*, which produces an array of writing style statistics, such as sentence length and the use of passive verbs and abstract words according to a private word list. However, as LM (2020) discuss, the proprietary nature of the Bog index limits its replicability and insights, and it mainly captures the writing style of a text which might not correspond to the ease of comprehension.

language being anticipated by the reader. The underlying intuition is straightforward: When the word, being read as a continuation of the preceding words (i.e., its context), is anticipated by the reader, he/she understands it with little effort; conversely, when it deviates from the reader's anticipation—that is, when this word results in a Bayesian surprise to the reader and triggers belief updating—comprehension becomes slower and more difficult (Anderson, 1990; Doya et al., 2007; Levy, 2008). DeLong et al. (2005) demonstrate this intuition in their seminal neural experimental study, in which they observed more intensive brain activities when readers encountered a more unexpected phrase, "an airplane," than "a kite" in finishing the sentence "The day was breezy so the boy went outside to fly…" [3] We provide more examples from our sample MD&As to illustrate predictable and unpredictable words in Appendix C.

Note that many factors potentially influence the misalignment, such as logic flow, text cohesion, syntactic complexity, grammatical mistakes, as well as the mismatch between readers and the writer in language abilities, domain knowledge, cultural backgrounds, habits of expression, and the purpose of reading/writing the text (Kutas et al., 2011). Importantly, measuring readability as language predictability aggregates all these factors. This conceptual construct is developed from well-established theories from several related or overlapping research areas in linguistics, psycholinguistics, cognitive science, and computational linguistics,[4] and is supported by ample behavioral and neural experimental evidence (e.g., Rayner and Well, 1996; Boston et al., 2008; Smith and Levy, 2013) (discussed in detail in Section 2).

[3] Delong et al. (2005) measure the brain activity level using N400 amplitude, which is one specific type of brain response that can be observed at the scalp by electroencephalography (EEG).

[4] Specifically, they include theories on the cloze procedure and probability (e.g., Taylor, 1953), surprisal theory (e.g., Shannon, 1948; Hale, 2001; Levy, 2008), theories and models of language predictability (see the review by Kutas et al., 2011; also Kuperberg and Jaeger, 2016), and n-gram language models (e.g., Bengio et al., 2003; Devlin et al., 2019; Wilcox et al., 2020).

Although language predictability, as a measure of readability, is conceptually appealing, the key challenge is how to operationalize it. We rely on surprisal theory, which posits that the reading difficulty is proportional to *Surprisal*, the negative log probability of word w conditional on its context c (i.e., $-\log p(w|c)$) and calculate language predictability as $(-1) \times$ *Surprisal* (Hale, 2001; Levy, 2008; Smith and Levy, 2013). Notably, its simple mathematical formula, $\log p(w|c)$, shows that language predictability addresses two disadvantages that are too important to be overlooked in the existing readability measures (e.g., the Fog and Bog indexes and file size). First, the existing measures ignore words' context while language predictability captures the intuition that words are easier to comprehend in contexts in which they are more predictable. Second, the existing measures ignore a target reader's specific characteristics, such as language ability, domain knowledge, cultural background, and habits of expression, all of which influence this reader's anticipation (i.e., $p(w|c)$).

Early linguistic studies measure predictability of a word (i.e., $p(w|c)$) using a Cloze procedure (Taylor, 1953), in which human subjects are asked to guess a masked word in context c. Next, $p(w|c)$ is calculated as the proportion of the participants selecting word w out of the total number of participants. However, this procedure is prohibitively expensive and nonviable for a large sample of text data.

Fortunately, obtaining $p(w|c)$ with an LLM that imitates human language comprehension—that is, asking AI to perform a cloze task—has become technologically feasible thanks to the recent advancement in LLMs (Devlin et al., 2019; Radford et al., 2019). By training on a vast corpus followed by fine-tuning using a domain-specific corpus, such as MD&As in our study, an LLM acquires a profound understanding of human language as well as world knowledge and domain-specific knowledge by learning the syntactic and semantic patterns of language in the

specific domain. Recent research in psycholinguistics has documented that the predictability of words estimated from LLMs is highly correlated with human expectations of upcoming words and, thus, is able to capture the difficulty in human language comprehension (Wilcox et al., 2020; Michaelov et al., 2021, 2022). We apply an open-source LLM, GPT-2, which was developed by OpenAI (Radford et al., 2019), to estimate the predictability of the words in MD&As.[5] GPT-2 is pre-trained by OpenAI on an extensive corpus of text from about 8 million web pages. We conduct fine-tuning using MD&A data to adapt the GPT-2 model to the MD&A corpus (described in detail in Appendix D). The fine-tuned GPT-2 models estimate the predictability of a given word in an MD&A as the logarithm of the likelihood of its occurrence given the preceding neighboring 256 words, which we consider the word's context. Finally, the language predictability score of an MD&A (denoted LPS) is calculated as the mean of the predictability of individual words in the MD&A (see Appendix B for an illustrative example).

We employ several approaches to validate our language predictability measure as a readability measure appropriate for corporate disclosures. First, our fine-tuned GPT-2 models exhibit human-like reading ability by "guessing" the next word correctly about 50% of the time while "reading" MD&As, which is equivalent to achieving about 90% comprehension (Bormuth, 1968). Second, we create a sample of incoherent MD&A text by randomly shuffling the pairing between the focal sentences and their contexts in the original MD&A text (i.e., coherent text), and our fine-tuned GPT-2 models accurately identify the falsified MD&A (i.e., incoherent text) as unreadable (in the bottom quintile of our entire sample). Notably, the existing readability

[5] We use GPT-2 because OpenAI does not provide open access to the features in its later versions that are required for the fine-tuning by our study. We also use the bidirectional encoder representations from transformers (BERT) model developed by Devlin et al. (2019) and pre-trained by Google, as an alternative LLM to estimate language predictability.

measures (the Fog index, the Bog index, and file size) cannot distinguish incoherent text from coherent text because both consist of the same words and sentences. Third, our measure identifies the repetitive and generic boilerplate sentences as extremely predictable (above the 95th percentile in our sample), which implies a low processing cost. This contrasts with the existing readability measures, which often categorize these sentences as unreadable because of their use of polysyllabic words and greater sentence length. Overall, the results of our validation tests demonstrate that LPS captures reader-text interactions as predicted by the theories and highlight the shortcomings of existing readability measures in ignoring human reading processes.

We also explore how the characteristics of a firm's business environment correlate with the language predictability of its MD&A text. We find that firms with lower earnings and a higher number of special items, those undergoing changes in business segments, and those with greater operational complexity tend to use less predictable language in their MD&As. Firms that use more forward-looking statements in their MD&As also contain less predictable language. This evidence suggests that the language predictability of MD&As reflects the information complexity of public firms' corporate disclosures.

We then investigate the effect of the language predictability of MD&As on the information environment after the 10-K filing date. We follow LM (2014) and Bonsall et al. (2017) and use stock market volatility and analysts' forecast dispersion and forecast accuracy to capture the post-filing information environment and examine their associations with LPS. In our empirical analyses, we include a variety of controls for industry and firm characteristics and year fixed effects that have been examined in the prior literature. Importantly, to mitigate the endogeneity concerns raised by Li (2010) and Bonsall et al. (2017) that time-invariant firm characteristics may drive the documented associations between the existing readability measures and processing

costs (Li, 2008; LM, 2014), we estimate the models with firm fixed effects. We find that lower language predictability in an MD&A is associated with higher post-filing stock return volatility, higher analyst forecast dispersion, and decreased forecast accuracy post-filing. These findings suggest that language predictability affects the processing costs of not only average market participants but also sophisticated information intermediaries. Notably, Bonsall et al. (2017) find that readability measured using the Fog and Bog indexes and file size is insignificant in explaining post-filing analyst forecast dispersion and accuracy and interpret the results as readability having no influence on more sophisticated market participants. By contrast, we find strong and consistent results that readability, when measured as language predictability, has a significant impact on the processing costs of sophisticated market participants. Therefore, we argue that the insignificant results from the Fog and Bog indexes and file size can be attributed to the failure of these measures to capture the target readers' characteristics, a particularly salient issue when readers are sophisticated or have domain-specific knowledge.

This is also the first study to disentangle the readability of boilerplate language from that of non-boilerplate language. Our findings show that analysts' forecast dispersion and accuracy are not affected by the readability of boilerplate sentences but both are influenced by the readability of non-boilerplate language, indicating that sophisticated market participants, such as analysts, are able to identify boilerplate content and recognize its irrelevant nature. This finding helps answer a question raised by Miller (2010) and Blankespoor et al. (2020) about whether sophisticated users are affected by boilerplate information.

Our main contribution is simple and focused. We respond to the pressing need in the accounting and finance research for a readability measure that is conceptually appropriate and theoretically grounded for capturing the processing costs of the target readers of corporate

disclosures and other business documents (Blankespoor et al. 2020; LM 2020). We introduce such a readability measure, one based on the concept of language predictability, which has strong theoretical underpinnings and experimental evidence supporting it. Our measure recognizes reader-text interaction, and overcomes the limitations of the existing readability measures (i.e., Fog index, Bog index, and file size), including their failure to capture target audience or consider the context of the words being read. Our measure is also technologically convenient for researchers to apply since it is computed by open-source and unsupervised LLMs. Hence, our study has important implications for the research community, corporations as information providers, and the SEC in promoting understandable and readable disclosures.

This study also contributes to the literature by providing new insights on how investors process qualitative disclosures. First, when corporate disclosures contain language that is unexpected by investors, they find the disclosures costly to process. Second, when firm fixed effects are included, we find that readability measured as language predictability explains analysts' processing costs while the Fog and Bog indexes and file sizes become insignificant. These results suggest that language predictability of corporate disclosures impacts sophisticated investors beyond the time-invariant complexity of the firm. Third, boilerplate disclosures may not affect sophisticated market participants' information processing costs because such investors are familiar with boilerplate language due to its repetitive nature.

Lastly, we contribute to the emerging stream of research on using LLMs to simulate certain human subgroups of interest. For example, Rozado (2024) demonstrates that by fine-tuning LLMs with modest amounts of textual content from left-leaning and right-leaning publications, these models can successfully pass the political orientation tests designed to assess human subjects' political preferences. Similarly, Garimella,et al. (2022) utilize blog posts and Reddit

comments authored by individuals from various genders and racial backgrounds to finetune LLMs, revealing that these models exhibit social biases reflective of the demographic groups. By showing that fine-tuning GPT2 with MD&A content can align the LLMs with sophisticated market participants, such as analysts who are familiar with corporate disclosures, we illustrate the potential of LLMs to enhance our understanding of the behaviors exhibited by specific subgroups of market participants.

2. Related literature

Most corporate disclosures are qualitative rather than quantitative. Research on qualitative disclosures has increased exponentially in recent years (see reviews by Li, 2010; LM, 2016, 2020; Gentzkow et al., 2019). The investing community is increasingly aware of the information contained in qualitative disclosures and uses it to value firms. Although corporate textual disclosures are generally accessible to the public, it is costly for market participants to process them (Liberti and Petersen, 2019; Blankespoor et al., 2020). Because market participants are constrained in their attention and expertise, the tasks of reading a text, comprehending its content, and evaluating its usefulness given their knowledge base are costly. However, as discussed by Blankespoor et al. (2020) and LM (2020), the extant literature on the processing cost of qualitative information is nascent and limited to several controversial measures of readability including the Fog index (and the modified Fog index), the Bog index, and file size (Li, 2008; LM, 2014; Bonsall et al., 2017). In this section, we review the existing readability measures, discuss their limitations, and introduce the theories underpinning our proposed measure of readability, the language predictability score (LPS).

2.1. Existing measures of readability for textual disclosures

There is a pressing demand in accounting and finance research for an appropriate proxy to measure the readability of corporate textual disclosures. The seminal work of Li (2008) measures the readability of firms' annual reports using the Fog index, developed by Gunning (1952), which is the sum of sentence length and the percentage of words with three or more syllables. Li (2008) argues that a higher Fog index indicates less readable text and finds a negative relationship between the Fog index and a firm's earnings performance. Subsequently, the Fog index has been widely used by researchers to measure the readability of business texts (e.g., Biddle et al., 2009; Lehavy et al., 2011; Dougal, et al., 2012; Bushee et al., 2018).

LM (2014, 2020) and others argue that the Fog index, developed in 1950s and originally intended for as assessing reading materials for school graders, is a poor measure of readability for corporate textual disclosures because it incorrectly assumes any words containing more than three syllables are complex. As such, text containing words such as "corporation" and "telecommunication," which are common in the business domain and easily understood by market participants, would be mislabeled as complex.

Instead, LM (2014) advocate measuring readability as the quantity of text in a financial document. Specifically, they recommend using the file size of the "complete submission text file" available on the SEC's EDGAR website as a proxy for the readability of a 10-K document. File size is an indirect measure of readability, because it focuses on the total amount of both quantitative and qualitative information released by the firm, rather than the challenges that investors encounter when trying to comprehend a text. Moreover, Bonsall et al. (2017) show that the variation in 10-K file size is driven by the file types (e.g., HTML, XML, or PDF) and the inclusion of content unrelated to the amount of text (e.g., tables and pictures). File size also can

be influenced by the increasingly large amount of boilerplate in 10-Ks, which investors tend to ignore because of its repetitive nature (Lang and Stice-Lawrence, 2015; Huang et al., 2021).

Bonsall et al. (2017) propose the Bog index as an alternative measure of readability. Bonsall et al. obtained the Bog index from a proprietary software, *StyleWriter*, which produces an array of writing style statistics, including sentence length and the use of passive verbs and abstract words according to a private word list. However, LM (2020) argue that the Bog index is intended to capture writing style rather than readability and that the lack of transparency in *StyleWriter's* proprietary list poses challenges for replication. LM further argue that *StyleWriter* suffers from mis-specification when applied to business texts. For instance, it categorizes some of the most frequently used words in financial disclosures, such as "approximately" and "generally", as less familiar to readers. Hence, the Bog index suffers from a similar problem as the Fog index; that is, neither sufficiently accounts for readers' existing knowledge.

Therefore, the need persists in the accounting and finance fields for a conceptually accurate and direct measure of readability for financial documents, such as 10-K or analyst reports. LM (2020) argue that such a measure must be designed to suit the target audience of the texts, who tend to possess specialized vocabularies, financial expertise, and investing experience.

2.2. The surprisal theory and language predictability

The broader literature on readability, including research in linguistics and cognitive science, has shown that formula-based readability measures such as the Fog index fail to capture the reader-text interaction. This interaction involves readers bringing their vocabularies, background knowledge, and purpose for reading to a particular text (Taylor, 1953; Bruce et al., 1981; Davison and Kantor, 1982; Stevens et al., 1992). It is to address this concern that we introduce

the language predictability score (LPS) as a new measure of readability. In this section, we review and discuss the theory underlying this measure, namely, surprisal theory.

Surprisal theory is rooted in psycholinguistics and cognitive science. Studies in these fields present strong evidence that text processing is both incremental and anticipatory. It is incremental in that readers integrate the meaning of a word or phrase as soon as it is encountered, using the context formed by preceding words (e.g., Tanenhaus et al., 1995).[6] In other words, readers do not "buffer" a word or phrase in memory and defer their interpretation at the end of the sentence. Text processing is also anticipatory in that readers actively anticipate upcoming words based on the context formed by previously encountered words (e.g., Eberhard et al., 1995; DeLong et al., 2005).[7] Such experimental evidence in psycholinguistic and cognitive science laid the foundation for surprisal theory.

According to surprisal theory, predictable words are easier to process because readers actively anticipate what comes next and integrate them into the preceding context more easily than unexpected words. Conversely, when highly anticipated words are not encountered, some additional processing may be triggered to reanalyze, revise, or override an accruing contextual representation, and the brain may even engage in learning a new language pattern (Hale, 2001; Levy, 2008; Kuperberg and Jaeger, 2016). Surprisal theory quantifies the degree of unexpectedness of a word as $Surprisal$, defined as $-\log p(w|c)$, where $p(w|c)$ is the conditional probability of word w given its context c. Various studies have shown $Surprisal$ to

[6] Tanenhaus et al. (1995) conducted an experiment in which individuals' eye movements were monitored as they looked at a visual scene while simultaneously listening to a set of instructions that referred to objects in that scene. They found that participants made saccadic eye movements to depicted objects without delay upon hearing the relevant input (often prior to the referent word itself), rather than after the instruction ended.

[7] DeLong et al. (2005) find that the magnitude of the N400 effect is inversely correlated with a word's cloze probability, suggesting that the information in words with a high cloze probability is anticipated before the participants read them.

be positively and linearly correlated with the amount of time it takes to read the word w, based on self-paced reading time, gaze duration in eye-tracking studies, and the N400 measures in EEG studies (e.g., Zakaluk and Samuels, 1988; Smith and Levy, 2013; Frank, et al., 2015; Goodkind and Bicknell, 2018). For example, Rayner and Well (1996) and Boston et al. (2008) conduct experiments tracking eye movements during various reading tasks and find that the reading times for words with high predictability are significantly shorter than those for words with low predictability.

Note that, because $Surprisal$ is inversely correlated with readability, we use language predictability calculated as $(-1) \times Surprisal$ to measure readability. To illustrate intuitively the relation between processing cost and language predictability, we provide in Appendix C examples in which the words encountered are predictable or unpredictable from our MD&A sample.

2.3. Methods for measuring language predictability in the extant literature

The key challenge in operationalizing language predictability is how to measure the predictability of individual words in context based on the intended audience. In one early and widely used method in the literature, a cloze task (Taylor, 1953), human participants are asked to guess the masked word in a sentence. The estimated predictability of the masked word is calculated as the proportion of the participants selecting this word as the continuation out of the total number of participants. However, the cloze task can be expensive for large datasets, which may require manual guessing of millions of words. Additionally, it may be subject to human biases (Smith and Levy, 2013) and may not provide reliable estimates for infrequent words (Levy, 2008).

An alternative to the cloze task is to estimate word probabilities statistically using n-gram language models, in which the conditional probability of a word given the preceding words in a specific corpus is calculated as follows:

$$P(w_i|w_{i-1}, \dots, w_{i-n}) = \frac{Count(w_i, w_{i-1}, \dots, w_{i-n})}{Count(w_{i-1}, \dots, w_{i-n})},$$

where $P(w_i|w_{i-1}, \dots, w_{i-n})$ is the ratio of the frequency of the sequence of words $(w_i, w_{i-1}, \dots, w_{i-n})$ to that of the sequence of words $(w_{i-1}, \dots, w_{i-n})$, and $Count(\cdot)$ refers to the number of specific sequence of words appearing in the corpus. While this method is mathematically accurate and straightforward, it normally uses a small n to represent context and ignores information outside the n-gram window. It becomes infeasible when n is sufficiently large to capture a meaningful context because of the curse of dimensionality—that is, a particular sequence of words, $(w_i, w_{i-1}, \dots, w_{i-n})$, can be too sparse even in a large corpus of text data.

With advancements in computational linguistics and machine learning in recent years, the research on language predictability has adopted deep learning-based language models to estimate the predictability of words, thereby alleviating the curse of dimensionality (Bengio et al., 2003).

2.4. Measuring language predictability using large language models

Using LLMs to estimate language predictability offers two benefits relevant to our goal of developing a conceptually appropriate readability measure for corporate textual disclosures. First, LLMs alleviate the curse of dimensionality and achieve higher accuracy in predicting the next word given its context. Second, they capture the readers' background knowledge by training on massive amounts of general human text and then fine-tuning with domain-specific text.

The seminal work by Bengio et al. (2003) developed the first neural-network-based language model (NNLM) and made important progress in alleviating the curse of dimensionality by using

fixed-length vectors called word embeddings to represent words. For example, assuming a vocabulary size of 10,000 words and n as the number of preceding words constituting the context, the dimensionality of an n-gram model is $(10{,}000^n - 1)$. By using fixed-length vectors of 30 adopted by Bengio et al. (2003) to represent each word, the dimensionality of an NNLM is reduced to $(30^n - 1)$. Bengio et al. also show that the accuracy of predicting the next word using an NNLM is greatly improved by the longer contexts afforded by reduced dimensionality.

Subsequent studies further improved NNLM performance using deep learning techniques. As the latest advancement in NNLMs, LLMs are typically built with a transformer architecture and trained on a massive amount of text data with millions or even billions of parameters (Vaswani et al., 2017). These models are designed to capture complex syntactic and semantic patterns in language, achieving performance on various language tasks similar to that of humans, including language generation, question answering, chatbots, and text summarization.

Several psycholinguistic and cognitive studies compare the language predictability derived from various n-gram models, NNLMs, and LLMs and show that the predictability of words estimated from LLMs has the highest correlation with human subjects' word processing time, as assessed by gaze duration via eye-tracking, self-paced reading time, and N400 neural responses (Goodkind and Bicknell, 2018; Wilcox et al., 2020; Michaelov et al., 2021). This evidence suggests that the language predictability estimated using LLMs has the potential to serve as a more accurate measure of readability than that based on other language models.

In addition to the benefit of improved accuracy in predicting language, LLMs also capture the background knowledge of the target audience. In our setting of corporate disclosures, a typical reader has general knowledge and domain-specific knowledge, including business knowledge, financial vocabulary and knowledge, industry knowledge, and even firm-specific

knowledge. As discussed, readability depends crucially on the reader–text interaction, specifically, how quickly readers with specific background knowledge predict the next word in its context. However, capturing human background knowledge has been virtually impossible until the emergence of LLMs.

In the first stage, LLMs are pre-trained on massive amounts of textual documents to optimize millions to trillions of parameters. For example, Google's BERT was pre-trained on a large corpus of over 3 terabytes of text data from books and web pages. OpenAI's GPT-2 was pre-trained on a dataset of more than 40 gigabytes of text data from various sources, including books, articles, and websites. Recent studies show that, after pre-training, the parameters of these LLMs appear to store vast amounts of knowledge similar to the human knowledge base, including factual knowledge, relational knowledge, and common sense (Petroni et al., 2019; Jiang et al., 2020). For instance, OpenAI's latest LLM, GPT-4, demonstrated performance in the 90th percentile on the Uniform Bar Exam, required in the United States to become a licensed attorney, and achieved success in multiple advanced placement exams, including on microeconomics and macroeconomics (OpenAI, 2023).

After pre-training, LLMs can be fine-tuned with a smaller, domain-specific dataset. This process adjusts the parameters of the pre-trained models based on the new dataset while retaining the general language understanding gained through the pre-training. Fine-tuning enhances LLMs' performance on specific language tasks according to the specific dataset. For example, a recent study showed that ChatGPT4 pass a CPA exam after fine-tuning, for which the model had failed without fine-tuning (Eulerich et al., 2023). Singhal et al. (2023) show that fine-tuning substantially enhances the performance of Google's PaLM LLM in answering questions based on the U.S. Medical Licensing Exam, nearly matching the performance of human clinicians.

In sum, we are confident that after pre-training and fine-tuning, the LLMs in this study will be able to simulate the understanding of domain-specific knowledge possessed by intended audiences of corporate disclosures.

3. Measuring language predictability of MD&As using large language models

3.1. The GPT-2 large language model

We apply one of the most advanced LLMs in the extant literature, GPT-2, which was developed by OpenAI (Radford et al. 2019), to measure the language predictability of words in corporate disclosures.[8] GPT-2 is based on the Transformer architecture (Devlin et al. 2019), which enables language models to capture the intricate relationships among words and generate more accurate representations of natural language. GPT-2, with 1.5 billion parameters, is pre-trained by OpenAI on an extensive corpus of about 8 million web pages. As mentioned, pre-training enables GPT-2 to acquire a profound understanding of natural language, which can then be fine-tuned with a domain-specific training dataset to imitate the language ability of a target audience. Upon its release, GPT-2 achieved state-of-the-art performance on a wide range of natural language processing (NLP) tasks. Recent studies have demonstrated the superiority of the Transformer-based large language model to other popular deep-learning language models (e.g., recurrent neural network and long short-term memory models) in measuring language predictability in psycholinguistic experiments (Wilcox et al., 2020; Michaelov et al., 2021).

We obtain the pre-trained GPT-2 model from OpenAI and fine-tune it using the textual data from our MD&A sample. The parameters of the pre-trained model are further optimized during the fine-tuning process. As discussed in Section 2.4, fine-tuning allows GPT-2's parameters to

[8] Our fine-tuning and LPS estimation require access to the parameters and internal functions of the LLM model. We use GPT-2 because OpenAI does not open such access for its later versions.

adapt to domain-specific knowledge. For each calendar year, we employ the MD&As from the preceding five years (excluding the current year) as the training data to fine-tune the GPT-2 model. This approach aims to mimic the language understanding of investors, whose domain-specific knowledge is likely acquired from the recent past and evolves over time.[9] By excluding the current year, we prevent the model from utilizing in-sample data to generate predictions. Appendix D provides technical details about the fine-tuning procedure.

3.2. Dataset construction

We start our dataset construction by downloading all 10-K filings available in EDGAR from 1995 to 2018, which include 199,965 10-K filings.[10] We keep only the first filing of each firm in each calendar year (dropping 10,561 filings) and require the filing to be issued at least 180 days after the prior filings (dropping another 481). We further eliminate 47,983 filings that either do not contain an MD&A section or do not use a standard format that allows the computer program to identify the MD&A content. We also require at least 250 words in the MD&A section (dropping 4,912 observations) following LM (2011). Because we use the prior five years of MD&As as training datasets to fine-tune the GPT-2 model for the current year, we cannot calculate the language predictability measures for the first five years of our sample (i.e., 1995–1999), resulting in the loss of another 23,601 observations. Our final sample thus contains 112,427 MD&As with available language predictability measures.

[9] We also tried fine-tuning the GPT-2 model using a shorter window of three years and obtained qualitatively similar results in the subsequent analyses.

[10] The Notre Dame Software Repository for Accounting and Finance provides a prepared and cleaned dataset with the SEC's Central Index Key number, filing date, form type, and file size for all 10-K filings at https://sraf.nd.edu/. At the time of our data retrieval, the dataset had been updated on 01/01/2019.

3.3. Calculating language predictability scores for MD&As

We use the following procedure to calculate the language predictability score (LPS) for each MD&A in our sample. First, for each word w_i in an MD&A, we obtain the preceding 256 words, $(w_{i-256}, \dots, w_{i-1})$ as its context.[11] Based on this context, the fine-tuned GPT-2 model estimates the probability of word w_i, $\mathrm{p}(w_i|w_{i-256}, \dots, w_{i-1})$. The language predictability of a word is calculated as the natural logarithm of the estimated conditional probability, $\log \mathrm{p}(w_i|w_{i-256}, \dots, w_{i-1})$.

Next, we calculate the LPS of the MD&A by aggregating the predictability of all of the words in an MD&A using their Z-score normalized mean as

$$Language\ Predictability\ Score\ (or\ LPS) = \frac{\left[\frac{\sum_{i}^{N} \log \mathrm{p}(W_i|W_{i-256}, \dots, W_{i-1})}{N}\right] - \mu}{\sigma},$$

where N is the total number of word predictions in the MD&A,[12] and μ and σ are the mean and standard deviation, respectively, of the language predictability of the words in the MD&As of the training dataset of the GPT-2 model.[13] We provide an example in Appendix B to illustrate the word-by-word language predictability estimations and the calculation of LPS.

[11] We use 256 words as the length of context because Keskar et al. (2019) show that using 256 words and using 512 words as context lengths to train LLMs produces similar performance in text generation tasks.

[12] N does not include the first 256 words in an MD&A because we require a word being predicted to have 256 preceding words as its context. N also excludes stop words, numbers, and punctuation. Stop words are commonly used words that carry little semantic meaning and are often necessary grammatical elements (e.g., 'a', 'the'). Numbers and punctuation marks usually serve specific functions and do not provide substantial contextual information. Note that, stop words, numbers, and punctuation are included in the context.

[13] Because we use year-specific GPT-2 models to estimate the language predictability of MD&As, the estimates may not be comparable across years because the statistical characteristics (e.g., means and standard deviations) may differ yearly.

4. Validation of the language predictability measure

4.1. Performance of the GPT-2 models in predicting the next word

A fundamental performance measure of a language model is its accuracy in predicting the next word given the word's context. This accuracy is calculated as the percentage of correct predictions. A prediction is considered correct when the word with the highest estimated probability according to the language model matches the actual next word written in the document.

The average accuracies of our GPT-2 models in the fine-tuning dataset (i.e., the in-sample dataset comprising five prior years of MD&As) and prediction dataset (i.e., the out-of-sample dataset comprising the current year's MD&As) are 50.2% and 49.8%, respectively. These accuracies are significantly higher than the baseline prediction accuracy of random guessing, which is about 1 out of 50,000, given that the dictionary of GPT-2 models contains about 50,000 unique words. These word prediction accuracies can also be translated into reading comprehension abilities. According to Bormuth (1968), accuracies of 44% and 57% in a cloze task are equivalent to 75% and 95% scores on the Gray Oral Reading Test (a popular English reading comprehension test for all ages in the US), which are the performance thresholds for readers who can effectively comprehend the content with instructions or independently, respectively. Hence, our GPT-2 models show strong comprehension ability of our MD&A data.

Moreover, our GPT-2 models' accuracies for the in-sample and out-of-sample datasets are similar, indicating that the risk of over-fitting the data is low.[14] The accuracy of our model is comparable to that obtained by Michaelov et al. (2022), who examine a set of GPT-based LLMs

[14] Overfitting occurs when a machine learning model fits its training data too closely, thereby learning too much noise and irrelevant information in the training dataset, and thus is unable to generalize well to new data or out-of-sample datasets.

shown to be highly effective in predicting words with reported accuracies ranging from 38.2% to 52.6% for general text.

4.2. Comparing the language predictability score of coherent and incoherent text: a falsification test

We conduct a falsification test to validate the effectiveness of our measures in reflecting human reading difficulty. Previous literature suggests that, when readers encounter semantically incoherent sentences, their ability to predict word occurrence decreases (Fitz and Chang, 2019). \Therefore, in this test, we falsify a text by shuffling the pairing of the focal words and their corresponding contexts. We then test whether the language predictability estimated by our GPT-2 models is higher for the original text than for the falsified one, which would indicate that our measure accurately reflects human reading difficulty due to a lack of logic flow.

Specifically, we randomly sample 10,000 focal sentences from our dataset along with their original contexts (i.e., the 256 words preceding the first word of the focal sentence) to form a coherent text dataset. We then use our GPT-2 models to calculate the LPS of the focal sentences in this coherent text dataset and find that the average LPS is -0.051, which is close to the median LPS for our entire sample (-0.063 reported in Table 2).

Next, we randomly shuffle the pairing between the focal sentences and their correct contexts, thereby creating a falsified dataset with incoherent matching between a sentence and its context. The average LPS of the focal sentences of this incoherent dataset is -0.717, which is below the 20^{th} percentile of our entire sample (-0.658, untabulated) and significantly lower than that of the coherent dataset ($p<0.01$). This result is consistent with our prediction that LPS captures human reading difficulty, which is greater for incoherent text than coherent text. By contrast, the Fog

index, the Bog index, and file size cannot distinguish the readability between coherent and noncoherent texts consisting of the same sentence.

4.3. Language predictability of boilerplate disclosures

Previous literature in psycholinguistics has demonstrated that readers tend to pay less attention to, or even skip, words in sentences when they have prior knowledge of the text content (Kaakinen and Hyönä, 2007; White et al., 2015). If we can find that the GPT-2 models generate higher LPS for such content, it provides further validation that our measure captures the processing cost of reading text by a target audience with prior knowledge.

In corporate disclosures, boilerplate content refers to generic and repetitive language that is commonly used across firms or remains unchanged for a firm over time (Brown and Tucker, 2011; Lang and Stice-Lawrence, 2015; Dyer et al., 2017). These disclosures are primarily driven by firms' litigation concerns or compliance needs and typically provide little new information (Huang et al., 2021). Hence, it is unclear that the target readers of financial reports incur significant processing costs when dealing with boilerplate disclosures owing to their redundant and repetitive nature (Blankespoor et al., 2020). As discussed in Section 2.2, according to surprisal theory, boilerplate disclosures should have a smaller $Surprisal$ effect, resulting in quicker processing for investors familiar with firm disclosures.

To conduct this test, we first separate the sentences in MD&As into boilerplate and non-boilerplate sentences. Following Lang and Stice-Lawrence (2015), we define a sentence as boilerplate when it contains four-word phrases (tetragrams) that occur in at least 60% of the MD&As in our sample. We find that the average ratio of boilerplate sentences in our sample MD&As is 12.3%. We then measure the language predictability of MD&As using boilerplate

and non-boilerplate sentences separately, denoted $LPS_{boilerplate}$ and $LPS_{non-boilerplate}$, respectively. We find that the average LPS estimated by the GPT-2 models for the boilerplate sentences is 1.320, which is higher than the 95th percentile of our entire sample (1.06673, untabulated), and also significantly higher than the average LPS for the non-boilerplate sentences (-0.334, the p-value of the differences between them is < 0.01). This result indicates that our measure correctly identifies generic and repetitive content as having a very low processing cost.

By contrast, we find that the Fog index of the boilerplate sentences is 29.26, significantly greater than 20.98 for the non-boilerplate sentences (p-value of differences between them is <0.01). This finding is not surprising given that boilerplate disclosures tend to consist of lengthy sentences and include many legal terms. In other words, using the Fog index as the readability measure, one would conclude that generic and repetitive content is more costly to process, which is inconsistent with the conjecture that investors tend to skip or skim through boilerplate disclosures. Similarly, readability measures based on the Bog index and file size suffer from the same concern because they increase along with the sentence length and the complex words involved in boilerplate disclosures.[15]

[15] The following is an example of a boilerplate sentence (from sec.gov/Archives/edgar/data/9346/0000009346-10-000009.txt): "*Investors are cautioned that such forward-looking statements involve risks and uncertainties including, without limitation, the following: (i) the Company's plans, strategies, objectives, expectations and intentions are subject to change at any time at the discretion of the Company; (ii) the Company's business is highly competitive and the entrance of new competitors into or the expansion of the operations by existing competitors in the Company's markets and other changes in the market for insurance products could adversely affect the Company's plans and results of operations; and (iii) other risks and uncertainties indicated from time to time in the Company's filings with the Securities and Exchange Commission.*" This sentence is a risk-factors boilerplate disclosure that barely changes year after year. It would be unintelligible to general audiences because of its excessive length and the use of jargon. Its Fog index is 50.18 and Bog index is 415, both of which are significantly higher than the average level observed in our sample (average Fog index: 21.160, average Bog index: 84.961). However, the target audience of annual reports (e.g., investors or analysts) likely find it easy to comprehend because they have read similar sentences many times. Consistent with the intuition, the LPS of this sentence is 1.404, which is higher than the 95th percentile of our entire sample (1.06673, untabulated), indicating that it is extremely predictable and easy to process.

This validity test highlights the limitations of the existing readability measures and the advantages of our approach. Whereas the existing readability measures ignore the target audiences' prior knowledge or familiarity with the text, our language predictability measure accounts for the reader–text interaction to provide a more accurate estimate of information processing costs. Given the increasingly lengthy boilerplate disclosures in corporate disclosures (Brown and Tucker, 2011; Lang and Stice-Lawrence, 2015; Dyer et al., 2017), the advantage conferred by our measure is particularly important for understanding the processing cost of corporate disclosures for investors and financial experts.

5. Empirical tests and results

5.1. Sample construction for the main tests

To prepare the sample for the main empirical tests, we follow LM (2014) to further filter the initial sample of 112,427 MD&A, as described in Section 3.3. We first match the sample with the CRSP database, which reduces it by 49,037 observations. Requiring a stock price on filing date *t*-1 of no less than $3 excludes another 17,144 observations.[16] Requiring a book-to-market value in COMPUSTAT removes another 2,814 observations. We are then left with a sample size of 43,432 observations. After excluding MD&As with fewer than 100 valid masked-word predictions and applying winsorization to LPS at the 0.5% and 99.5% levels to mitigate the impact of outliers, we further eliminate 462 observations. The final sample consists of 42,970 MD&A observations from 6,843 unique firms. Table 1 summarizes our sample selection procedure.

[INSERT TABLE 1 ABOUT HERE]

[16] We include this requirement because, as LM (2011, 2014) point out, removing low-priced firms minimizes the role of the bid-ask bounce when assessing market reactions to the 10-K.

5.2. Descriptive statistics of LPS and other readability measures

Panel A of Table 2 reports the descriptive statistics for the key variables used in our study. Appendix A provides detailed variable definitions. The mean of LPS is negative (-0.085) because we use the mean value of the in-sample LPS to normalize the out-of-sample LPS, the latter of which is, on average, lower than the former.

[INSERT TABLE 2 ABOUT HERE]

Panel B of Table 2 reports the correlations between LPS and other measures of readability. We observe a very low correlation between LPS and the $Bog\ Index$ and $Fog\ Index$ (0.004 and 0.004, respectively, both insignificant), thereby confirming the fundamental difference between LPS and these measures. We also find a negative correlation between LPS and $Log(Word\ Count)$ (-0.297, significant at the 0.01 level). Because LPS is based on the average language predictability of individual words contained in an MD&A, by construction, it is not affected by the length of the document. Hence, its negative relation with $Log(Word\ Count)$ suggests that, when an MD&A is longer, it tends to contain words that are, on average, less predictable, consistent with the finding that larger and more complex firms have longer MD&As (Li, 2008). When we remove boilerplate sentences from the MD&As, the correlation between $LPS_{non-boilerplat}$ and $Log(Word\ Count)$ remains negative (-0.288, significant at the 0.01 level).

5.3. Determinants of the language predictability of MD&As

As the first study to introduce language predictability as a much-needed readability measure for business documents, we investigate the economic determinants that drive the variation in the language predictability of MD&As. The purpose of this multivariate regression analysis is not to exhaust the economic determinants of LPS but to confirm that LPS can reflect the processing

costs of firms' information. Hence, we examine the following economic factors that typically increase the processing costs of firms' disclosures.

- *Negative earnings:* Negative earnings imply a higher processing cost because they are less persistent than positive earnings (Hayn, 1995; Lawrence et al., 2018). It is difficult for investors to understand the valuation implications of negative earnings because they cannot easily be incorporated into valuation models (Lundholm and Sloan, 2023). The economic factors that lead to negative earnings often imply greater uncertainty about the firms' future, such as financial distress and operational disruptions. As a result, analysts' forecast errors and forecast dispersions are larger for loss firms (Brown, 2001; Liu and Natarajan, 2012). We measure negative earnings ($NegEarnings$) as an indicator variable that equals one when a firm reports negative earnings in the current fiscal year and zero otherwise.

- *Operational complexity:* Some firms have complexities (e.g., more product lines or services, complicated supply chains or distribution networks, subsidiaries, complicated business models) that increase the processing cost of their disclosures. We measure the operational complexity ($NBSEG$) using the logarithm of the number of business segments at the end of the fiscal year.[17]

- *Significant firm events:* Firms undergoing significant and unusual events, such as changes in their operating segments resulting from merged or disposed businesses, face higher uncertainty and are likely to have more complex disclosures. We measure these events using

[17] We do not use the number of geographic segments as a measure of operational complexity because Li (2008) suggests that this variable is not associated with disclosure complexity. Additionally, many values for this variable are missing.

$Change_NBSEG$ (the natural logarithm of one plus the absolute change in a firm's number of business segments compared to the previous fiscal year).

- *Forward-looking disclosures*: Forward-looking statements in MD&As are less predictable than statements about firms' past performance because the future is uncertain, and investors are less likely to anticipate managers' future strategies. Muslu et al. (2015) find that firms provide more forward-looking statements when their information environments are poorer. Therefore, we include FLS, calculated as the ratio of the forward-looking sentences in the current MD&A, as a determinant of language predictability.

- *Special items*: Firms with a significant amount of special items are more likely to experience some unusual events (Li, 2008). The occurrence of special items represents extraordinary transactions or events that are not part of a company's regular operations (Johnson et al., 2011). Thus, the presence of special items may lead to the language used in annual reports becomes less predictable as companies strive to explain and disclose the impact of these irregular transactions or events accurately. Hence, we include SI as a determinant of language predictability, which is calculated as the natural logarithm of one plus the firm's total amount of special items in the current fiscal year.

We investigate the relation between LPS and the above factors using a multivariate regression model. We also include the following control variables used by LM (2014) and Bonsall et al. (2017) as additional determinants: (1) book-to-market ratio ($Log(BtoM)$), measured as the natural logarithm of the book-to-market ratio of year t, (2) market value ($Log(MVE)$), measured as the natural logarithm of market capitalization of year t, (3) pre-filing stock momentum ($PreALPHA$), measured as the intercept from a Fama-French three-factor

model estimated prior to the filing date, (4) pre-filing stock volatility ($PreRMSE$), measured as the pre-filing return root mean squared error measured from a Fama-French three-factor model using trading days [−252, −6], and (5) $NASDAQ$, an indicator variable that equals one if the firm traded on NASDAQ and zero otherwise. We also include year and Global Industry Classification Standard (GICS) industry fixed effects as determinants to control for the unobservable characteristics that are common for all firms in a given year, and those common for all firms within an industry, respectively.

Column (1) of Table 3 presents the results of regressing MD&A language predictability on the potential determinants. Consistent with our intuition, we find that the economic factors leading to higher processing costs are associated with less predictable language in MD&As. Specifically, we find significant and negative coefficients on $NegEarnings$, $NBSE,\ SI$ and $Change_NBSEG$ in the LPS regression (at least at the 0.05 level). These results suggest that firms with lower earnings, having greater operational complexity, with more special items, and experiencing greater changes in business segments tend to use less predictable language in their MD&As. Additionally, our finding of a significant and negative coefficient on FLS (at the 0.01 level) suggests that MD&As with more forward-looking statements exhibit less predictable language.

Columns (2) of Table 3 report the regression results with firm fixed effects. Interestingly, we find that all of the economic determinants are significantly related to LPS in the predicted directions. These results suggest that our constructed LPS is not limited to capturing firm-invariant characteristics about language complexity but can also effectively reflect time-variant patterns.

[INSERT TABLE 3 ABOUT HERE]

5.4. Consequences of the language predictability of MD&As

In this section, we investigate how the language predictability of MD&As affects the information environment after the 10-K filing date. We predict that, when the language used in MD&As is less predictable, it is more difficult for market participants to extract value-relevant information from such disclosures, leading to a higher stock return volatility in the period immediately following the filing of the disclosures. We also predict that disclosures with lower language predictability can be more challenging for analysts to process, resulting in greater analyst forecast dispersion and lower forecast accuracy. As argued by LM (2014), the post-filing stock return volatility is a market-based, all-inclusive measure that captures the processing costs of all investors, whereas analyst forecast dispersion and forecast accuracy capture the processing costs of sophisticated market participants, represented by financial analysts.

Following LM (2014) and Bonsall et al. (2017), we test the predictions by examining the association between MD&A language predictability (LPS) and the post-filing stock return volatility ($PostRMSE$), analyst forecast dispersion ($Dispersion$), and analyst earnings forecast accuracy (AFE). Specifically, $PostRMSE$ is the root mean squared error (RMSE) of the post-filing returns, estimated from a Fama-French three-factor model using trading days [6, 28], multiplied by 100. $Dispersion$ is the standard deviation of the post-filing analyst forecasts for the next fiscal quarter's earnings divided by the stock price immediately before the 10-K filing. AFE is the absolute value of the post-filing analyst forecast error of the next fiscal quarter's earnings scaled by the stock price at the fiscal quarter end. Notably, higher values of AFE indicate less accurate analyst forecasts. Appendix A contains detailed variable definitions.

To remain consistent with prior studies, we use the same set of control variables in LM (2014) and Bonsall et al. (2017): (1) the book-to-market ratio ($Log(BtoM)$), (2) market value

($Log(MVE)$), (3) pre-filing alpha ($Pre\ ALPHA$), (4) pre-filing RMSE ($PreRMSE$), (5) absolute filing period abnormal return ($Abs(CAR_{0,1})$ measured as the absolute value of the 2-day buy-and-hold abnormal return from the filing date (day 0 to day +1), and (6) NASDAQ.[18] When the dependent variables are analyst forecast dispersion and accuracy, we include an additional control variable for the number of analysts following the firm ($Num_Analysts$). We also include year fixed effects to control for temporal trends.

More importantly, we include firm fixed effects in all regressions to address the endogeneity concerns raised by Li (2010) and Bonsall et al. (2017) that potential time-invariant firm characteristics may influence the relation between readability measures and the market outcome variables capturing investor processing difficulty. Indeed, Bonsall et al. (2017) show that, when they replace industry fixed effects with firm fixed effects, the relation between the existing readability measures, including the Fog index and the Bog index, and processing costs, measured as analyst dispersion and analyst forecast accuracy, become insignificant, and the relation between readability, measured as file size and total number of words and analyst forecast accuracy, also becomes insignificant. Bonsall et al. (2017) interpret these results as readability not influencing sophisticated market participants. Therefore, we include firm fixed effects in all empirical tests so that we can document a reliable relation between readability and the processing costs of market participants driven by the variations across different MD&As within the same firm.

We also run the above regression separately for the other existing readability measures to compare the effect of LPS on capital market participants (Li, 2008; LM, 2014; Bonsall et al.,

[18] The associations between LPS and post-filing information environment remain significant when we include all determinants in Section 5.3.

2017), namely, (1) the Fog index of MD&As ($Fog\ Index$), (2) the Bog index of the report ($Bog\ Index$), (3) the logarithm of the 10-K file size ($Log(File\ Size)$), and (4) the logarithm of the number of words in the MD&As ($Log(Word\ Count)$).

As shown in Section 4.3, about 12.3% of MD&A content is boilerplate sentences. Investors, particularly sophisticated information intermediaries, likely recognize the repeating and irrelevant nature of this boilerplate language. Thus, we rerun all regressions using language predictability based on the non-boilerplate sentences (i.e., $LPS_{non-boilerplat}$). If, indeed, investors skim through or even skip boilerplate sentences, the relation between MD&A readability and market participants' processing costs can be captured more accurately by focusing on the non-boilerplate content in MD&As.

5.4.1. Impact of the predictability of MD&A language on subsequent stock return volatility

Table 4 reports the results of regressing $PostRMSE$ on LPS and other readability variables. Column (1) of Table 4 shows, after including firm fixed effects, the coefficient on LPS is significantly negative (at the 0.01 level), suggesting that less-predictable language in MD&As increases post-filing stock return volatility. In terms of economic significance, we find that, when LPS increases by one standard deviation (i.e., 0.712), $PostRMSE$ decreases by 2.95% of its standard deviation (i.e., -0.0659*0.712/1.593).

Column (2) of Table 4 shows the results with $LPS_{non-boilerpl}$ replacing LPS as the independent variable. We observe that $LPS_{non-boilerplate}$ is significantly and negatively associated with $PostRMSE$ (at the 0.01 level), suggesting that, if we assume that investors can identify and skim through or even skip boilerplate content, less predictable language in the non-boilerplate content of MD&As increases post-filing stock return volatility. Prior studies on

readability of corporate disclosures do not disentangle boilerplate from non-boilerplate content; thus, our results provide new insights to the literature that the negative relation between readability measured as language predictability and processing costs is not affected by boilerplate language.

Columns (3)–(6) of Table 4 show that, after including firm fixed effects, $Fog\ Index$ becomes insignificant in explaining future stock return volatility, but $Bog\ Index$, $Log(File\ Size)$, and $Log(Word\ Count)$ remain significant, suggesting that investors' processing cost is affected by language predictability, writing style, and information quantity.

Untabulated results show that, when other readability measures are included as control variables, the significant negative association between LPS and $PostRMSE$ remains significant at the 0.01 level, suggesting that LPS has an incremental impact on investors' processing costs of MD&As beyond those explained by the existing measures of readability.

[INSERT TABLE 4 ABOUT HERE]

5.4.2. Impact of language predictability on analysts' processing of information

It is unclear in the extant literature whether the readability of MD&As has any influence on sophisticated market participants. Bonsall et al. (2017) show that once firm fixed effects are controlled, most of the relation between analysts' processing costs and readability measured via the Fog index, Bog index, file size, and total number of words becomes insignificant. In this study, we argue that the insignificant results documented by Bonsall et al. (2017) could be due to the failure of the existing readability measures to capture the target readers' prior knowledge, which becomes a particularly salient issue when these readers, such as analysts, are sophisticated and experienced in processing these disclosures. Because LPS takes into account readers' prior

knowledge, we predict that, when the language in an MD&A is less predictable, analysts find it difficult to process; that is, analysts have more diverse interpretations of a firm's disclosure and face greater challenges to integrating the information into earnings forecasts. Following LM (2014) and Bonsall et al. (2017), we regress analyst forecast dispersion ($Dispersion$) and analyst earnings forecast accuracy (AFE) on the readability measures.

Tables 5 and 6 report the results of the regressions using $Dispersion$ and AFE as the dependent variables, respectively. In both tables, Column (1) shows that, after controlling for firm fixed effects, LPS is negatively and significantly associated with both $Dispersion$ and AFE (at least at the 0.05 level). This result is consistent with our prediction that less predictable language in MD&As increases analysts' processing costs. Specifically, a one-standard-deviation increase in LPS leads to decreases in $Dispersion$ and AFE that are 5.45% (-0.0452*0.712/0.591) and 2.69% (-0.0736*0.712/1.947) of their standard deviations, respectively. By contrast, Columns (3)–(5) show that there are no significant relations (at the 0.05 level) between $Fog\ Index$, $Bog\ Index$, or $Log(File\ Size)$ and analysts' processing costs (i.e., $Dispersion$ and AFE). These results are consistent with the findings of Bonsall et al. (2017). However, Column (6) shows that $Log(Word\ Count)$ is significantly related to analysts' processing costs. Untabulated results show that, when we include all readability measures in the regressions of analyst dispersion and forecast accuracy, LPS is consistently significant (at least at the 0.1 level). Therefore, we reach a different conclusion from Bonsall et al. (2017) that readability measured as language predictability has a significant impact on the processing costs of sophisticated market participants.

In Tables 5 and 6, Column (2) shows that, when we replace LPS with $LPS_{non-boilerpla}$, its coefficients are significant and negative (at least at the 0.05 level). This result suggests that the

language predictability of non-boilerplate content has a significant impact on analysts' costs of processing the disclosures. Interestingly, untabulated results show no significant relation between $LPS_{boilerplate}$ and either analyst forecast dispersion or accuracy. Together, these results speak to the question raised by Miller (2010) and Blankespoor et al. (2020) about whether sophisticated users can comprehend the importance (or lack of importance) of boilerplate information. Our finding that analysts are not affected by the readability of boilerplate sentences is consistent with sophisticated market participants, such as analysts, being able to identify and recognize the irrelevant nature of such content.

[INSERT TABLE 5 ABOUT HERE]

[INSERT TABLE 6 ABOUT HERE]

5.5. Additional analysis: the role of fine-tuning of GPT-2 in measuring language predictability

By design, fine-tuning can adapt LLMs to specific domain and thus further improve the performance of LLMs in various tasks involving domain knowledge. As we have shown in the validity tests and main empirical tests, our GPT-2 models after fine-tuning imitate investors who are familiar with the language patterns of MD&As and generate *LPS* estimates for MD&As that explain the processing costs of investors and financial analysts.

To understand the role of fine-tuning of GPT-2 in measuring language predictability of MD&As, we repeat our empirical tests using language predictability scores estimated by the pre-trained GPT-2 model from OpenAI. First, the overall accuracy of the pre-trained GPT-2 model in the masked word prediction task for our entire MD&A dataset, is 42.9%. As discussed in Section 4.1, an accuracy of 44% in a cloze task is equivalent to scoring 75% on the Gray Oral Reading Test (Bormuth 1968). Therefore, this performance of 42.9% indicates a reasonably strong ability

of pre-trained GPT-2 model in understanding the MD&As language, probably because GPT-2 is pre-trained with massive amount of training data (8 million Web pages) that already include various business or financial text, such as financial news, company Web pages, online discussion forums and encyclopedia websites related to business topics.

After fine-tuning the GPT-2 models using merely five-year's MD&A texts, its accuracy in the same masked word prediction task increases 16.08% to 49.8% (as reported in Section 4.1). This enhancement is substantial and indicates that the fine-tuning procedure is important for GPT-2 models to learn the language patterns of MD&As that are different from those in general business text. In other words, fine-tuning enables the GPT-2 models to imitate an expert who understands MD&As' language much more than an investor who is exposed to only general business texts.

The importance of fine-tuning is further demonstrated in Table 7, which reports the results of the main tests using the language predictability scores for MD&As estimated by the pre-trained GPT-2 models (denoted as LPS_{Pre_train}). Table 7 shows that, the coefficient on LPS_{Pre_train} is significant and negative (at the 0.01 level) in explaining post filing stock return volatility, but insignificant in the regressions of post-filing analyst forecast dispersion and accuracy. These results, together with our main results reported in Section 5.4, suggest that without fine-tuning, our readability measure explains the processing costs of overall investors in the market, but not those of sophisticated market participants, such as analysts. Stated differently, only when an MD&A contains words unexpected by "an expert" with domain knowledge of MD&As, would analysts find it difficult to process. Hence, this analysis highlights the importance of measuring readability in the context of target audience.

[INSERT TABLE 7 ABOUT HERE]

5.6. Robustness test: measuring language predictability using the BERT LLM

To examine whether our empirical findings are robust to different implementations of LLMs, we replicate our analysis using the BERT LLM, which is developed by Google and pre-trained on an extensive corpus of over 3.3 billion words from books and web pages (Devlin et al., 2019). Both BERT and GPT-2 LLMs are based on the transformer architecture, which has shown superior performance on various natural language tasks compared with other language model architectures. GPT-2 and BERT differ in their pre-training procedures: GPT-2 only uses next-word prediction to train the model; BERT is trained by performing both next-word and next-sentence predictions. Moreover, GPT-2 uses only the preceding text as the context in predicting masked words whereas BERT uses both the preceding and the following text as the context of the masked words. These differences indicate that GPT-2 is conceptually more aligned with the reading process we attempt to capture in our study, which is incremental and anticipatory, than BERT.

As a robustness check, we follow the same fine-tuning approach as for GPT-2 and calculate the LPS of each MD&A as described in Section 3. We then replicate our analysis as described in Section 5 using the LPS estimated with the BERT model and find that all results remain qualitatively similar. The implementation details and regression results are reported in Appendix E. In sum, we show that our readability measure is robust across two popular LLM implementations.

6. Conclusions

Most corporate disclosures are qualitative. However, the current understanding of the processing costs of qualitative information is limited (Blankespoor et al., 2020). Because human language involves high-dimensional and unstructured information, the attempt to quantify

readability—the relative ease with which a reader comprehends a text—is challenging both conceptually and empirically. Readability of textual disclosures is a necessary part of investors' processing costs and a determinant of effective communication and disclosure quality. Its importance is evidenced by the enormous number of studies of readability published in academic journals (see reviews by Li, 2010; LM, 2016, 2020). However, readability measures widely used in the extant literature, which are based on readability formulas (e.g., the Fog index) or the total quantity of the information (e.g., file size) are controversial (LM, 2020).

Answering a call by Blankespoor et al. (2020) and LM (2020), we propose a new measure of readability based on the concept of language predictability, which has strong theoretical roots and experimental support in psycholinguistics and cognitive science. We argue that it is conceptually appropriate for measuring the readability of corporate disclosures because it considers the context in which words are read and gauges the difficulty of comprehension from the perspective of the target audience. Specifically, reading difficulty for a text increases when the words encountered as the continuation of the preceding words (i.e., the context) are more unexpected for a target audience with certain types of language ability, domain knowledge, and backgrounds. Implementing the concept has become technologically feasible thanks to the recent advancement of LLMs such as GPT-2 and BERT that are sufficiently powerful to imitate the language ability of a general population or, with fine-tuning strategies, the ability of a special population.

Using a large sample of MD&As, we show how to implement language predictability using GPT-2. In the validation tests, we find that our readability measure, LPS, distinguishes coherent text from incoherent text that consists of the exact same words and classifies the boilerplate content (repeating and generic language) in MD&As as predictable, while the existing measures

(i.e., the Fog and Bog indexes and file size), by contrast, are not able to do so. We show further that the LPS of MD&As has a significant impact on the post-filing information environment. Specifically, after controlling for firm fixed effects, we find that lower language predictability in the MD&A section of an annual report is associated with larger post-filing stock return volatility as well as greater dispersion and lower accuracy in analysts' earnings forecasts. Contrary to the arguments put forth by Bonsall et al. (2017) that the clarity of disclosures does not affect the processing cost of sophisticated information intermediaries such as analysts, we find strong and consistent evidence that the readability of corporate disclosures measured by language predictability indeed impacts analysts' processing costs. We also find that the readability of non-boilerplate content has a significant influence on analysts while that of boilerplate content does not. This finding suggests that analysts can identify boilerplate language and may consider it irrelevant.

Overall, this new measure of readability contributes important new insights to the literature on processing costs. Given its conceptual appeals and technological feasibility, we believe this measure has great potential to broaden textual research in accounting and finance. Future research can assess its usefulness in measuring readability of other qualitative business documents, such as conference call transcripts, news articles, social media text, and regulatory announcements.

Reference

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Erlbaum;Hillsdale,NJ.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, 3(Feb), 1137-1155.

Biddle, G. C., Hilary, G., & Verdi, R. S. (2009). How does financial reporting quality relate to investment efficiency?. *Journal of Accounting and Economics*, 48(2-3), 112-131.

Blankespoor, E., deHaan, E., & Marinovic, I. (2020). Disclosure processing costs, investors' information choice, and equity market outcomes: A review. *Journal of Accounting and Economics*, 70(2-3), 101344.

Bonsall IV, S. B., Leone, A. J., Miller, B. P., & Rennekamp, K. (2017). A plain English measure of financial reporting readability. *Journal of Accounting and Economics*, 63(2-3), 329-357.

Bormuth, J. R. (1968). Cloze test readability: Criterion reference scores. *Journal of Educational Measurement*, 5(3), 189-196.

Boston, M. F., Hale, J., Kliegl, R., Patil, U., & Vasishth, S. (2008). Parsing costs as predictors of reading difficulty: An evaluation using the Potsdam Sentence Corpus. *Journal of Eye Movement Research*, 2(1).

Brown, L. D. (2001). A temporal analysis of earnings surprises: Profits versus losses. *Journal of Accounting Research*, *39*(2), 221-241.

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

Bruce, B., Rubin, A., & Starr, K. (1981). Why readability formulas fail. *IEEE Transactions on Professional Communication*, (1), 50-52.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712.

Bushee, B. J., Gow, I. D., & Taylor, D. J. (2018). Linguistic complexity in firm disclosures: Obfuscation or information?. *Journal of Accounting Research*, *56*(1), 85-121.

Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4), 359-394.

Clarkson, P. M., Kao, J. L., & Richardson, G. D. (1999). Evidence that management discussion and analysis (MD&A) is a part of a firm's overall disclosure package. *Contemporary Accounting Research*, 16(1), 111-134.

Davison, Alice, and Robert N. Kantor, 1982, On the failure of readability formulas to define readable texts: A case study from adaptations, *Reading Research Quarterly* 17, 187–209.

Dechow, P. M., & Dichev, I. D. (2002). The quality of accruals and earnings: The role of accrual estimation errors. *The Accounting Review*,77(s-1), 35-59.

Dechow, P., Ge, W., & Schrand, C. (2010). Understanding earnings quality: A review of the proxies, their determinants and their consequences. *Journal of Accounting and Economics*, 50(2-3), 344-401.

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature Neuroscience*, 8(8), 1117-1121.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (pp. 4171-4186).
Dougal, C., Engelberg, J., Garcia, D., & Parsons, C. A. (2012). Journalists and the stock market. *The Review of Financial Studies*, 25(3), 639-679.
Doya, K. (Ed.). (2007). Bayesian brain: Probabilistic approaches to neural coding. *MIT press*.
Dyer, T., Lang, M., & Stice-Lawrence, L. (2017). The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics*, 64(2-3), 221-245.
Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, 24(6), 409-436.
Efretuei, E., & Hussainey, K. (2023). The fog index in accounting research: contributions and challenges. *Journal of Applied Accounting Research*, 24(2), 318-343.
Eulerich, M., Sanatizadeh, A., Vakilzadeh, H., & Wood, D. A. (2023). Is it All Hype? ChatGPT's Performance and Disruptive Potential in the Accounting and Auditing Industries. *SSRN Electronic Journal*.
Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, 41(4), 469-495.
Fitz, H., & Chang, F. (2019). Language ERPs reflect learning through prediction error propagation. *Cognitive Psychology*, 111, 15-52.
Francis, J., Nanda, D., & Olsson, P. (2008). Voluntary disclosure, earnings quality, and cost of capital. *Journal of Accounting Research*, 46(1), 53-99.
Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140, 1-11.
Garimella, A., Mihalcea, R., & Amarnath, A. (2022, November). Demographic-aware language model fine-tuning as a bias mitigation technique. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing* (Volume 2: Short Papers) (pp. 311-319).
Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535-574.
Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics* (CMCL 2018) (pp. 10-18).
Goodman, K. S. (1984). Chapter IV: Unity in Reading. *Teachers College Record*, 85(5), 79-114.
Gunning, R. (1952). Technique of clear writing.
Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 2001 (Vol. 2, pp. 159-166).
Hayn, C. (1995). The information content of losses. *Journal of Accounting and Economics*, *20*(2), 125-153.
Hirshleifer, D., & Teoh, S. H. (2003). Limited attention, information disclosure, and financial reporting. *Journal of Accounting and Economics*, 36(1-3), 337-386.
Hirshleifer, David, Seongyeon Lim, Sonya, Hong Teoh, Siew, 2009. Driven to distraction: extraneous events and underreaction to earnings news. *Journal of Finance* 64, 2289–2325.

Huang, A. H., Shen, J., & Zang, A. Y. (2021). The unintended benefit of the risk factor mandate of 2005. *Review of Accounting Studies*, 1-37.
Jiang, Z., Xu, F. F., Araki, J., & Neubig, G. (2020). How can we know what language models know?. *Transactions of the Association for Computational Linguistics*, 8, 423-438.
Kaakinen, J. K., & Hyönä, J. (2007). Perspective effects in repeated reading: An eye movement study. *Memory and Cognition*, 35, 1323-1336.
Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., & Socher, R. (2019). Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv*:1909.05858.
Kim, C., Wang, K., & Zhang, L. (2019). Readability of 10-K reports and stock price crash risk. *Contemporary Accounting Research*, 36(2), 1184-1216.
Kuperberg, G. R., & Jaeger, T. F. (2016). What do we mean by prediction in language comprehension?. Language, *Cognition and Neuroscience*, 31(1), 32-59.
Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. *Predictions in the Brain: Using Our Past to Generate a Future*, (pp. 190–207). Oxford University.
Lang, M., & Stice-Lawrence, L. (2015). Textual analysis and international financial reporting: Large sample evidence. *Journal of Accounting and Economics*, 60(2-3), 110-135.
Lawrence, A., Sloan, R., & Sun, E. (2018). Why are losses less persistent than profits? Curtailments vs. conservatism. *Management Science*, *64*(2), 673-694.
Lehavy, R., Li, F., & Merkley, K. (2011). The effect of annual report readability on analyst following and the properties of their earnings forecasts. *The Accounting Review*, 86(3), 1087-1115.
Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, 106(3), 1126-1177.
Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, 45(2-3), 221-247.
Li, F. (2010). Textual analysis of corporate disclosures: A survey of the literature. *Journal of Accounting Literature*, 29(1), 143-165.
Liberti, J. M., & Petersen, M. A. (2019). Information: Hard and soft. *Review of Corporate Finance Studies*, *8*(1), 1-41.
Liu, X. G., & Natarajan, R. (2012). The effect of financial analysts' strategic behavior on analysts' forecast dispersion. *The Accounting Review*, *87*(6), 2123-2149.
Lim, E. K., Chalmers, K., & Hanlon, D. (2018). The influence of business strategy on annual report readability. *Journal of Accounting and Public Policy*, 37(1), 65-81.
Loughran, T., & McDonald, B. (2014). Measuring readability in financial disclosures. *Journal of Finance*, 69(4), 1643-1671.
Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, 54(4), 1187-1230.
Loughran, T., & McDonald, B. (2020). Textual analysis in finance. *Annual Review of Financial Economics*, 12, 357-375.
Lundholm, R. J., & Sloan, R. G. (2023). *Equity Valuation and Analysis.* Independently published.
Michaelov, J. A., Bardolph, M. D., Coulson, S., & Bergen, B. (2021). Different kinds of cognitive plausibility: Why are transformers better than RNNs at predicting N400 amplitude?. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 43, No. 43).

Michaelov, J. A., Coulson, S., & Bergen, B. K. (2022). So cloze yet so far: N400 amplitude is better predicted by distributional information than human predictability judgements. *IEEE Transactions on Cognitive and Developmental Systems*.

Miller, B. P. (2010). The effects of reporting complexity on small and large investor trading. *The Accounting Review*, 85(6), 2107-2143.

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

OpenAI (2023). Gpt-4 technical report. *arXiv preprint arXiv*:2303.08774.

Peng, L. (2005). Learning with information capacity constraints. *Journal of Financial and Quantitative Analysis*, 40(2), 307-329.

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H., & Riedel, S. (2019). Language models as knowledge bases?. *arXiv preprint arXiv:1909.01066*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.

Rayner, K., & Well, A. D. (1996). Effects of contextual constraint on eye movements in reading: A further examination. *Psychonomic Bulletin & Review*, 3(4), 504-509.

Rayner, K., Pollatsek, A., Ashby, J., & Clifton Jr, C. (2012). Psychology of reading. *Psychology Press*.

Rozado, D. (2024). The political preferences of LLMs. *PloS one*, 19(7), e0306621.

Securities and Exchange Commission, 1998. A Plain English Handbook: How to Create Clear SEC Disclosure. (http://www.sec.gov/pdf/handbook.pdf).

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., ... & Natarajan, V. (2023). Large language models encode clinical knowledge. *Nature*, *620*(7972), 172-180.

Sims, C. A. (2010). Rational inattention and monetary economics. In *Handbook of Monetary Economics* (Vol. 3, pp. 155-181). Elsevier.

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

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3), 302-319.

Stevens, K. T., Stevens, K. C., & Stevens, W. P. (1992). Measuring the readability of business writing: The cloze procedure versus readability formulas. *The Journal of Business Communication (1973)*, *29*(4), 367-382.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217), 1632-1634.

Taylor, W. L. (1953). "Cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4), 415-433.

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N; Kaiser, Łukasz; Polosukhin, Illia (2017). "Attention is All you Need". *Advances in Neural Information Processing Systems*. Curran Associates, Inc. 30

Wilcox, E. G., Gauthier, J., Hu, J., Qian, P., & Levy, R. P. (2020). On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society* (CogSci 2020) (p. 7).

White, S. J., Warrington, K. L., McGowan, V. A., & Paterson, K. B. (2015). Eye movements during reading and topic scanning: Effects of word frequency. *Journal of Experimental Psychology: Human Perception and Performance*, 41(1), 233.

You, H., & Zhang, X. J. (2009). Financial reporting complexity and investor underreaction to 10-K information. *Review of Accounting studies*, 14, 559-586.
Zakaluk, B. L., & Samuels, S. J. (1988). Readability: Its Past, Present, and Future. *International Reading Association*

Table 1. Sample Selection

Purpose	Source/Filter	Sample Size	Observations Removed
Generating language predictability scores	10-Ks and 10-K405s from EDGAR from 1995 to 2018	199,965	-
	Include only the first filing per firm in each calendar year	189,404	10,561
	With at least 180 days from prior filings	188,923	481
	Successfully detect the MD&A section	140,940	47,983
	With at least 250 words in the MD&A	136,028	4,912
	Exclude observations filed before 2000	112,427	23,601
For archival-based capital market tests	Require a CRSP PERMNO match and as ordinary common equity	63,390	49,037
	CSRP market capitalization data to be available and price on filing date day minus one $\geq$ \$3	46,246	17,144
	Book-to-market and market value to be available in the COMPUSTAT dataset and to have a positive book value	43,432	2,814
	Require at least 100 valid masked-word predictions and remove outlier of the estimated language predictability	42,970	462

Table 2. Summary Statistics and Pairwise Correlation

Panel A: This table reports descriptive information for the key variables in our study. Detailed variable definitions are provided in Appendix A.

Variable Name	Obs	Mean	Median	SD	P25	P75
LPS	42,970	-0.085	-0.063	0.712	-0.532	0.388
$Log(Word\ Count)$	42,970	9.010	9.081	0.636	8.716	9.397
$Fog\ Index$	42,970	21.160	21.029	1.899	20.039	22.097
$Log(File\ Size)$	42,970	14.903	14.670	1.521	13.812	16.320
$Bog\ Index$	42,739	84.961	85.000	6.994	80.000	89.000
$LPS_{boilerplate}$	42,661	1.320	1.289	1.015	0.650	1.958
$LPS_{non-boilerplate}$	42,970	-0.334	-0.309	0.768	-0.808	0.172
$PreRMSE$	42,970	2.685	2.278	1.656	1.575	3.335
FLS	42,970	0.177	0.165	0.074	0.129	0.209
$Log(BtoM)$	42,970	-0.814	-0.716	0.851	-1.266	-0.255
$Log(MVE)$	42,970	6.452	6.375	1.792	5.192	7.620
$PreALPHA$	42,970	0.048	0.029	0.185	-0.049	0.118
$NASDAQ$	42,970	0.585	1.000	0.493	0.000	1.000
$NegEarnings$	42,806	0.182	0.000	0.386	0.000	0.000
$Abs(CAR_{0,1})$	42,552	0.032	0.018	0.045	0.007	0.038
SI	41,624	1.473	0.787	1.746	0.000	2.555
$PostRMSE$	40,185	2.528	2.128	1.593	1.473	3.095
$Change_NBSEG$	38,907	0.706	0.000	1.319	0.000	1.000
$NBSEG$	36,114	1.812	1.386	0.602	1.386	2.303
$Num_Analysts$	22,173	4.471	3.000	4.638	1.000	6.000
AFE	21,901	0.520	0.173	1.947	0.062	0.449
$Dispersion$	15,803	0.243	0.096	0.591	0.039	0.232

Panel B: Pairwise correlations among readability measures

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1) LPS	1.000						
(2) $LPS_{boilerplate}$	0.657	1.000					
(3) $LPS_{non-boilerplate}$	0.947	0.511	1.000				
(4) $Log(Word\ Count)$	-0.297	-0.231	-0.288	1.000			
(5) $Fog\ Index$	0.004	-0.082	-0.006	0.051	1.000		
(6) $Bog\ Index$	0.004	-0.001	0.006	0.256	0.284	1.000	
(7) $Log(File\ Size)$	-0.169	-0.062	-0.137	0.415	0.168	0.316	1.000

Table 3. Determinants of MD&A Language Predictability

The dependent variable is language predictability scores (LPS). The t-statistics are reported in parentheses below the coefficient. ***/**/* means significance at 0.01/0.05/0.10 level.

VARIABLES	(1) LPS	(2) LPS
$NegEarnings$	-0.0246**	-0.0277***
	(-2.550)	(-3.492)
SI	-0.0181***	-0.00908***
	(-7.992)	(-5.336)
$NBSEG$	-0.144***	-0.0576***
	(-24.01)	(-8.074)
$Change_NBSEG$	-0.0171***	-0.0112***
	(-6.802)	(-6.236)
FLS	-2.122***	-2.333***
	(-43.05)	(-50.94)
$Log(BtoM)$	-0.0336***	-0.0247***
	(-7.180)	(-5.152)
$Log(MVE)$	-0.0453***	-0.0106**
	(-16.35)	(-2.440)
$PreALPHA$	0.0240	0.0230*
	(1.287)	(1.828)
$PreRMSE$	-0.00430	-0.00326
	(-1.494)	(-1.428)
$NASDAQ$	0.130***	-0.0355**
	(16.53)	(-1.985)
Industry Fixed Effect	Yes	No
Firm Fixed Effects	No	Yes
Year Fixed Effect	Yes	Yes
Observations	37,608	36,341
Adj. R-Squared	0.261	0.746

Table 4. Impact of MD&A Language Predictability on Subsequent Stock Return Volatility
The dependent variable is the post-event return volatility, $PostRMSE$ (defined as the root-mean square error from a Fama-French model estimated using trading days [6, 28]). The t-statistics are reported in parentheses below the coefficient, where standard errors are clustered by firm. ***/**/* means significance at 0.01/0.05/0.10 level.

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	$PostRMSE$	$PostRMSE$	$PostRMSE$	$PostRMSE$	$PostRMSE$	$PostRMSE$
LPS	-0.0659***					
	(-4.253)					
$LPS_{non-boilerplate}$		-0.0544***				
		(-4.059)				
$Fog\ Index$			-0.00611			
			(-1.328)			
$Bog\ Index$				0.0152***		
				(3.371)		
$Log(File\ Size)$					0.0521***	
					(4.791)	
$Log(Word\ Count)$						0.110***
						(6.473)
$Log(BtoM)$	-0.115***	-0.115***	-0.113***	-0.121***	-0.116***	-0.118***
	(-7.238)	(-7.232)	(-7.122)	(-7.432)	(-7.264)	(-7.415)
$Log(MVE)$	-0.184***	-0.184***	-0.184***	-0.187***	-0.188***	-0.188***
	(-10.54)	(-10.55)	(-10.49)	(-10.67)	(-10.66)	(-10.73)
$PreALPHA$	-0.629***	-0.630***	-0.632***	-0.633***	-0.629***	-0.625***
	(-13.72)	(-13.73)	(-13.78)	(-13.78)	(-13.73)	(-13.64)
$PreRMSE$	0.256***	0.256***	0.257***	0.254***	0.256***	0.253***
	(13.19)	(13.19)	(13.23)	(13.07)	(13.18)	(13.03)
$Abs(CAR_{0,1})$	2.408***	2.411***	2.413***	2.399***	2.413***	2.392***
	(6.018)	(6.025)	(6.022)	(5.957)	(6.027)	(5.987)
$NASDAQ$	0.0212	0.0200	0.0198	0.0256	0.0203	0.0276
	(0.449)	(0.422)	(0.420)	(0.542)	(0.431)	(0.585)
Firm Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Year Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Observations	38,672	38,672	38,672	38,463	38,672	38,672
Adj. R-Squared	0.651	0.651	0.651	0.651	0.651	0.651

Table 5. Impact of MD&A Language Predictability on Analyst Forecast Dispersion
The dependent variable is the analyst forecast dispersion, $Dispersion$ (defined as the standard deviation of analysts' forecasts for the next fiscal quarter's earnings). The t-statistics are reported in parentheses below the coefficient, where standard errors are clustered by firm. ***/**/* means significance at 0.01/0.05/0.10 level.

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	$Dispersion$	$Dispersion$	$Dispersion$	$Dispersion$	$Dispersion$	$Dispersion$
LPS	-0.0452***					
	(-2.692)					
$LPS_{non-boilerplate}$		-0.0341**				
		(-1.973)				
$Fog\ Index$			-0.000347			
			(-0.118)			
$Bog\ Index$				-0.000716		
				(-0.320)		
$Log(File\ Size)$					0.00611	
					(0.853)	
$Log(Word\ Count)$						0.0367***
						(2.656)
$Log(BtoM)$	0.00367	0.00436	0.00559	0.00590	0.00522	0.00383
	(0.266)	(0.318)	(0.409)	(0.427)	(0.381)	(0.277)
$Log(MVE)$	-0.134***	-0.133***	-0.133***	-0.134***	-0.133***	-0.133***
	(-9.341)	(-9.321)	(-9.330)	(-9.358)	(-9.339)	(-9.372)
$Abs(CAR_{0,1})$	0.308*	0.310*	0.309*	0.310*	0.309*	0.308*
	(1.943)	(1.956)	(1.944)	(1.938)	(1.940)	(1.939)
$PreALPHA$	-0.501***	-0.501***	-0.503***	-0.505***	-0.503***	-0.502***
	(-9.508)	(-9.523)	(-9.491)	(-9.440)	(-9.491)	(-9.493)
$PreRMSE$	0.144***	0.145***	0.145***	0.145***	0.145***	0.144***
	(6.092)	(6.094)	(6.101)	(6.054)	(6.098)	(6.099)
$NASDAQ$	-0.104**	-0.104**	-0.104**	-0.103**	-0.104**	-0.102**
	(-2.329)	(-2.316)	(-2.301)	(-2.276)	(-2.307)	(-2.279)
$Num_Analysts$	0.0056***	0.0056***	0.0057***	0.0057***	0.0057***	0.0057***
	(4.514)	(4.497)	(4.546)	(4.514)	(4.545)	(4.546)
Firm Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Year Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Observations	14,754	14,754	14,754	14,687	14,754	14,754
Adj. R-Squared	0.321	0.321	0.321	0.321	0.321	0.321

Table 6. Impact of MD&A Language Predictability on Analyst Forecast Accuracy
The dependent variable is the analyst forecast accuracy, AFE (defined as the absolute value of the analysts' forecast error of next fiscal quarter's earnings). The t-statistics are reported in parentheses below the coefficient, where standard errors are clustered by firm. ***/**/* means significance at 0.01/0.05/0.10 level.

VARIABLES	(1) AFE	(2) AFE	(3) AFE	(4) AFE	(5) AFE	(6) AFE
LPS	-0.0736**					
	(-2.127)					
$LPS_{non-boilerplate}$		-0.0584**				
		(-2.086)				
$Fog\ Index$			0.0139*			
			(1.768)			
$Bog\ Index$				0.000960		
				(0.247)		
$Log(File\ Size)$					0.0185	
					(0.747)	
$Log(Word\ Count)$						0.0756**
						(2.083)
$Log(BtoM)$	0.0299	0.0303	0.0311	0.0318	0.0314	0.0293
	(0.890)	(0.902)	(0.925)	(0.943)	(0.931)	(0.869)
$Log(MVE)$	-0.402***	-0.402***	-0.401***	-0.403***	-0.403***	-0.403***
	(-9.260)	(-9.259)	(-9.212)	(-9.212)	(-9.224)	(-9.261)
$Abs(CAR_{0,1})$	-0.764***	-0.765***	-0.766***	-0.770***	-0.765***	-0.763***
	(-5.727)	(-5.730)	(-5.733)	(-5.741)	(-5.732)	(-5.705)
$PreALPHA$	0.261***	0.261***	0.262***	0.263***	0.261***	0.259***
	(5.722)	(5.721)	(5.727)	(5.723)	(5.731)	(5.689)
$PreRMSE$	0.695	0.698	0.697	0.708	0.694	0.687
	(1.530)	(1.538)	(1.534)	(1.553)	(1.528)	(1.515)
$NASDAQ$	-0.0950	-0.0946	-0.0953	-0.0948	-0.0973	-0.0912
	(-0.997)	(-0.991)	(-0.996)	(-0.989)	(-1.015)	(-0.952)
$Num_Analysts$	-0.0076**	-0.0076**	-0.0075**	-0.0074**	-0.0075**	-0.0075**
	(-2.025)	(-2.040)	(-1.996)	(-1.976)	(-1.994)	(-2.001)
Firm Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Year Fixed Effect	Yes	Yes	Yes	Yes	Yes	Yes
Observations	20,700	20,700	20,700	20,600	20,700	20,700
Adj. R-Squared	0.231	0.231	0.231	0.230	0.231	0.231

Table 7. Impact of MD&A Language Predictability (estimated using GPT-2 without fine-tuning) on Information Environment after the 10-K Filing Date

The dependent variables are the post-event return volatility ($PostRMSE$), analyst forecast dispersion ($Dispersion$) and analyst forecast accuracy (AFE), respectively. LPS_{Pre_train} is measured using GPT-2 without fine-tuning and has been normalized to have a standard deviation of 1. The t-statistics are reported in parentheses below the coefficient, where standard errors are clustered by firm. ***/**/* means significance at 0.01/0.05/0.10 level.

	(1)	(2)	(3)
VARIABLES	$RMSE$	$Dispersion$	AFE
LPS_{Pre_train}	-0.0329***	-0.0138	-0.0131
	(-3.239)	(-1.352)	(-0.423)
Control Variables	Included	Included	Included
Firm Fixed Effect	Yes	Yes	Yes
Year Fixed Effect	Yes	Yes	Yes
Observations	38,672	14,754	20,700
Adj. R-Squared	0.651	0.231	0.231

Appendix A. Variable Definitions

Variable Names	Definitions
LPS	The average language predictability of the words contained in an MD&A, normalized using the mean and standard deviation of language predictability of words in the training dataset. Language predictability of a word is estimated by the fine-tuned GPT-2 models and calculated as the natural logarithm of the conditional probabilities of this word given its preceding 256 words (i.e., its context). This variable is winsorized at the top and bottom 1% levels.
$LPS_{boilerplate}$	LPS of an MD&A, in which only words in boilerplate sentences are used for calculation. This variable is winsorized at the top and bottom 1% levels. We follow Lang and Stice-Lawrence (2015) to identify boilerplate sentences.
$LPS_{non-boilerplate}$	LPS of an MD&A, in which only words in non-boilerplate sentences are used for calculation. This variable is winsorized at the top and bottom 1% levels.
LPS_{pre_train}	Same as LPS but estimated by the pre-trained GPT-2 model before fine-tuning, winsorized at the top and bottom 1% levels.
$Log(Word\ Count)$	The natural logarithm of the number of words in the MD&A section.
$Fog\ Index$	The Gunning (1952) Fog index calculated using Python's textstat package.
$Bog\ Index$	A plain English measure. The data are from Bonsall et al. (2017).
$Log(File\ Size)$	The logarithm of the 10-K file size.
$PreRMSE$	Prior period return RMSE measured with a Fama-French three-factor model using trading days [−252, −6], multiplied by 100. At least 60 observations of daily returns must be available to be included in the sample.
FLS	The ratio of forward-looking statement (FLS) sentences in MD&A. We follow Li (2010) and classify sentences as forward-looking or non-forward-looking statement.
$Log(BookToMarket)$	The natural logarithm of book value divided by market value.
$Log(Market\ Value)$	The natural logarithm of the market value of equity.
$PreALPHA$	Prior period return alpha measured with a Fama-French three-factor model using trading days [−252, −6], multiplied by 100. At least 60 observations of daily returns must be available to be included in the sample.
$NASDAQ$	An indicator variable that equals one when a firm is listed on NASDAQ at the filing time of the 10-K, and zero otherwise.
$NegEarnings$	An indicator variable that equals one when the firm reports negative earnings in the current fiscal year, and zero otherwise.
$Abs(CAR_{0,1})$	The absolute value of the filing date excess return, where the filing date excess return is measured as the buy-and-hold return starting on the filing date (day 0) through day +1 minus the buy-and-hold return of the CRSP value-weighted index over the same two-day period.
AFE	The absolute value of the analysts' forecast error of the next fiscal quarter's earnings, where the forecast error is measured as (actual earnings – average analysts' forecast earnings) normalized by the stock price from before the next fiscal quarter's announcement date. When a given analyst has more than one recorded forecast, only the forecast closest to the earnings announcement

	is included in the sample. Only forecasts occurring after the 10-K filing date are included.
$Dispersion$	The standard deviation of analysts' forecasts for the next fiscal quarter's earnings divided by the stock price from before the next fiscal quarter's announcement date. Only forecasts occurring after the 10-K filing date are included. Firms with fewer than two analyst forecasts are assigned a missing value.
$Num_Analysts$	The number of analysts issuing forecasts for the earnings of the firm in the next fiscal quarter. Only forecasts occurring after the 10-K filing date are included.
$PostRMSE$	Post-filing return RMSE measured with a Fama-French three-factor model using trading days [6, 28], multiplied by 100. At least 10 observations of daily returns must be available to be included in the sample.
SI	The natural logarithm of the number of special items at the end of the fiscal year.
$NBSEG$	The logarithm of the number of business segments at the end of the fiscal year.
$Change_NBSEG$	The natural logarithm of one plus the absolute change in a firm's number of business segments compared with the previous fiscal year.

Appendix B. Examples of Using GPT-2 to Calculate *LPS*

Preceding words (*italics*) and next word (<bold>)	**Language predictability of next word**	**Natural logarithm of Language predictability of next word**
…human resource and administrative functions of the acquired companies. Such efforts, to date, have not been successful and our failure to **<successfully>**…	0.032959	-3.412
…resource and administrative functions of the acquired companies. Such efforts, to date, have not been successful and our failure to successfully **<integrate>**…	0.46782	-0.759
…and administrative functions of the acquired companies. Such efforts, to date, have not been successful and our failure to successfully integrate **<the>**…	Stop word (omitted)	-
…administrative functions of the acquired companies. Such efforts, to date, have not been successful and our failure to successfully integrate the **<acquired>**…	0.788002	-0.238
…functions of the acquired companies. Such efforts, to date, have not been successful and our failure to successfully integrate the acquired **<businesses>**…	0.592787	-0.522
…of the acquired companies. Such efforts, to date, have not been successful and our failure to successfully integrate the acquired businesses **<creates>**…	0.000924	-6.986
…the acquired companies. Such efforts, to date, have not been successful and our failure to successfully integrate the acquired businesses creates **<a>**…	Stop word (omitted)	-
…acquired companies. Such efforts, to date, have not been successful and our failure to successfully integrate the acquired businesses creates a **<substantial>**…	0.044709	-3.107
…companies. Such efforts, to date, have not been successful and our failure to successfully integrate the acquired businesses creates a substantial **<risk>…**	0.234232	-1.451

Note: The *text is from https://www.sec.gov/Archives/edgar/data/1460329/000095012310098381/g25003e10vk.htm*

This table provides an illustrative example of how the *LPS* is calculated for a nine-word segment of text, i.e., "…*successfully integrate the acquired business creates a substantial risk…*". The first column shows the preceding words (i.e., context) based on which GPT-2 calculates the conditional probability of next word, denoted as "**<bold>**." The second column shows the estimated probability of the next word given the preceding words using GPT-2. In this example, the *LPS* is calculated as the natural logarithm of seven estimated probabilities ("**the**" and "**a**" are omitted because they are stop words) normalized by *LPS*'s mean and STD in the training sample, -2.373 and 0.338, respectively:

$$LPS = \frac{\frac{(-3.412)+(-0.759)+(-0.238)+(-0.522)+(-6.986)+(-3.107)+(-1.451)}{7} - (-2.373)}{0.338} = 0.057$$

Appendix C. Language Predictability Examples in MD&A

We extract four examples from our MD&A sample and use our fine-tuned GPT-2 models to predict the masked word given the preceding words (i.e., context), shown in column 1. The top five highly predictable words are shown in column 2 in the decreasing order of their conditional probabilities estimated by our GPT-2 models, which are shown in column 3.

Examples of Predictable Words

Preceding words	Top candidate words	Estimated probability
In 2009, we did not pay our lease payment obligations for many of the elevator leases in our Jingli Shanghai elevator operations, which resulted in termination of many of these elevator leases. If a significant number of our elevator leases are **<MASKED WORD>**	terminated	0.251895
	not	0.072988
	closed	0.042995
	cancelled	0.034856
	renewed	0.028269

In this example, the actual word in the masked-word position is "**terminated**." It is the highest-probability (0.251895) word according to GPT-2's output. The context preceding the masked word mentions the termination of many elevator leases. This leads to a logical expectation that the masked word would relate to such activities, such as "**terminated**." Other top words are also reasonable predictions: "**not**" suggests a negative outcome, while "**closed**" and "**cancelled**" imply similar conclusions about the status of the leases. This example shows the effectiveness of GPT-2 in understanding the language patterns of MD&A text. It is also consistent with high-predictability words incurring low processing costs because they are anticipated by the audience even before seeing them and easy to integrate into the context.

Preceding words	Top candidate words	Estimated probability
Such efforts, to date, have not been successful and our failure to successfully integrate the acquired businesses creates a substantial risk that we may not be able to fully realize the anticipated **<MASKED WORD>**	benefits	0.439675
	results	0.073413
	future	0.034301
	benefit	0.030459
	cash	0.025745

In this example, the actual word in the masked-word position is "**benefits**." It is also the word with the highest conditional probability (0.439675) according to GPT-2's output. Examining the context, we observe that this company invested some effort, acquiring business, in this case, and anticipated economical returns; "**benefits**" is a commonly used word after "anticipated" in such contexts in annual reports. This example is also consistent with high-predictability words incurring low processing costs.

Examples of Unpredictable Words

Preceding words	Top candidate words	Estimated probability
During the second quarter of 1999, the Company concluded its negotiations on certain restaurants included in the fourth quarter special charge on a favorable basis compared to amounts previously accrued. As a result, the Company recognized a $730,000 special credit during the three months ended June 30, 1999. In January 2000, an additional property was **<MASKED WORD>**	acquired	0.151417
	placed	0.128296
	sold	0.080897
	opened	0.061917
	leased	0.05746

In this example, the actual word in the masked-word position is "**exited**," with an estimated probability of 0.0008. By examining the context, while the word "negotiation" gives some hints about acquisition activity, it is not clear whether it is "**acquired**" or "**sold**," both of which are top words with high probabilities. The actual word "**exited**" is not predictable because it is not a commonly used word describing "property." Therefore, it is a "surprise" to the audience and likely incurs high processing costs because an accruing contextual representation must be reanalyzed or overridden.

Preceding words	Top candidate words	Estimated probability
Revenue from the sale of livestock health management and quality enhancement products, equine imaging cameras and NutriCharge, is recognized on shipment to the customer. Revenue from our information management products is recognized in the period in which the information or analysis is delivered to the customer, normally on a monthly basis. These **<MASKED WORD>**	products	0.181573
	revenue	0.101154
	revenues	0.097357
	arrangements	0.061669
	services	0.039228

In this example, the actual word in the masked-word position is "**assets**," with an estimated probability of 0.0009. The context discusses revenue generated by selling certain products. Intuitively, the word in the masked position should be one of the top words in the table. However, the actual word is "**assets**." The discrepancy between the model predictions, which represent the audience's anticipation, and the actual word triggers some form of extra processing, such as to check whether the conflict is due to processing errors, to restructure the contextual representations, and to learn new language patterns.

Appendix D. Implementation of GPT-2 Model Training

During the fine-tuning, the pre-trained GPT-2 model is used as a classification model and is trained to predict the focal word based on its preceding 256 tokens. The deviation between the model's prediction and the exact word occurrence is measured using the cross-entropy loss: $Loss_i = -\log P(w_i|w_{i-256}, \dots, w_{i-1})$, where w_i refers to the masked word in position i, $(w_{i-256}, \dots, w_{i-1})$ is the words preceding the masked word (i.e., the context); $P(w_i|w_{i-256}, \dots, w_{i-1})$ is the model's predicted probability of the word, w_i, conditional on the preceding words $(w_{i-256}, \dots, w_{i-1})$.

As the training progresses, the parameters of the GPT-2 model are gradually updated to minimize the predictions' cross-entropy loss, improving the model's overall ability to predict the next word in the corpus.

We use the Transformers library in Python to develop our program. Specifically, we use the *class transformers.GPT2ForSequenceClassification* and the following parameters to train our language models:

- Number of preceding tokens used for prediction = 256: the number of preceding tokens used as context for predicting the next word's occurrence.
- Number of epochs = 20: the number of times that the model goes through the corpus to train the parameters.
- Number of <context, word> pairs from each document = 20: the number of times that a <context, word> pair is randomly sampled from each document in the corpus.
- Number of GPUs = 3: the number of GPUs used for training the model. We use three A800 GPUs for training.
- Batch size = 120: the number of samples processed in each GPU before the model is updated.
- Learning rate=0.0001: the extent to which the model updates its parameters after each prediction.

We use default values for other parameters in *transformers.TrainingArguments*. We adopt the "gpt2" pre-trained model, which was initially trained by OpenAI, and used Pytorch to implement the fine-tuning. The GPT-2 model that we use is provided by HuggingFace. Before training starts, we apply the tokenizer of GPT-2 to transform the text into tokens.

For each MD&A in the training sample, we randomly sample 400 <context, word> pairs. The context refers to the 256 tokens preceding the predicting word. Predicting words that have less than 256 preceding tokens or that require representation by splitting tokens are excluded from the sampling process. Additionally, we exclude stop words, numbers, and punctuation as predicting words in the sampling procedure.

Appendix E. BERT Large Language Model

E1. Implementation of BERT Model Training

We use the Transformers library in Python to develop our program. Specifically, we leverage the *class transformers.BertForSequenceClassification* and the following parameters to train our language models:

- Number of preceding tokens used for prediction = 256: the number of preceding tokens used as context for predicting the next word's occurrence.
- Number of epochs = 20: the number of times that the model goes through the corpus to train the parameters.
- Number of <context, word> pairs from each document = 20: the number of times that a <context, word> pair is randomly sampled from each document in the corpus.
- Number of GPUs = 5: the number of GPUs used for training the model. We use five 2080Ti GPUs for training.
- Batch size = 24: the number of samples processed in each GPU before the model is updated.
- Learning rate=0.0001: the extent to which the model updates its parameters after each prediction.

We use default values for other parameters in *transformers.TrainingArguments*.

We adopt the 'bert-base-uncased' pre-trained model, which was initially trained by Google, and use Pytorch to implement the fine-tuning. The BERT model that we use is provided by HuggingFace. Natural language processing (NLP) researchers from HuggingFace have made the PyTorch version of BERT available, which is compatible with Google's pre-trained checkpoints and is capable of reproducing their results. [19]

[19] https://github.com/google-research/bert/blob/master/README.md

E2. Impact of MD&A Language Predictability Estimated by BERT on Subsequent Stock Return Volatility

The dependent variable is the post-event return volatility: $PostRMSE$ (defined as the root-mean-square error from a Fama-French model estimated using trading days [6, 28]). The t-statistics are reported in parentheses below the coefficient, where standard errors are clustered by firm. ***/**/* means significance at 0.01/0.05/0.10 level. LPS_BERT and $LPS_BERT_{non-boilerplat}$ are measured using the BERT large language model and have been normalized to have a standard deviation of 1.

VARIABLES	(1) $PostRMSE$	(2) $PostRMSE$
LPS_BERT	-0.0594***	
	(5.489)	
$LPS_BERT_{non-boilerplate}$		-0.0551***
		(-5.009)
$Log(BtoM)$	-0.115***	-0.115***
	(-7.242)	(-7.242)
$Log(MVE)$	-0.186***	-0.185***
	(-10.62)	(-10.61)
$PreALPHA$	-0.623***	-0.625***
	(-13.60)	(-13.62)
$PreRMSE$	0.255***	0.255***
	(13.12)	(13.14)
$Abs(CAR_{0,1})$	2.413***	2.416***
	(6.018)	(6.025)
$NASDAQ$	0.0268	0.0256
	(0.566)	(0.541)
Firm Fixed Effect	Yes	Yes
Year Fixed Effect	Yes	Yes
Observations	38,695	38,695
Adj. R-Squared	0.651	0.651

E3. Impact of MD&A Language Predictability Estimated by BERT on Analyst Forecast Dispersion

The dependent variable is the analyst forecast dispersion. $Dispersion$ (defined as the standard deviation of analysts' forecasts for the next fiscal quarter's earnings). The t-statistics are reported in parentheses below the coefficient, where standard errors are clustered by firm. ***/**/* means significance at 0.01/0.05/0.10 level. LPS_BERT and $LPS_BERT_{non-boilerpl}$ are measured using the BERT large language model and have been normalized to have a standard deviation of 1.

VARIABLES	(1) $Dispersion$	(2) $Dispersion$
LPS_BERT	-0.0216**	
	(2.372)	
$LPS_BERT_{non-boilerplate}$		-0.0231**
		(-2.437)
$Log(BtoM)$	0.00414	0.00397
	(0.301)	(0.287)
$Log(MVE)$	-0.133***	-0.133***
	(-9.301)	(-9.302)
$Abs(CAR_{0,1})$	0.325**	0.326**
	(2.025)	(2.029)
$PreALPHA$	-0.501***	-0.501***
	(-9.492)	(-9.492)
$PreRMSE$	0.147***	0.147***
	(6.192)	(6.195)
$NASDAQ$	-0.104**	-0.104**
	(-2.332)	(-2.335)
$AnalystCount$	0.00572***	0.00570***
	(4.583)	(4.579)
Firm Fixed Effect	Yes	Yes
Year Fixed Effect	Yes	Yes
Observations	14,785	14,785
Adj. R-Squared	0.321	0.321

E4. Impact of MD&A Language Predictability Estimated by BERT on Analyst Forecast Accuracy

The dependent variable is the analyst forecast accuracy. AFE (defined as the absolute value of the analysts' forecast error of the next fiscal quarter's earnings); The t-statistics are reported in parentheses below the coefficient, where standard errors are clustered by firm. ***/**/* means significance at 0.01/0.05/0.10 level. LPS_BERT and $LPS_BERT_{non-boilerplat}$ are measured using the BERT large language model and have been normalized to have a standard deviation of 1.

VARIABLES	(1) AFE	(2) AFE
LPS_BERT	-0.0662***	
	(2.973)	
$LPS_BERT_{non-boilerplate}$		-0.0727***
		(-3.417)
$Log(BtoM)$	0.0264	0.0256
	(0.783)	(0.761)
$Log(MVE)$	-0.404***	-0.405***
	(-9.298)	(-9.314)
$Abs(CAR_{0,1})$	-0.763***	-0.764***
	(-5.714)	(-5.718)
$PreALPHA$	0.260***	0.260***
	(5.721)	(5.722)
$PreRMSE$	0.704	0.705
	(1.549)	(1.554)
$NASDAQ$	-0.0943	-0.0945
	(-0.996)	(-0.997)
$AnalystCount$	-0.00739**	-0.00746**
	(-1.983)	(-2.001)
Firm Fixed Effect	Yes	Yes
Year Fixed Effect	Yes	Yes
Observations	20,722	20,722
Adj. R-Squared	0.229	0.229