Bloated Disclosures: Can ChatGPT Help Investors Process Information?

Alex G. Kim, Maximilian Muhn, and Valeri Nikolaev
APRIL 2023

Becker Friedman Institute

FOR ECONOMICS AT **UCHICAGO**

Bloated Disclosures: Can ChatGPT Help Investors Process Information?

Alex G. Kim[*]   Maximilian Muhn[†]   Valeri V. Nikolaev[‡]

April 20, 2023

Abstract

Generative AI tools such as ChatGPT are expected to disrupt numerous industries and could fundamentally alter the way economic agents process information. We probe the economic usefulness of these tools in extracting information from complex corporate disclosures using the stock market as a laboratory. We use the GPT language model to summarize textual information disclosed by companies in their annual reports (MD&A) and during conference calls. Unconstrained summaries are dramatically shorter compared to the original disclosures, whereas their information content is amplified. When the originals have a positive (negative) sentiment, the summary becomes more positive (negative). More importantly, the summaries' are more effective in explaining stock market reactions to the disclosed information. Motivated by these findings, we propose a novel measure of disclosure "bloat." We show that bloated disclosure is associated with adverse capital market consequences, such as lower price efficiency and higher information asymmetry. Finally, we show that the model is effective at targeted summaries that distinguish between financial and non-financial (ESG) performance.

Keywords: ChatGPT, GPT, generative AI, large language model, informativeness, textual sentiment, MD&A, conference calls, summarization, bloated disclosure, information asymmetry

JEL Codes: C45, D80, G3; G11, G12, G14, M41

I Introduction

Textual data has become ubiquitous in economics and finance (e.g., Gentzkow et al., 2019; Goldstein et al., 2021). Yet, the vast amounts of textual data and their unstructured nature make it inherently difficult for economic agents to process information efficiently. Advanced language models such as GPT (ChatGPT) potentially enable decision-makers to analyze and interpret large volumes of text data to make more informed decisions. While the popularity of these tools has been growing exponentially since November 2022, their economic usefulness remains to be understood. Anecdotal evidence indicates a sharp interest of financial firms in language models. In a recent interview, Citadel's founder and CEO Ken Griffin indicated that ChatGPT technology fundamentally affects their business and that the company is negotiating an enterprise-wide license (Doherty and Marques, 2023). In this paper, we probe large language models' usefulness in extracting the most relevant information from corporate disclosures. In particular, we use GPT 3.5 Turbo to summarize information communicated by companies to their stakeholders. We then explore the information content of these summaries and construct a measure of the degree of redundant or less relevant textual information in corporate disclosures.

Corporate disclosures offer one of the best settings to study the usefulness of language modeling from a user's perspective. Preparers of financial statements have long been concerned about information overload in corporate filings, i.e., their excessive length and complexity (e.g. Loughran and McDonald, 2014a; Dyer et al., 2017). The usefulness of textual information is further diminished due to management's incentives to obfuscate negative information by providing irrelevant or immaterial details (e.g., Li, 2008). Regulators and standard setters have recognized these concerns and proposed a number of initiatives to promote more concise, accessible, and informative disclosures (SEC, 2013). These include the "Plain English" initiative and the development of more effective disclosure frameworks with the ultimate goal of improving the relevance of information communicated to stakeholders. Nevertheless, concerns about disclosure complexity and information overload still persist as financial statements have become longer and longer over the past 20 years and market reactions to their release have been substantially delayed (Cohen et al., 2020).

The GPT-3.5 (Generative Pre-trained Transformer) model, which has been the foundation for ChatGPT, is particularly well-suited for analyzing corporate disclosures due to its ability to summarize relevant information in a concise, effective, and understandable by a human manner. The model is pre-trained on a vast language corpus and then

fine-tuned for specific tasks, such as summarization or sentiment analysis. GPT is known to outperform other existing models in summarization tasks (Bhaskar et al., 2022; Goyal et al., 2022).[1] The transformer architecture relies on so-called attention mechanisms to identify relationships between words, sentences, and paragraphs in a document. This feature allows the model to generate summaries that retain the most relevant information.[2]

We use this technology to address the following questions. How effective are the language models, and GPT-3.5 in particular, at reducing the length of corporate disclosures? How does the information content of the condensed document compare to that of the original? To what extent do companies differ in terms of informational "bloat"? Are there capital market consequences of providing redundant information? Finally, can the model construct meaningful targeted summaries when investors are interested in a specific topic, such as ESG activities?

To answer these questions, we focus on two primary types of narrative corporate disclosures: Management Discussion and Analysis (MD&A) and earnings conference calls. The MD&A is a mandatory disclosure with the goal of helping investors view a company's performance through the eyes of its management. The SEC expressed repeated concerns with the informativeness of MD&A disclosures, e.g., in relation to boilerplate information or excessive complexity. In contrast, conference calls are voluntary disclosures held by companies to help investors process reported earnings and answer questions from analysts.[3] They are less subject to boilerplate concerns but possibly harder to summarize due to their less structured format. They also enable analysts to ask for clarification of technical aspects of financial disclosures or in response to the lack of clarity. Overall, both types of disclosures carry important information and are complementary.

We begin our analysis with a random sample that constitutes about 20% of the population of MD&As. We also randomly sample about 20% of conference call transcripts. We then instruct GPT-3.5 Turbo to produce an unrestricted summary of each document without referencing information from other documents or external sources. In subse-

---

[1]Prior-generation summarization models such as BART or PEGASUS require task-specific fine-tuning to achieve a reasonable performance (Zhang et al., 2020). GPT-3.0 overcame this challenge by dramatically increasing the size of the training data corpus. With these versatile language processing capabilities, GPT models can now perform zero-shot long text summarization on par with humans (Bhaskar et al., 2022; Goyal et al., 2022).

[2]We use GPT-3.5-turbo in our main analysis, which is known to produce very reliable summaries on general documents without further fine-tuning. Fine-tuning the model specifically to summarize corporate disclosures may yield even higher-quality summaries. Note, however, that fine-tuning GPT-3.5 is prohibited by OpenAI at this time.

[3]Even though conference calls are voluntary, more than 99% of US public firms host this event (NIRI, 2014).

quent analysis, we narrow down the prompt to retrieve information related to financial performance or ESG-related activities only.

We find that the model produces unconstrained summaries that are, on average, less than 20% of the original document's length. This is a striking result, which, taken at face value, points to potentially large gains in information processing. The key question here is whether the summaries, which are bound to omit many details, lose the bulk of their information content. In fact, we observe a slight reduction, on average, in the readability of the summarized document compared to the original. To determine whether this loss is consequential, we next focus on examining the information content.

From the perspective of an unboundedly rational investor with unconstrained information processing capacity (an analogy to a machine), the unconstrained document has at least as much information as the summary document. Such an investor will learn to filter out noise (like the model does) and undo any of possible biases in disclosed information. In this case, length, complexity, and information overload are irrelevant. For this reason, we focus on a scenario where an investor reads the document with the objective of extracting its general sentiment. A plausible scenario is that, when dealing with a lengthy document, an investor will largely focus on identifying positive vs. negative content. This approach is most closely captured by Loughran and McDonald (2011) measure of sentiment, which relies on positive/negative word counts in financial documents. We thus adopt this measure.[4]

We show that the average sentiment of the original MD&A (conference call) documents is negative (positive). This is intuitive as investor relation departments tend to be conservative in written disclosures, whereas executives prefer to convey confidence and optimism. More importantly, we find that the sentiment of the summarized document is more pronounced relative to the original. In particular, when the original sentiment is positive (negative), the summarised document is even more positive (negative). Such results can happen, for example, if companies (executives) "hedge" their views by precautionary statements that do not contain much information or are largely boilerplate.

While the summarized document can lose information, the above discussion suggests that summary-based sentiment may also be a cleaner measure of the true sentiment.[5] To test for information content, we examine stock market reactions to corporate disclosures.

---

[4]We specifically do not use the state of art measures of sentiment, e.g., machine learning-based sentiment scores (e.g., Frankel et al., 2022; Huang et al., 2023), because the objectives of such measures trained on datasets with billions of observations is to filter out redundant information, which is not what human readers are capable of doing.

[5]This would be the case when the original comingles positive and negative words or the same information is repeated multiple times.

Specifically, we regress short-window abnormal stock returns measured around the disclosure days on the sentiment scores. We find strong evidence that the summary-based sentiment is more informative in explaining stock market reactions to the disclosed information. As expected, we find a sharper increase in the informativeness for the MD&A summaries, which are more subject to concerns related to obfuscation or boilerplate language.

Our findings reveal a remarkable ability of the language model to condense information while maintaining and, in fact, uncovering its information content. The results also point to potentially considerable informational "bloat" in corporate disclosures. The next logical step is to ask whether companies differ in how bloated their disclosures are and whether the bloat causes some adverse capital market consequences. To address this question, we use the relative amount by which a document's length is reduced as a measure of the degree of redundant or irrelevant information, referred to as *Bloat*.[6]

We examine whether *Bloat* varies in time, by industry, and at the firm level by conducting a variance decomposition. For both MD&A and conference call samples, time, industry, and the interaction between time and industry fixed effects explain around 30-40% of the variation, whereas 60-70% of the variation in *Bloat* is firm-specific. Firm fixed effects explain around 26-30% of the firm-specific variation, which implies a substantial firm-year component. Indeed, we show that Bloat varies considerably from year to year within the same firm. We find that *Bloat* tends to be higher when a firm reports losses, has negative sentiment, and experiences negative stock market reactions, which are consistent with the obfuscation hypothesis in Li (2008).

We show that bloated disclosures are associated with adverse capital market consequences. We find that measures of stock price efficiency and information asymmetry deteriorate in the presence of bloated reporting. These results continue to hold when we control for conventional proxies of readability, which highlights the notion that our measure captures a different construct – the one that directly measures the relevance and redundancy of information instead of readability.

In our final set of tests, we explore the usefulness of the GPT model to produce

---

[6]Temperature parameter of the model may slightly influence the length of GPT-generated summaries. High temperature means that the model is allowed more discretion in generating summaries, while zero temperature means that the model will generate exactly the same summary no matter how many times we repeat the prompt. However, low temperature achieves high reproducibility at the cost of reliability. In our study, we set the temperature parameter to be 0.5. To show that the length of a summary does not dramatically vary across trials, we randomly select one conference call transcript with an original length of 41,252 and ask the model to summarize it 50 times. The mean value of summary length is 7,835 with a standard deviation of 150. The minimum length is 7,425, while the maximum length is 8,095. Once we scale the summary length by the original length, one standard deviation is equivalent to 0.36% of the original length.

targeted summaries. Combining its ability to "comprehend" text, GPT is capable of generating query-based summaries (in a way that is analogous to ChatGPT) to extract theme-specific information without further training (Yang et al., 2023). In particular, investors may be interested in understanding information concerning a specific topic, such as environmental impact or regulatory uncertainty. To explore this, we create prompt-based summaries that retain information about (1) financial performance and (2) "ESG" based on the original summaries. In fact, we do not clarify the meaning of "ESG" but instead choose to rely on the model's ability to successfully comprehend this concept. Intuitively, we find an increasing time trend in ESG-related content throughout our sample period in both MD&A and conference call communications. Furthermore, we find that both summaries are incrementally informative and thus capture different dimensions communicated by the firms. As one would expect, the sentiment from ESG-specific summaries is increasingly more important in determining stock market reactions over time. This finding is consistent with prior studies showing that ESG risks are priced in more recent years (Giglio et al., 2021; Sautner et al., 2023). Overall, the model shows significant promise in extracting targeted information from lengthy and hard-to-read financial disclosures.

We contribute to the literature in four ways. First, we contribute by establishing the economic usefulness of generative AI-based techniques in analyzing and summarizing unstructured textual data in a useful way.[7] Due to technological constraints, providing reliable summaries of long text documents has not been feasible until recently. We provide preliminary evidence of dramatic reductions in the length of disclosed information while maintaining and enhancing its information content. In that sense, our findings are relevant not only to academics but also to regulators and investors. As investors often face significant disclosure information processing costs (e.g., Sims, 2003; Blankespoor et al., 2020; Cohen et al., 2020), generating and consuming automatic summaries of complex financial disclosures via ChatGPT could support their decision-making. We show that these summaries retain and amplify information content, thus, allowing investors to focus on information that truly matters. This suggests that generative AI systems have the potential to be the next meaningful advance in financial reporting technology, in line with prior advances such as the introduction of EDGAR by the SEC (e.g., Gao and Huang, 2019; Chang et al., 2022).[8]

---

[7]Recent studies using non-generative large language models for different purposes, such as creating improved sentiment measures or conducting other optimization tasks via trained BERT or GPT-3.0 models include Kim and Nikolaev (2022); Huang et al. (2023) and Bernard et al. (2023)

[8]Our results are related to but different from Cardinaels et al. (2019), who show in an experimental setting that automatically generated summaries of earnings releases via LexRank are more neutral than

Second, we contribute by developing a novel measure of the degree to which textual information contains redundancies and excessive details. Most prior studies on textual disclosures focus on readability or linguistic complexity (e.g., Li, 2008; Loughran and McDonald, 2014a; Bonsall IV et al., 2017). In contrast, disclosure "bloat" is an intuitive, easy-to-implement, and distinct construct with a substantial firm-level variation. As discussed by Loughran and McDonald (2014a), linguistic complexity or readability commingles both textual and content complexity. Our measure is subject to this issue to a much lesser extent because GPT is trained to understand and summarize complex content while omitting less relevant information. The measure can be easily applied in various corporate contexts and is of interest to investors and regulators.

Third, and more broadly, our study adds to the literature on disclosure quality and its economic consequences (e.g., Leuz and Verrecchia, 2000; Balakrishnan et al., 2014) We show that companies with bloated disclosures exhibit lower (higher) price efficiency (information asymmetry). This aspect of disclosure quality has not received attention in the prior literature. Finally, we contribute by establishing the value of language models in extracting targeted and standardized information, e.g., environmental impact, from general-purpose corporate disclosures (e.g., Hassan et al., 2019; Sautner et al., 2023; Florackis et al., 2023). For example, measuring a company's environmental performance is a highly complex task, and AI-based language models show promise in helping investors and regulators understand and process it more efficiently.

II Generating Summaries with GPT

GPT is a large language model with billions of parameters trained on a vast corpus of text data. Its goal is to predict the next word in a sentence (text) conditional on the prior words. In this section, we describe how the GPT generates summaries. We then motivate and provide the rationale for our disclosure *Bloat* measure.

A The Transformer Architecture

GPT is based on the highly influential Transformer architecture developed by Vaswani et al. (2017), Radford et al. (2018, 2019), and Brown et al. (2020). The Transformer is a type of neural network capable of modeling long-range dependencies among words in a text (or sequence). Each word (token) is represented by an m-dimensional vector, $x_k = (x_k^1, x_k^2, \cdots, x_k^m)$, referred to as word embedding. The model thus treats a text as $n \times m$ matrix, $X = (x_1, x_2, \cdots, x_n)'$. The number of rows in a matrix corresponds to the

the original document (which then lowers the participants' valuation of the firm).

number of words (tokens), and the number of columns is the dimension of word embeddings. For instance, the sentence "*Compared to our competitors, our company is dedicated to promoting sustainable technologies such as renewable energies and net-zero plan.*" has nineteen tokens. Assuming that a token is modeled as a 100-dimensional vector $(x_i^1, x_i^2, \cdots, x_i^{100})$ $(1 \le i \le 100)$, this sentence is represented by a 19×100 matrix.

The central component of the Transformer architecture is the so-called self-attention mechanism. It enables the model to learn the relevance (relative importance) of each word in an input text by considering its positional and contextual relationships with other words. To capture the relationships between different words and extract the most relevant information, GPT calculates self-attention scores. This process involves query (Q), key (K), and value (V) matrices parametrized as: $Q = X \cdot W_Q$, $K = X \cdot W_K$, and $V = X \cdot W_V$. Each row in the query matrix (i.e., a query vector) corresponds to a token for which we want to calculate the attention score. The query vector is used to compare the current token to other tokens based on their key vectors. Accordingly, each row in the key matrix (i.e., a key vector) represents a token that we want to compare the current query token against. Lastly, each row in value matrix represents the information contained in the corresponding token. The model learns the weight matrices $W_Q (m \times \dim Q)$, $W_K (m \times \dim K)$, and $W_V (m \times \dim V)$ from the pre-training phase (note that $\dim Q$ and $\dim K$ are chosen to be the same).

As an illustration, consider the 19-token sentence we saw previously. To calculate the self-attention of the word *sustainable*, the model measures its relation with every other word (including itself). Specifically, the query vector corresponding to *sustainable* is compared to 19 key vectors by calculating the dot products for each *query-key* pair. The dot-product captures the correlation between two vectors, i.e., the semantic similarity between a pair of tokens. Mathematically, this is expressed as $Q \cdot K'$, which is, in our case, a 19×19 relation score matrix. The element of this matrix at the intersection of ith row and jth column, $[Q \cdot K']_{ij}$, measures the similarity between the ith and jth words.

More formally, the model calculates the attention matrix to capture inter-relatedness among tokens:

$$\text{Score}(Q, K) = \text{softmax}\left(\frac{Q \cdot K'}{\sqrt{\dim K}}\right) \tag{1}$$

where $Q \cdot K'$ is a $k \times k$ matrix of semantic proximity between queries and keys, $\sqrt{\dim K}$ is a normalizer, and softmax is a function that maps row vectors into weights that sum up to one. The attention matrix above is then post-multiplied by the value matrix, $\text{Score}(Q, K) \cdot V$, to obtain the weighted sum of the value vectors and is an output of

the attention layer.

The model is pre-trained on a large corpus to learn word embeddings and to be able to calculate self-attention matrices for variable-length text sequences. This information is passed on to the "decoder loop" to generate a sequence of words to be included in a text, e.g., the summary. In doing so, the model searches for the most probable next word and it does so autoregressively, i.e., conditional on all prior words while relying on the pre-calculated self-attention scores and the entire word corpus $\mathcal{L}$. More formally, the model thus calculates the conditional probability distribution over y_{n+1} given by $p(y_{n+1}|y_1, y_2, \cdots, y_n, \mathcal{L})$, where $(y_1, y_2, \cdots, y_n)$ represent words already included in the summary, and chooses the most likely next word.

Returning to our example, the summary sentence starts with the word *"Our"* because it has the highest self-attention score.[9] The second most likely word, conditional on the first, is "*company*" (subject) and so on. At some point, the most likely token predicted by the model is the end of the sentence token, which completes the summary and renders the following sentence: "*Our company promotes sustainable technologies.*"[10] This result retains the most relevant information and leaves out redundancies or unnecessary details.

B Conceptual Underpinning of Disclosure Bloat

The process of constructing an effective (optimal) summary involves a trade-off between relevant information (signal) and unnecessary details or redundancies (noise). Intuitively, more relevant words should be included and less relevant words should be omitted. GPT's self-attention mechanism enables this trade-off when producing an optimal summary. To capture this idea, we offer a stylized model of the optimal information summarization, which, in turn, motivates our measure of informational bloat.

Let the k-th row $[QK]_k$ ($1 \leq k \leq n$) of the matrix $\text{Score}(Q, K) = \text{softmax}\left(\frac{Q \cdot K'}{\sqrt{\dim K}}\right)$

---

[9]It is very natural that *Our* receives a high self-attention score. *Our* is associated with the main verb *dedicate* and its objective in the sentence. Furthermore, considering that it is the possessive pronoun of the subject in the sentence, its positional importance is also very high.

[10]GPT allows a researcher to adjust its temperature. Higher temperature means that GPT has a higher degree of freedom in selecting the words from the pre-trained corpus. In this example, when we allow GPT a higher temperature, it searches its own vocabulary dictionary $\mathcal{L}$. Now, the summary becomes "*Our company considers environmental issues seriously.*" Note that the words *environmental* or *seriously* did not appear in the original text. However, the model chooses the most appropriate words from its pre-trained corpus to complete a sentence. Setting a higher temperature, therefore, may yield powerful and informative summaries. However, an excessively high temperature may make the model place too much weight on its own corpus rather than focus on the pre-calculated self-attention scores from the given text. This setting might make the summary inaccurate. Therefore, there is a trade-off between the model's creativeness and reliability when increasing the temperature parameter. We revisit this issue in Section III.

be the self-attention score vector corresponding to word embedding x_k. Denote $s(k) : R^{\dim V} \to [0, 1]$ to be the arithmetic mean of the elements in the self-attention score vector of x_k:

$$s(k) = \frac{1}{\dim V} \sum_{i=1}^{\dim V} q_{ki} \qquad (2)$$

By construction, $s(k)$ varies between 0 and 1, with higher values indicating that x_k has higher information relevance. Now we sort the elements of function $s(k)$ to have an ascending order, i.e., $\{s(1), \cdots, s(n)\} \to \{\bar{s}(1), \cdots, \bar{s}(n)\}$ such that $\bar{s}(k) \leq \bar{s}(k+1)$.

We can then define the "signal-to-noise ratio" for a text that includes k tokens as:

$$SNR(k) = \frac{\sum_{i=1}^{k} \bar{s}(i)}{\sum_{i=1}^{k} (1 - \bar{s}(i))} \qquad (3)$$

The marginal benefit of including a word k is an increase in the total "signal" by $\bar{s}(k)$ while the marginal cost is the inevitable increase in "noise" by $1 - \bar{s}(k)$. An optimal summarization algorithm would include words as long as the marginal benefit exceeds the marginal cost. Since $\bar{s}(k)$ is by construction an increasing function of k, there exists a shortest optimal length n^* that maximizes $SNR(k)$.

Figure 1 visually illustrates the relationship between summary length and the signal-to-noise ratio. As the model starts to include word embeddings with the highest informational value, the overall signal-to-noise ratio begins to increase. This continues until we reach the inflection point n^* such that $\bar{s}(n^*) = 1 - \bar{s}(n^*)$. If the summary becomes longer than n^*, the overall signal-to-noise ratio declines.

Note that the optimal length n^* depends on the functional form of $\bar{s}(k)$. For example, if $\bar{s}(k)$ is convex, the optimal summarization point n^* is smaller than $\frac{1}{2}n$. Intuitively, convex $\bar{s}(k)$ implies that there are relatively fewer important tokens than unimportant tokens, leading to a shorter summary. For a concave $\bar{s}(k)$ the optimal summarization point is larger than $\frac{1}{2}n$. Indeed, when many word tokens convey relevant information, the model generates a longer summary.[11]

Bloated disclosure has a high frequency of tokens with less important information content, thereby leading to a convex $\bar{s}(k)$. Based on this idea, we construct the *Bloat*

---

[11]Consider some hypothetical cases that help to understand our algorithm. When $\bar{s}(0) = \cdots = \bar{s}(n) = 1$, all word embeddings convey perfect signals. In this case, the model finds a corner solution and includes all word embeddings in the summary. When $\bar{s}(0) = \cdots = \bar{s}(n-1) = 0$ and only $\bar{s}(n) = 1$ (i.e., extreme convexity), the last token conveys all the signal while other tokens convey only noise. In this case, the model uses the last token only to generate a summary.

measure as follows:

$$Bloat = \left(\frac{n - n^*}{n} \right) \tag{4}$$

The spirit of this measure is to capture how far the document deviates from its optimal summary length n^*, i.e, the length that maximizes the signal-to-noise ratio. Higher *Bloat* implies that the original disclosure text contains more relatively uninformative, repetitive, or irrelevant content.

III   Data and GPT Processing

Sample Selection. To probe the value of generative AI for processing financial information, we focus on the two most prominent types of narrative disclosures: the management discussion and analysis (MD&As) section in firms' annual reports and earnings conference calls. Both types of narrative disclosures are known to contain relevant information (e.g. Hassan et al., 2019; Cohen et al., 2020; Li et al., 2021). To construct our sample, we obtain machine-readable MD&A sections of annual reports and earnings conference call transcripts for all US public firms from 2009 to 2020. We remove firms from the financial sector (SIC code starting with 6) or firms with missing values for key variables (e.g., stock returns or analyst forecasts). These filtering steps leave us with an initial sample of 8,699 MD&As and 42,238 conference calls.

Due to GPT processing constraints, we then draw a random sample of about 20% relative to the population described above. This results in 1,790 MD&As reported by 339 firms and 8,907 conference calls held by 367 firms. Additional details on each type of disclosure are provided below.

Management Discussion and Analysis. MD&As are included as "Item 7" of the 10-K filings. We download all 10-K filings (including 10-K/A and 10-KSB) from EDGAR and use regular expressions to extract Item 7 (Kim and Nikolaev, 2022). Out of 10,588 10-K statements, we successfully retrieve 8,699 MD&A sections (approximately 82.16%). Several firms do not have correct item numbers for MD&A sections or simply include a hyperlink that directs viewers to an external website. These outliers are not collected by our automated MD&A extraction algorithm. Our retrieval rate is comparable to other studies (e.g. Cohen et al., 2020). We delete figures, tables, html script, and XBRL tags.[12]

---

[12]Unlike other natural language models such as BERT, we do not need to replace numbers with placeholder tokens as GPT is capable of understanding the contextual meaning of numbers used in texts.

Earnings Conference Calls. Earnings conference calls are quarterly events voluntarily held by firms. Almost all US public firms hold quarterly earnings conference calls (Hassan et al., 2019). We use the entire conference call transcripts as inputs. We do not exclude operator instructions or questions from analysts as they also convey information about the speakers and the purpose of the speech.[13]

Other Data. We use the CRSP database for daily stock and market returns. We rely on Trades and Quote (TAQ) database to calculate the probability of informed trades (*PIN*). Financial characteristics such as total asset, net income, filing dates, and earnings announcement dates are from Compustat. Analyst forecasts are from I/B/E/S Details file. Finally, the institutional holdings data is from Thomson Reuters 13-F filings.

GPT Processing. We use GPT-3.5-Turbo API provided by the OpenAI ChatCompletion endpoint to construct summaries.[14] GPT-3.5-Turbo allows 4,096 tokens per request. For longer documents, following Ramshaw and Marcus (1999), we divide the document into chunks not to exceed 4,096 tokens and process each chunk separately.[15] We then concatenate the generated summaries of each chunk and create a single summary for the entire document.

We use a prompt that instructs the model to summarize the input text using only the information included in the text.[16] We instruct the model not to restrict the length of a summary. See our internet appendix for the exact prompts that we use for each task.[17]

Random Sampling. As GPT involves significant time and resources to generate summaries, we only process a random sample of approximately 20% of all documents. Specifically, we obtain a list of the 1,694 (1,834) unique firms in the MD&A sample (conference call sample) and then randomly choose 20% of these unique firms (339 and 367 firms, respectively). We then retain all documents associated with these firms for GPT processing.[18]

---

[13]Since transcripts contain utterances or onomatopoeia (e.g., 'um', 'ph') in square brackets, we delete all expressions within the square brackets from the transcripts.

[14]GPT-3.5 underpinned the initial version of ChatGPT from November 2022. A more recent version from March 2023 relies on GPT-4.0

[15]Using SpaCy sentence tokenizer, we do not allow a single sentence to be divided into two different chunks. Furthermore, for conference call transcripts, since each turn-at-talk is marked with a new line separator, we do not allow a single turn-at-talk to be included in two separate chunks.

[16]Our prompt only exploits the user role of the API and does not specify a system or assistant role.

[17]For our model parameters, we set none for max_tokens, 0.5 for temperature, 1.0 for top_p, 0.0 for presence penalty, and 0.0 for frequency penalty.

[18]Compared to randomization on the document level, this cluster randomization allows us to track

Table 1 provides summary statistics for our sample vs. the population, reflecting firm size and its information environment. In both the MD&A and conference call samples, the randomly chosen sets of firms do not significantly differ from their populations and hence are representative. Additionally, the MD&A and conference calls samples do not differ significantly from each other. For example, the mean log(*ME*) of the MD&A sample is 14.58, while the mean log(*ME*) of the conference calls sample is 14.68.

IV How effective are the summaries?

In this section, we provide descriptive evidence on the attributes of the summarized disclosures. We subsequently analyze the information content of the summaries relative to that of the original documents.

A Length, Sentiment, and Readability

Measurement. We investigate the textual properties of GPT summaries by focusing on three dimensions: length, sentiment, and readability (Fog index and plain English Measure).[19] Length is measured by the number of words contained in a given document. The sentiment is based on financial keyword dictionaries provided by Loughran and McDonald (2011):

$$Sentiment = \frac{\sum_{x \in \mathcal{D}} \mathbf{1}[x \in \mathcal{P}] - \sum_{x \in \mathcal{D}} \mathbf{1}[x \in \mathcal{N}]}{\sum_{x \in \mathcal{D}} \mathbf{1}[x \in \mathcal{P}] + \sum_{x \in \mathcal{D}} \mathbf{1}[x \in \mathcal{N}]} \tag{5}$$

where $\mathcal{D}$ is a set of given document, $\mathcal{P}$ is a set of positive financial keywords, $\mathcal{N}$ is a set of negative financial keywords, x is an individual word, and $\mathbf{1}[\cdot]$ is an indicator function. A higher proportion of positive keywords corresponds to a higher *Sentiment* score.

We use *Fog* index (Gunning et al., 1952), which measures the percentage of complex words per sentence, as a measure of readability (e.g. Li, 2008; Loughran and McDonald, 2016). A higher *Fog* indicates less readable disclosures. We supplement Fog index by the "plain English" measure (*Plain_Eng*) calculated in accordance with the narrative disclosure guidelines set forth by the SEC.[20] Analogous to *Fog*, hihger *Plain_Eng* indicates

---

changes in the summaries within-firm across time.

[19]Another commonly used proxy, the Bog index, is only readily available for full 10-Ks via Bonsall IV et al. (2017). Due to processing constraints of the commercial software *StyleWriter*, it is not trivial to create this measure for our set of MD&A and earnings call transcripts.

[20]Specifically, following Loughran and McDonald (2014b), we calculate the following six components: (i) the average number of words per sentence, (ii) the average number of characters per word, (iii) the number of passive voice verbs, (iv) number of pronouns, (v) number of legal expressions, and (vi) other characteristics (e.g., number of the word "respectively", number of negative expressions, etc.). We then

lower readability.

Linguistic Attributes. We calculate the linguistic attributes for both raw and summarized documents and present them in Table 2 and Figure 2.

Panels A and B of Table 2 and Figure 2(a) focus on the MD&A sample. Panel A indicates that the GPT model reduces the document length of MD&As by more than 80%, on average. The average (median) length of the original documents is 79,310 (71,264) words whereas it goes down to 14,974 (13,543) for summarized documents. This difference is economically large and indicates potentially sizeable efficiency gains for a human reader.

Next, we observe that the average textual sentiment of raw MD&As is negative (-0.249) and that the average tone becomes even more negative in the summarized documents (-0.300). More importantly, the standard deviation of sentiment increases from 0.203 (raw) to 0.265 for summarized documents. This widening of the distribution of sentiment is indicative of sentiment becoming more clear-cut.

In terms of textual complexity, the average Fog index of raw documents is 10.03 and the average plain English measure of raw documents is -0.30. Based on common rule-of-thumb rules, both scores indicate that the documents are "highly complex". Summarized documents seem to become slightly more complex in general, with an average Fog index of 10.50 and a plain English measure of -0.21. This phenomenon likely arises because summaries inevitably need to include financial jargon in a relatively shorter document.

Panel B of Table 2 also reveals substantial heterogeneity in the changes in length, sentiment, and readability across MD&As.

We repeat the same analysis for the conference calls sample and report the results in Panels C and D of Table 2 as well as Figure 2(b). One notable difference is that conference call transcripts have, on average, positive textual sentiment unlike MD&As. Overall, the inferences remain qualitatively similar. Notably, the summarized sentiment continues to have a higher standard deviation than raw sentiment and the summaries become slightly less readable.

B   Detecting Positive vs. Negative Sentiment

The widening of the distribution of sentiment in the summary documents raises the question of whether the summaries capture the sentiment of the original document in a more definitive (precise) way. To test whether there is an amplification of the sentiment in the summarized document, we split our sample based on the median value of raw sentiment.

---

standardize and add all six elements to obtain *Plain_Eng*.

Table 3 reports these results for MD&As (Panel A) and conference calls (Panel B). Panel A indicates that when the original's sentiment is above its median value, the summarized document becomes more positive compared to the original. The last column of the panel indicates that this difference is statistically significant. In contrast, when raw sentiment is below the median value, the average summarized sentiment is significantly more negative than the original sentiment. Figure 2(a) helps to visualize these findings. As illustrated, we find that the GPT summarization makes relatively positive documents more positive and relatively negative documents more negative. These results support the notion that GPT summaries amplify the tone of the underlying raw documents.

In terms of readability measures, for the above-median sentiment documents, the summaries exhibit a higher Fog and plain English measure (10.61 and -0.23, respectively) compared to the raw documents (8.19 and -0.36, respectively). Interestingly, for the below-median documents, the summaries are now less complex in terms of Fog index. The average Fog index of summaries is 10.40, which is smaller than that of raw documents (11.86). Comparing the two MD&A partitions yield some more insights into this finding. In line with the managerial obfuscation hypothesis in (e.g. Li, 2008), firms issue longer and more complex (raw) reports when they deliver negative news. At the same time, the summaries exhibit a relatively more consistent length and readability across the two sub-samples. Combining these observations with previous findings, GPT appears to successfully navigate through the fog in corporate disclosures producing consistent-quality summaries.

Next, we turn attention to the analysis of the conference call reported in Panel B of Table 3 and Figure 2(b). As in the MD&A sample, we observe that documents with above-median raw sentiment become more positive in their summaries and vice versa. Specifically, the average summarized sentiment of the above-median group is 0.554, while the average raw sentiment of the same group is 0.329 (with the difference being statistically significant at the 1% level). In contrast, the average summarized sentiment of the below-median group 0.140 is slightly smaller than the corresponding raw sentiment (0.17). Overall, we observe the same asymmetric effect of summarization on sentiment in the conference call sample.

Like in the MD&A sample, the summarized documents also exhibit an increase in their Fog index. The resulting summaries are ultimately also fairly comparable in terms of their length (7,807 for above-median and 7,518 for below-median) and readability (Fog index of 11.45 versus 11.351 and plain English measure of -0.69 versus -0.72).

Taken together, we find evidence that summarization provides a "high-resolution" textual sentiment. As in prior studies (e.g. Li, 2008), we document that narrative dis-

closures dealing with negative news tend to be lengthy and complex. GPT appears to filter out this noise by making negative documents more negative and positive documents more positive, i.e., amplifying their information content. Summarized documents become slightly less readable, however, likely due to financial jargon inevitably being included in shorter-length summaries. To shed further light on the informational value of the summarized disclosures, we examine their information content more formally in the following subsection.

C Informativeness of Summarized Disclosures

A natural question arising from our previous findings is whether GPT summaries are indeed more informative compared to the originals. As discussed previously, this question is only meaningful from a standpoint of a user with a limited information processing capacity (e.g. Maćkowiak et al., 2023)). We take a perspective of an average investor who reads MD&As (conference call transcripts) with the objective of learning its general sentiment.

Test Design. To test this, we compare raw document sentiment vs. summarized document sentiment in explaining stock price reactions to disclosures. Specifically, we estimate the following regression:

$$CAR_{[0,1]it} = \beta Sentiment_{it}^{j} + \gamma \mathbf{X}_{it} + \delta_t + \theta_i + \varepsilon_{it} \tag{6}$$

where $CAR_{[0,1]it}$ is firm i's cumulative abnormal returns over the two-day window (starting from the 10-K filing date for MD&As and conference call date for conference calls) at time t. We calculate abnormal returns by subtracting the value-weighted market returns. $Sentiment_{it}^{j}$ is either $Sentiment_{it}^{Raw}$ (raw sentiment) or $Sentiment_{it}^{Sum}$ (summarized sentiment) of firm i at time t. $\mathbf{X}_{it}$ is a vector of firm-level control variables that include the natural logarithm of market capitalization (log(ME)), the natural logarithm of book-to-market ratio (log(BE/ME)), institutional holdings ($Inst\_Own$), and scaled earnings surprise (SUE). δ_t represents time-fixed effects (year fixed effects for the MD&A sample and year-quarter fixed effects for the conference call sample). θ_i represents either firm fixed effects or industry fixed effects at the two-digit Standard Industry Classification (SIC) code level. All continuous variables are winsorized at 1% and 99% to mitigate the influence of outliers. Standard errors are clustered at the industry level.

Results. Table 4 reports our findings. We start with the MD&A sample (Panel A). Columns (1)-(4) examine the informativeness of raw sentiment. We observe weak and

mostly insignificant associations of raw sentiment with contemporaneous abnormal returns, in line with Frankel et al. (2022). This can happen if MD&As are too "noisy," or if the market had already anticipated and incorporated all information contained in MD&As into prices before its release. The following analysis indicates that the former is a more appropriate explanation.

Columns (5)-(8) rely on the summary-based sentiment as our explanatory variable. We find a striking result that $Sentiment^{Sum}$ exhibits economically strong and highly statistically significant associations with abnormal stock returns. This holds across all models. For example, the coefficient on $Sentiment^{Sum}$ is 0.030 (t-value = 3.69) for the model that includes year and industry fixed effects and 0.027 (t-value = 3.99) when we include year and firm fixed effects. The results also show considerable economic magnitudes. A one standard deviation increase in $Sentiment^{Sum}$ is associated with a 0.087 standard deviation increase in abnormal returns (or 80 basis points). We also find similar results when we partition the sample based on the sign of summarized sentiment.[21]

Overall, we observe a remarkable contrast between the sentiment of the summarized versus the original document in explaining market reactions. Note that, unlike more sophisticated machine-learning-based measures of sentiment (e.g. Frankel et al., 2022), our sentiment was not pre-trained to explain stock returns.

We then turn to the analysis with the conference calls sample tabulated in Panel B. In this sample, raw sentiment exhibits significant positive associations with stock price movement around the conference call date. One exception is that, when we partition the sample based on the sign of raw sentiment, the raw sentiment loses its statistical significance for the below median sub-sample (column (4)).

Nevertheless, columns (5)-(8) indicate that the summary sentiment has stronger associations with abnormal returns, highly significant throughout all specifications. For example, in our most stringent specifications with firm fixed effects (column 2 vs. column 6), the coefficient goes up from 0.065 to 0.077, and the t-value increases from 9.51 to 14.35. The economic magnitude of summarized sentiment implies that a one standard deviation increase in $Sentiment^{Sum}$ is associated with a 0.25 standard deviation increase in abnormal returns. Furthermore, there is a remarkable increase in adjusted R-squared from 4.2% (raw sentiment) to 12.9% (summarized sentiment) for the same model.

We visualize these findings by plotting the average cumulative abnormal returns across sentiment quintiles in Figure 3. For each period, we partition the sample into

---

[21]Since the summarized tone of MD&A explains the contemporaneous cumulative abnormal returns, it is not likely that all the information embedded in MD&A is already reflected in market prices on the filing date.

quintiles based on the value of raw or summarized sentiment. Then we calculate the average cumulative abnormal returns for each quintile and plot the results. The dotted line in Figure 3(a) (the MD&A sample) represents raw sentiment, and the solid line represents the summary sentiment. We do not observe a clear trend in the raw sentiment. In contrast, there is a clear trend in the summary-based sentiment. We repeat this exercise with the conference calls sample (Figure 3(b)). Here, we observe a positive slope for both lines, consistent with our regression results. However, the solid line is steeper, thus confirming a stronger positive association between summarized sentiment and abnormal returns.

Taken together, we find strong support for the claim that language models generate economically useful summaries representing complex textual disclosures in a concise way while retaining and amplifying the information content.

V Measuring Disclosure Bloat

The flip side of GPT's impressive performance in summarizing complex disclosures while preserving the information content is that one can quantify the degree of redundant (or less relevant) information in the original disclosure. Motivated by the conceptual underpinnings in Section II.B, we introduce such a measure. Specifically, *Bloat* takes the difference between the length of the original document and that of its summary (scaled by the length of the original). Higher *Bloat* is associated with a higher degree of noise in the original document. In this section, we analyze *Bloat* and explore its determinants and capital market consequences.

A Sources of variation in Bloat

Descriptive Statistics. In Table 5, Panel A, we provide descriptive statistics for our measure. The average *Bloat* for the MD&A (conference call) sample is 0.81 (0.82), i.e., 81% (82%) of the original. In other words, GPT considers approximately 80% of the disclosed content to be less relevant when trading off its "signal" and "noise" components.[22] We also observe a meaningful variation in *Bloat*, which is 0.088 (0.102) for the MD&A sample (conference call sample).

Figures 4(a) and 4(b) plot the average *Bloat* for the MD&A conference calls samples, respectively. We observe several interesting patterns. First, although there is no clear increasing or decreasing time trend, there are over-time fluctuations in *Bloat* for both

---

[22]Note that this does not necessarily indicate that 80% of the disclosure is pure noise.

samples. Second, we do not observe a high correlation between the changes in disclosure length and *Bloat*. Third, *Bloat* measures computed for MD&A and conference call samples exhibit some correlation over time (the Spearman correlation coefficient between annual numbers is 0.44). Lastly, in 2020, we note a steep decline in *Bloat* and a steep increase in length for both samples. This phenomenon is likely driven by corporate disclosures highlighting the impact of COVID-19. Since COVID-19 cannot be "stale information" at the time and since it often had a material impact on firms' business, GPT is likely to find such information relevant and include it in its summaries.[23]

Bloat Variance Decomposition. We next examine how much variation in *Bloat* can be attributed to time-, industry-, or firm-specific factors. Panel B of Table 5 reports incremental adjusted R-squared after adding different sets of fixed effects. For the MD&A (conference call) sample, year fixed effects explain only 0.56% (1.32%) of the total variation in *Bloat*. The inclusion of industry fixed effects increases explained variation to 7.90% (3.93%). A more sizable portion of variance, however, is explained by the interactions between industry and time fixed effects, which account for an additional 26.34% (20.59%) of the total variance. This implies that the remaining 68.20% (74.16%) of the total variance is attributable to firm-level factors. We zoom in on this firm-level variation at the bottom of Panel B. We show that firm-fixed effects, i.e., time-invariant firm characteristics explain about 32.33% (26.04%) of firm-level variance. This means that roughly two-thirds (three quarters) of firm-level variation in *Bloat* it time-varying.[24].

Stickiness in Bloat. To shed additional light on firm-level variation in Bloat, we rank firms into quintiles (each period) and then measure the frequencies with which they transition across quintiles in the subsequent period. Panel C focuses on the MD&A reports, whereas Panel D focuses on the conference calls. For instance, the *i*-th row and *j*-th column (c_{ij}) of Panel C shows the fraction of firms that moved from the *i*-th quintile in year $t-1$ to the *j*-th quintile. The diagonal elements show the frequency with which a firm stays in the same quintile.

---

[23]Indeed, we manually check several summaries for 2020 and find that almost all of them include COVID-related information to some extent. We interpret this finding as anecdotal evidence that GPT is capable of extracting new, material information and including it in its summaries.

[24]However, one alternative explanation for large unexplained variation by firm fixed effects is that our measure potentially has high measurement error. To partially address this issue, we follow Hassan et al. (2019) and estimate the measurement error associated with *Bloat*. Specifically we regress $Bloat_{it}$ on $Bloat_{it-1}$ and obtain the coefficient $\hat{\beta}^{OLS}$. Then we use $Bloat_{it-2}$ as an instrument of $Bloat_{it-1}$ and obtain two-stage least squares estimator $\hat{\beta}^{IV}$. The measurement error is $1 - \frac{\hat{\beta}^{OLS}}{\hat{\beta}^{IV}}$. The estimated measurement error is 7.46% for MD&A *Bloat* and 6.15% for conference call *Bloat*.

Both Panels C and D indicate that only between a quarter to a third of companies stay within the same quintile from period to period. Therefore, bloat in corporate disclosures tends to change considerably within the same firm from period to period.

B Bloat Determinants

Test Design. We next examine the economic determinants of *Bloat* by estimating the following OLS regression:

$$Bloat_{it} = \gamma \mathbf{X}_{it} + \delta_t + \theta + \varepsilon_{it} \tag{7}$$

where $\mathbf{X}_{it}$ is a vector of firm-level determinants. We include the natural logarithm of market capitalization (log(*ME*)) and the natural logarithm of book-to-market (log(*BE/ME*)) as proxies for firm size or growth opportunities, respectively. We also include the number of analysts following a firm (*N_Analyst*) and institutional holdings (*Inst_Own*) to capture differences in the demand for information across companies. We use the lag between the fiscal period end and the filing date (*Report_Lag*) as a proxy for reporting timeliness and complexity. More complex reports are likely to have longer reporting lags. Next, we include variables that capture firms' performance and its variability: buy-and-hold returns around the filing or conference call date (*One_Day_Ret*), the textual sentiment of raw disclosure (*Sentiment^{raw}*), an indicator for whether a firm reports negative earnings (*Loss*), and earnings volatility (*Earn_Vol*). Broadly, the goal of adding these variables is to capture reporting complexities that come with differences in performance and also possible incentives to obfuscate disclosures when things are not going well. We also include textual attributes: *Fog* and *Plain_Eng* proxy for readability, and *Log_Length* proxies for document complexity. Finally, δ_t denotes time (year or year-quarter) fixed effects and θ denotes industry or firm fixed effects. All continuous variables are winsorized at 1% and 99%. Standard errors are clustered at the industry level.

Results. We report the results in Table 6. Columns (1) and (2) focus on the MD&A sample. Based on column (1), which conditions on industry and year fixed effects, we find statistically significant coefficients on log(*ME*), *N_Analyst*, *Loss*, *Plain_Eng*, *Sentiment*, *One_Day_Ret*, *Report_Lag*, and *Log_Length*. A negative coefficient on log(*ME*) implies that larger firms have less bloated disclosures on average. We further find a positive coefficient on *Loss* and negative coefficients on *One_Day_Ret* and *Sentiment*. Collectively, these relations suggest that *Bloat* increases as performance declines. Note, however, that there are two plausible explanations for this finding. *Bloat* may increase as if negative per-

formance is more difficult to explain. Alternatively, managers may obfuscate the bad performance (e.g. Li, 2008; Loughran and McDonald, 2014a)).

We find a positive coefficient for *N_Analyst*, which is somewhat counterintuitive and could potentially be driven by a firm's need to cater to multiple audiences (and, hence, being more inclined to include more redundant information in their narrative disclosures). Lastly, we observe that document length, *Log_Length*, has a positive and significant relation with *Bloat*. However, economically, the effect is still modest, as a 10% increase in length corresponds to a 0.002 increase (or 0.025 standard deviations) in *Bloat*.

Jointly, the above variables explain only about 16.6% of the variation in the MD&A *Bloat*, which indicates considerable firm-level residual variation in disclosure bloat. In column (2), we replace industry fixed effects with firm fixed effects. While the coefficient estimates are generally similar regardless of the fixed effect structure, the adjusted R-squared increases to 44.5%.

For the conference calls analysis, tabulated in columns (3) and (4), the results are generally similar except that we find a positive and statistically significant coefficient on *Earn_Vol*, which is likely to reflect disclosure complexity. Compared to the MD&A sample, the determinants of conference calls' *Bloat* jointly explain a somewhat lower portion of its variation (only about 11.7% without and 22.2% with firm fixed effects).

Overall, we find preliminary evidence that *Bloat* is associated with the financial circumstances of a firm in intuitive ways, which helps to establish its validity. We also find that managers are more likely to release bloated disclosures when their firm performs worse, which is consistent with the managerial obfuscation hypothesis.

C   Capital Market Consequences

Rich cross-sectional and over-time variation in corporate disclosure bloat, in conjunction with incentives for obfuscation of actual performance, makes it interesting to study the effect of *Bloat* on capital market outcomes. In theory, low disclosure quality is associated with lower liquidity and higher cost of capital (e.g. Leuz and Verrecchia, 2000; Lambert et al., 2007). Specifically to our measure, the presence of redundant and irrelevant details is likely to slow down price discovery and introduce a layer of informational asymmetries among investors trading the stock.

Test Design. We use three proxies that jointly capture the degree of price informativeness and information asymmetry. First, we use intraperiod timeliness (*IPT*), following Butler et al. (2007), to measure the speed of price discovery. *IPT* is calculated over a five-

day window relative to filing or conference call dates based on the following formula:

$$IPT_{[0,5]} = \sum_{i=0}^{4} \left(\frac{CAR_{[0,i]}}{CAR_{[0,5]}} \right) + \frac{1}{2} \tag{8}$$

where $CAR_{[0,i]}$ denotes cumulative abnormal returns from day 0 to day i. We use market-adjusted abnormal returns in the calculation of $CAR_{[0,i]}$. Intuitively, a higher *IPT* indicates faster price discovery after a release of certain information.

Second, we calculate the probability of informed trade (*PIN*) (Easley et al., 1996) by following the algorithm suggested in Brown and Hillegeist (2007).[25]

As our third proxy, we use daily bid-ask spread measured on the announcement day following Corwin and Schultz (2012).

To test for informational frictions created by bloated reporting, we estimate the following ordinary least squares regression:

$$Info\ Friction_{it} = \beta Bloat_{it} + \gamma \mathbf{X}_{it} + \delta_t + \theta + \varepsilon_{it} \tag{9}$$

where *Info Friction*$_{it}$ is either $IPT_{[0,5]it}$, PIN_{it}, or *Spread*$_{it}$, $\mathbf{X}_{it}$ is a set of firm-level control variables, and δ_t stands for time fixed effects and θ stands for either firm or industry fixed effects. We use the same set of control variables as in Section IV.B, except that we also include the absolute value of earnings surprise (*abs_SUE* to control for the news component of the announcement. As before, the standard errors are clustered by industry and all continuous variables are winsorized at 1% and 99% levels.

Results. Table 7 presents the results. Panel A focuses on the MD&A sample. In line with our expectations, we find that *Bloat* exhibits a negative association with $IPT_{[0,5]}$ and shows positive associations with the probability of informed trade and bid-ask spread. In terms of economic magnitudes, a one standard deviation increase in *Bloat* is associated with a 0.16% increase in the probability of informed trading, an 8.8% point decrease in the speed of price discovery, and a 17.6% point increase in bid-ask spread (all values computed using coefficient values in columns (1), (3), and (5)). We use industry and year fixed effects in columns (1), (3), and (5), but our results are similar when we replace industry fixed effects with firm fixed effects in columns (2), (4), and (6).

In Panel B, we repeat the same analysis for the conference calls sample. The infer-

---

[25]We follow Lee and Ready (1991)'s algorithm to infer the directions of daily trades from the Trade and Quote (TAQ) database. The calculation is performed as follows: $PIN = \frac{\mu\alpha}{\mu\alpha + 2\varepsilon}$, where α is the probability of an information event, μ is the trading intensity informed traders, and ε is the trading intensity of uninformed traders.

ences are qualitatively similar. A one standard deviation increase in *Bloat* is associated with a 0.17% increase in the probability of informed trading, a 5.0% point decrease in the speed of price discovery, and a 45.5% point increase in bid-ask spreads (computed using columns (1), (3), and (5)).

Taken together, our results are strongly in line with the prediction that disclosure bloat hinders effective information transfer between companies and information users.

VI Analysis of Theme-Specific Summaries

Our final analysis probes another relevant application of GPT by exploiting query-based summaries. Query gives a specific set of instructions to a machine when performing a summary. GPT is known to successfully perform query-based summaries without further fine-tuning or training (Yang et al., 2023). We instruct GPT to prepare theme-specific summaries related to financial and non-financial performance and examine their informativeness.

A Obtaining Theme-Specific Summaries

We use the unconstrained summaries performed by GPT as a starting point and further instruct the model to extract information about (1) financial performance and (2) ESG activities. For this analysis, we focus on conference call transcripts because ESG discussions are not as common within MD&A. In contrast, conference call transcripts have a broader scope and often feature discussions of environmental and social issues (Hassan et al., 2019; Sautner et al., 2023). We provide our prompts in Appendix B. We do not provide any additional explanations about what the definition of financial performance or ESG activities is when writing the prompts. Instead, we rely on the notion that GPT "understands" these concepts.[26]

Appendix C provides several snippets from ESG- and financial-performance-related summaries. In the case of ESG, we observe that firms discuss greenhouse gas emissions, environmental sustainability, and renewable energy, etc. In the case of financial performance, the discussion is mainly related to operations, earnings, cash flows, and fluctuations in revenue, etc. Based on these examples, as expected, theme-specific sum-

---

[26]As a validation exercise, we ask GPT to compose an artificial earnings conference call transcript that features the discussion of ESG (or financial performance). We include the full text of this "constructed" transcript in Appendix B. As can be seen from the example, the model understands the concept of ESG and financial performance without any further directions to narrow down its answers. Additionally, we make sure that the model does not make up or refer to other topics of information when it fails to find ESG- or financial-performance-related information from a given text.

maries appear to perform reasonably well.

B   Informativeness of Theme-Specific Summaries

Time Trend.

In Table 8, Panel A, provides summary statistics illustrating the information content of the financial and ESG-related summaries. %ESG indicates the percentage of non-empty ESG-related summaries (an empty summary is generated when GPT determines there is no relevant information to summarize), whereas %Fin indicates the percentage of non-empty financial-performance-related summaries. *len*ESG and *len*Fin are the length of ESG-related and financial-performance-related summaries, respectively, scaled by the length of the original summarized document. As expected, we find that summaries of financial performance are almost always non-empty. At the same time, the frequency of non-empty ESG-related summaries varies between 20% and 45% over the years. Only 22.7% of the original summaries contained ESG-related information in 2009. However, this number gradually increased to 43.7% in 2020, consistent with the heightened attention in ESG issues. Similarly, the length of ESG-related summaries has increased gradually from 2.2% of the original summary in 2009 to 4.3% of the original summary in 2020. In contrast, the length of financial performance-related summaries does not exhibit a systematic trend.

We visualize these results in Figure 5(a). The left-hand-side depicts the time trend of %ESG and *len*ESG and the right-hand-side depicts %Fin and *len*Fin. One can observe a positive time trend in ESG-related summary characteristics, while this is not the case for financial summaries.

Overall, these results are consistent with prior evidence of an increasing time trend climate risk exposure based on conference call transcripts (Sautner et al., 2023) and help us to validate prompt-based summaries further.

Informativeness. To examine the incremental informativeness of theme-specific summaries to investors, we use them to generate ESG-related ($Sentiment^{ESG}$) and financial-performance-related sentiment ($Sentiment^{Fin}$), in line with the definition in Section IV.A. Then we estimate the following ordinary least squares estimation by year:

$$CAR_{[0,1]it} = \beta_1 Sentiment_{it}^{ESG} + \beta_2 Sentiment_{it}^{Fin} + \gamma \mathbf{X}_{it} + \delta_t + \theta_j + \varepsilon_{it} \qquad (10)$$

where $CAR_{[0,1]it}$ is the cumulative abnormal return on the conference call date, $\mathbf{X}_{it}$ is the same set of firm-level controls, δ_t stands for quarter fixed effects and θ_j stands for

industry fixed effects. Standard errors are clustered by industry and all continuous variables at winsorized at 1% and 99%.

Table 8, Panel B reports the results. Columns (1) and (2) report annual coefficient estimates and *t*-statistics for *Sentiment*^{ESG}. In 2009, we find a negative and insignificant coefficient. However, the coefficient flips its sign in 2010 and steadily increases in magnitude over time. Ultimately, in 2020, we observe a coefficient of 0.076 with a *t*-value of 2.27. At the bottom of the panel, we also report the pooled (full sample) estimation. We observe a positive coefficient of 0.048 (*t*-value = 1.88) on *Sentiment*^{ESG}. The last row shows that the observed time trend is statistically significant, indicating a steady increase in the importance of ESG-related information from 2010 to 2020.[27]

Columns (3) and (4) report the estimates of β_2 by and their corresponding *t*-statistics. Unlike *Sentiment*^{ESG}, we observe a positive and statistically significant coefficient on *Sentiment*^{Fin} starting from 2009. Although there are some fluctuations in coefficient magnitudes and *t*-values over time, we see consistent statistical significance throughout our sample period. In the full sample analysis, we also find a positive and statistically significant coefficient (0.063) on *Sentiment*^{Fin} (*t*-value = 2.10). This *t*-statistics is slightly higher than that of *Sentiment*^{ESG}. However, we obtain a positive yet statistically insignificant time trend (0.031 with a *t*-value of 1.63).

Figure 5(b) visualizes these findings. The left-hand side of the figure shows a time trend in *t*-statistics over time. The solid line, which represents yearly *t*-values of *Sentiment*^{ESG}, shows a strong increasing time trend. In contrast, the dotted line that represents yearly *t*-values of *Sentiment*^{Fin}, does not show a notable trend. On the right-hand-side, we observe that the coefficient values of both *Sentiment*^{ESG} and *Sentiment*^{Fin} have an increasing time trend.

VII Conclusion

We probe the economic usefulness of large language models using financial markets as a laboratory. By summarizing a large sample of corporate disclosures with GPT-3.5-Turbo, we show that the length of the summaries is shortened by as much as 80%, on average. Importantly, the obtained summaries appear to provide more relevant insights as compared to the underlying raw documents. Specifically, we show that summarized sentiment better explains cumulative abnormal returns around disclosure dates than raw sentiment. Building on this insight, we construct a novel and easy-to-implement measure

---

[27]We estimate the following model: $tvalue_t = \gamma_0 + \gamma_1 Year_t + \epsilon_t$ and report $\hat{\gamma}_1$ accompanied by its robust *t*-statistics. $\hat{\gamma}_1$ (0.159) is highly significant (*t*-value = 2.80).

of the degree of "bloat" in textual disclosures. Disclosure bloat exhibits rich heterogeneity across firms and over time and varies intuitively with its economic determinants. We show that bloated disclosures are associated with slower price discovery and higher information asymmetry, thus implying negative capital market consequences. Finally, we show that GPT is useful to investors interested in targeted summaries related to important topics, such as a summary of ESG-related activities.

Our results indicate that investors can utilize generative AI systems to cut through the clutter of corporate disclosures. Over the past decades, corporate disclosures have been increasing in length and complexity and investors often do not have the capacity to fully process disclosed information (e.g. Blankespoor et al., 2020; Cohen et al., 2020). Large language models show promise in dealing with this information overload. Summaries generated by GPT are significantly shorter while they retain and amplify the main message. Such AI tools should be beneficial for investors in making more informed investment decisions. Although investors can seek out tools like ChatGPT themselves, regulators or information intermediaries could build the necessary infrastructure to readily provide such summaries in a timely fashion.

Finally, by relying on recent advancements in generative AI, we develop a simple and intuitive strategy to measure the degree to which textual information contains redundancies and irrelevant or excessive details. Due to its straightforward nature, our methodology can be easily implemented for any other type of corporate information or context (e.g., press releases, job postings, and websites). Additionally, as textual data has become increasingly relevant in many disciplines (e.g. Gentzkow et al., 2019), our strategy can be easily adapted to other non-corporate settings as well (e.g., news articles).

References

Balakrishnan, K., Billings, M.B., Kelly, B., Ljungqvist, A., 2014. Shaping liquidity: On the causal effects of voluntary disclosure. The Journal of Finance 69, 2237–2278.

Bernard, D., Blankespoor, E., de Kok, T., Toynbee, S., 2023. Confused readers: A modular measure of business complexity. University of Washington Working Paper .

Bhaskar, A., Fabbri, A.R., Durrett, G., 2022. Zero-shot opinion summarization with GPT-3. arXiv preprint arXiv:2211.15914 .

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, 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, 329–357.

Brown, S., Hillegeist, S.A., 2007. How disclosure quality affects the level of information asymmetry. Review of Accounting Studies 12, 443–477.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Advances in Neural Information Processing Systems 33, 1877–1901.

Butler, M., Kraft, A., Weiss, I.S., 2007. The effect of reporting frequency on the timeliness of earnings: The cases of voluntary and mandatory interim reports. Journal of Accounting and Economics 43, 181–217.

Cardinaels, E., Hollander, S., White, B.J., 2019. Automatic summarization of earnings releases: attributes and effects on investors' judgments. Review of Accounting Studies 24, 860–890.

Chang, Y.C., Hsiao, P.J., Ljunqvist, A., Tseng, K., 2022. Testing disagreement models. The Journal of Finance 77, 2239–2285.

Cohen, L., Malloy, C., Nguyen, Q., 2020. Lazy prices. The Journal of Finance 75, 1371–1415.

Corwin, S.A., Schultz, P., 2012. A simple way to estimate bid-ask spreads from daily high and low prices. The Journal of Finance 67, 719–760.

Doherty, K., Marques, F., 2023. Citadel negotiating enterprise-wide ChatGPT license, Griffin says. Bloomberg .

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, 221–245.

Easley, D., Kiefer, N.M., O'hara, M., Paperman, J.B., 1996. Liquidity, information, and

infrequently traded stocks. The Journal of Finance 51, 1405–1436.

Florackis, C., Louca, C., Michaely, R., Weber, M., 2023. Cybersecurity risk. The Review of Financial Studies 36, 351–407.

Frankel, R., Jennings, J., Lee, J., 2022. Disclosure sentiment: Machine learning vs. dictionary methods. Management Science 68, 5514–5532.

Gao, M., Huang, J., 2019. Informing the Market: The Effect of Modern Information Technologies on Information Production. The Review of Financial Studies 33, 1367–1411.

Gentzkow, M., Kelly, B., Taddy, M., 2019. Text as data. Journal of Economic Literature 57, 535–74.

Giglio, S., Kelly, B., Stroebel, J., 2021. Climate finance. Annual Review of Financial Economics 13, 15–36.

Goldstein, I., Spatt, C.S., Ye, M., 2021. Big Data in Finance. The Review of Financial Studies 34, 3213–3225.

Goyal, T., Li, J.J., Durrett, G., 2022. News summarization and evaluation in the era of GPT-3. arXiv preprint arXiv:2209.12356 .

Gunning, R., et al., 1952. Technique of clear writing .

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

Huang, A.H., Wang, H., Yang, Y., 2023. Finbert: A large language model for extracting information from financial text*. Contemporary Accounting Research Forthcoming.

Kim, A.G., Nikolaev, V.V., 2022. Context-based interpretation of financial information. University of Chicago Working Paper .

Lambert, R., Leuz, C., Verrecchia, R.E., 2007. Accounting information, disclosure, and the cost of capital. Journal of accounting research 45, 385–420.

Lee, C.M., Ready, M.J., 1991. Inferring trade direction from intraday data. The Journal of Finance 46, 733–746.

Leuz, C., Verrecchia, R.E., 2000. The economic consequences of increased disclosure. Journal of Accounting Research , 91–124.

Li, F., 2008. Annual report readability, current earnings, and earnings persistence. Journal of Accounting and Economics 45, 221–247.

Li, K., Mai, F., Shen, R., Yan, X., 2021. Measuring corporate culture using machine learning. The Review of Financial Studies 34, 3265–3315.

Loughran, T., McDonald, B., 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-Ks. The Journal of Finance 66, 35–65.

Loughran, T., McDonald, B., 2014a. Measuring readability in financial disclosures. The

Journal of Finance 69, 1643–1671.

Loughran, T., McDonald, B., 2014b. Regulation and financial disclosure: The impact of plain english. Journal of Regulatory Economics 45, 94–113.

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

Maćkowiak, B., Matějka, F., Wiederholt, M., 2023. Rational inattention: A review. Journal of Economic Literature 61, 226–73.

NIRI, 2014. Earnings call practices report .

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al., 2018. Improving language understanding by generative pre-training .

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al., 2019. Language models are unsupervised multitask learners. OpenAI blog 1, 9.

Ramshaw, L.A., Marcus, M.P., 1999. Text chunking using transformation-based learning. Natural Language Processing Using Very Large Corpora , 157–176.

Sautner, Z., van Lent, L., Vilkov, G., Zhang, R., 2023. Firm-level climate change exposure. Journal of Finance Forthcoming.

SEC, 2013. Report on review of disclosure requirements in regulation s-k. http://www.sec.gov/news/studies/2013/reg-sk-disclosure-requirements-review.pdf , 5514–5532.

Sims, C.A., 2003. Implications of rational inattention. Journal of Monetary Economics 50, 665–690.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in Neural Information Processing Systems 30.

Yang, X., Li, Y., Zhang, X., Chen, H., Cheng, W., 2023. Exploring the limits of ChatGPT for query or aspect-based text summarization. arXiv preprint arXiv:2302.08081 .

Zhang, J., Zhao, Y., Saleh, M., Liu, P., 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization, in: International Conference on Machine Learning, PMLR. pp. 11328–11339.

Appendix A. Variable Descriptions

Name	Description
Bloat	The difference between the length of original document and summarized document, scaled by the length of the original document.
Length	The number of words contained in a document.
Sentiment^{Raw}	Textual sentiment of a raw document calculated using financial keyword dictionaries by Loughran and McDonald (2011).
Sentiment^{Sum}	Textual sentiment of a summarized document calculated using financial keyword dictionaries by Loughran and McDonald (2011).
Fog	Fog index based on Gunning et al. (1952) with higher values indicating less readable text.
Plain_Eng	Plain English measure based on Loughran and McDonald (2014b) with higher values indicating less readable text.
$CAR_{[0,1]}$	Market adjusted cumulative abnormal returns from the filing date to one day after the filing date.
$IPT_{[0,5]}$	5-day interperiod timeliness calculated following Butler et al. (2007).
PIN	Probability of informed trades calculated following Easley et al. (1996). Buy and sell directions are imputed using Lee and Ready (1991)'s algorithm. We use quarterly *PIN* for the conference calls sample and annual *PIN* for the MD&A sample.
Spread	One-day bid-ask spread.
log(*ME*)	The natural logarithm of the market capitalization.
log(*BE/ME*)	The natural logarithm of book-to-market ratio.
N_Analyst	The number of analysts following a firm calculated as the number of distinct analyst reports.
Inst_Own	Institutional ownership calculated from 13-F filings.
Report_Lag	The number of days between fiscal period end and earnings announcement date.
Earn_Vol	Earnings volatility calculated as the standard deviation of five past earnings scaled by total assets. For quarterly volatility, we use earnings of the same quarter in last five years.
Loss	An indicator that equals one when a firm reports negative earnings.
One_Day_Ret	Raw buy-and-hold return on the filing date.
Price	The end price of a stock on the filing date.
SUE	Earnings surprise calculated as the difference between the reported earnings-per-share and analyst consensus, scaled by the price at fiscal period end date.
abs_SUE	Absolute value of earnings surprise.
Friday	An indicator that equals one when the reporting date is Friday.
Industry Classification	Two-digit standard industry classification (SIC) codes

Appendix B. Theme-Specific Summaries

For financial and ESG summaries, we include additional prompts to prevent the model from including other topics of information when it fails to identify financial or ESG-related information from a given transcript. Without further directions, the model likely generates some "made-up" answers when there is insufficient information (e.g., "since there is no information about ESG in the given text, we summarize the financial performance of this firm"). Explicitly instructing the model to print out NA when there is no information available and giving it a related example (one-shot training) significantly improves the accuracy of theme-specific summaries.

One might still be concerned that the model does not understand the concept of ESG or financial performance absent specific examples. However, we find that providing specific examples to the model adversely harms the performance by confining the summary only to information that is similar to the provided examples. We further validate that GPT understands the used concepts by prompting it to write a paragraph about ESG and financial performance in a hypothetical conference call. Our prompt was "Imagine that you are a CEO of a company. You are holding an earnings conference call. Write a paragraph of conference call transcript that talks about ESG (financial) performance."

Topic	Generated Script
ESG	Ladies and gentlemen, thank you for joining us on today's earnings conference call. I am pleased to share that our company has made significant progress in our Environmental, Social, and Governance (ESG) performance this quarter. Through our investments in sustainable technologies and resource management, we have managed to reduce our carbon footprint by 18% year-on-year, surpassing our initial target of 12%. Our commitment to diversity and inclusion has resulted in a 25% increase in underrepresented groups within our workforce, and we have implemented comprehensive training programs to foster a culture of respect and equal opportunity. In terms of governance, we have strengthened our board's oversight with the addition of two independent directors, who bring with them deep expertise in sustainable business practices. Our focus on ESG not only underscores our commitment to responsible corporate citizenship, but also strategically positions our company for long-term growth and profitability in a rapidly evolving global market.
Financial	Ladies and gentlemen, thank you for joining us on today's earnings conference call. I am delighted to share that our company has delivered a robust financial performance this quarter, driven by strong operational execution and a continued focus on innovation. Our revenues have grown by 15% year-on-year, reaching $4.2 billion, with a 20% increase in net income at $850 million. This outstanding performance reflects the success of our strategic investments in high-growth markets and the expansion of our product and service offerings. Furthermore, our adjusted EBITDA margin has improved by 250 basis points to 30%, primarily due to the ongoing optimization of our cost structure and efficiency gains across our operations. Our balance sheet remains healthy, with a net debt-to-EBITDA ratio of 1.5x, providing ample liquidity and financial flexibility to support our future growth initiatives. We are confident in our ability to continue delivering value to our shareholders, and as a testament to this confidence, we are raising our full-year guidance for both revenue and earnings per share.

Appendix C. Sample Summaries

Panel A. Excerpt from a sample summary (ESG-specific)
[...] In the Energy market, Woodward expects long-term growth in demand and is well-positioned to benefit from the shift to natural gas from coal. The company's net sales for the second quarter of 2012 were $469 million, with the energy segment's sales increasing by 14% to $244 million. [...]
[...] Capella has maintained a financial responsibility score of 3.0, the highest possible score for educational institutions approved to distribute federal financial aid. [...]
[...]The company is focused on environmental sustainability and accountability, and has shifted their portfolio towards financial and environmental sustainability. They are developing green aluminum contracts and believe there is potential for a premium for low-carbon producers. They have seen a reduction in value-add shipments in the second quarter, but are seeing some strength in the foundry markets going into the third quarter. [...]
[...]Unitil has decreased its fugitive emissions from natural gas distribution and is exploring opportunities to add renewable natural gas to its supply portfolio. The company is committed to creating a sustainable future through its employees and providing a safe and respectful workplace.[...]
[...] The company has a good relationship with Qatar Petroleum and Chevron Phillips Chemical, but each transaction stands on its own. Chevron does not flare gas and has flow assurance covered, with plans to increase export capacity. The company is also focused on reducing carbon emissions and has investments in wind and renewable natural gas to lower the carbon intensity of operations. [...]
Panel B. Excerpt from a sample summary (Financial-performance-specific)
[...] Revenue and adjusted earnings per share were above expectations in Q1 2014, with revenue growing 5% over the prior year quarter. Small Business Services saw growth of almost 9%, and Financial Services also saw over 2% revenue growth. The company also generated strong operating cash flow and increased their balance sheet cash position. The company's brand awareness campaign and cost reduction efforts contributed to their success. [...]
[...] The company expects revenue in the range of $2.50 billion to $2.80 billion in the third quarter, with earnings per share in the range of $0.29 to $0.39, including expected restructuring charges. The company is encouraged by its revenue moving back toward end demand and its improving profitability. The company's mission is to grow revenue and maintain a 25% operating expense level. [...]
[...] The company expects EPS to increase only 5% or 4% next year, despite substantial growth in EPS in the fourth quarter of last year and this year. It discusses the financial outlook and projections for QUALCOMM for fiscal year 2010, including a decline in revenue due to average revenue decline on chipsets and increased pricing pressure in the chipset business. [...]

Figure 1. Visual Illustration of Disclosure Bloat

This figure illustrates the concept of our bloat measure. The horizontal axis is the length of each document with n being the original length. The length of the summarized document is denoted as n^*. The vertical axis is hypothetical noise-to-information ratio.

Figure 1. Visual Illustration of Disclosure Bloat
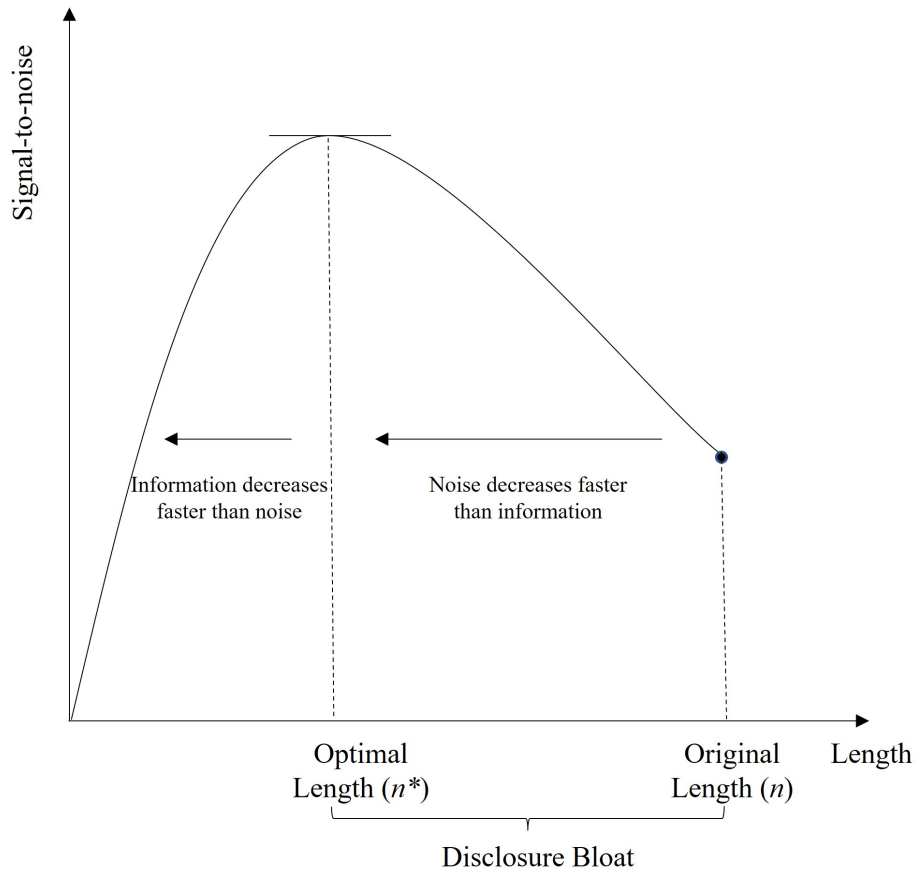
Figure 2. Changes in Average Length and Sentiment

This figure illustrates the changes in average length and sentiment before and after the summarization. AboveMed (BelowMed) refers to the observations that are above (below) the median value of sentiment. Figure 2(a) uses the MD&A sample and Figure 2(b) uses the conference call sample.

Figure 2(a). Changes in Average Length and Sentiment (MD&A sample)
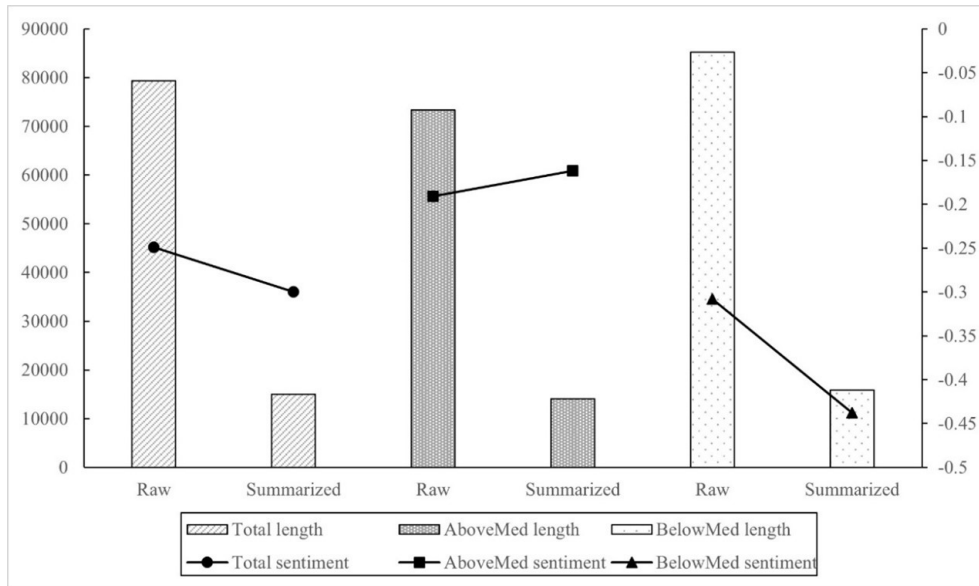
Figure 2(b). Changes in Average Length and Sentiment (conference call sample)
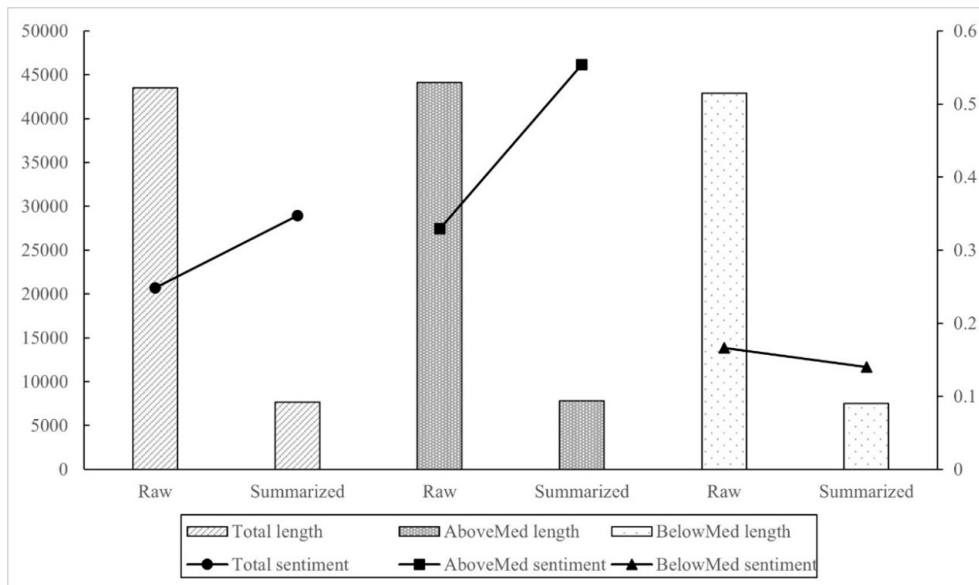
Figure 3. Abnormal Returns and Sentiment

This figure illustrates the averaged cumulative abnormal returns over two days (filing date and one day after) depending on the quintiles of sentiment. Sentiment is calculated using the Loughran and McDonald (2011) keyword dictionaries. Dotted line represents sentiment from the raw documents. Solid line represents sentiment from the summarized documents. Figure 3(a) uses the MD&A sample and Figure 3(b) uses the conference call sample.

Figure 3(a). Abnormal Returns and Sentiment Quintiles (MD&A sample)
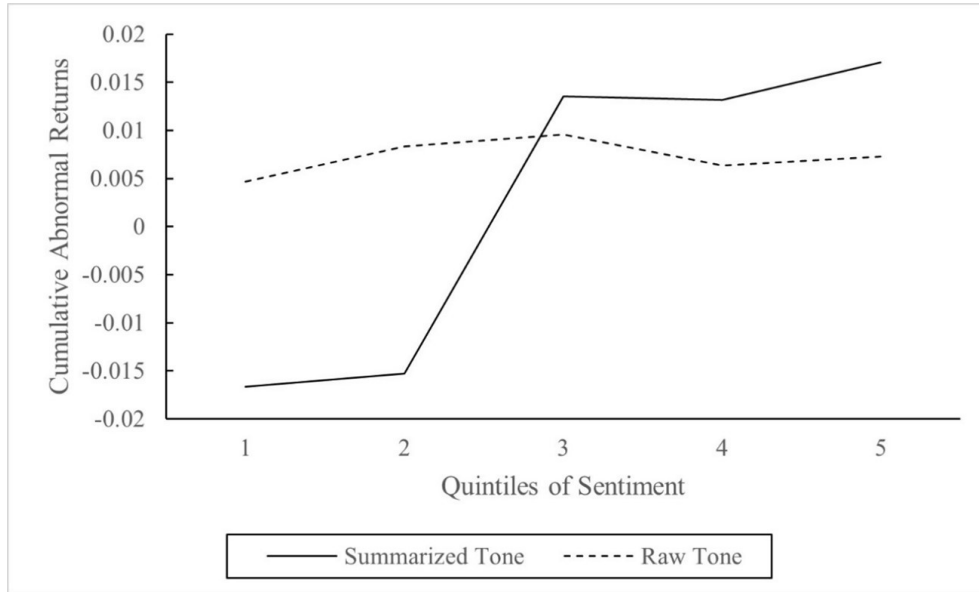
Figure 3(b). Abnormal Returns and Sentiment Quintiles (conference call sample)
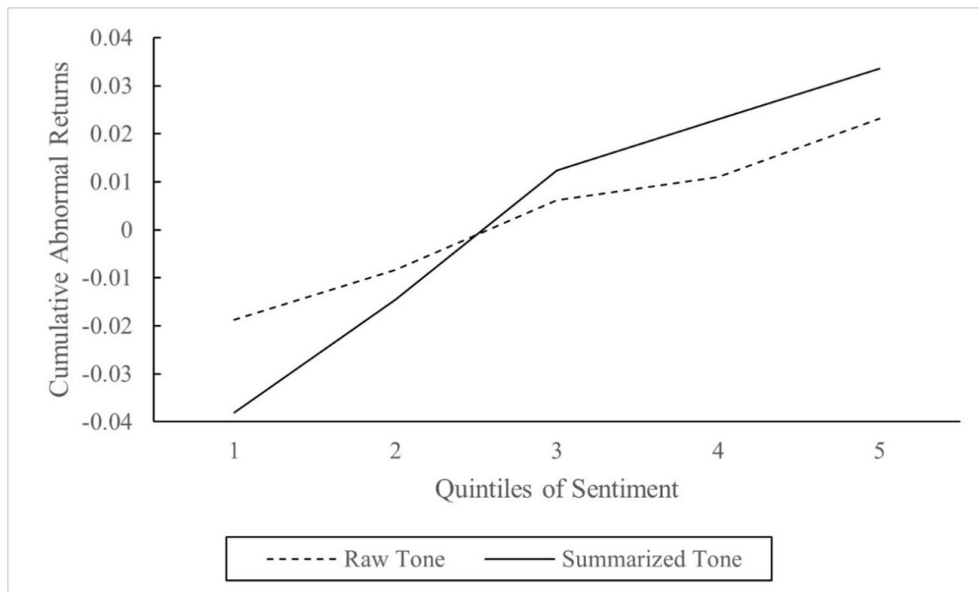
Figure 4. Average Length and Disclosure Bloat

This figure illustrates the average length and average *Bloat* of each document on an annual basis. Length is the number of words included in each document. Figure 4(a) uses the MD&A sample and Figure 4(b) relies on the conference call sample.

Figure 4(a). Average Length and Disclosure Bloat (MD&A sample)
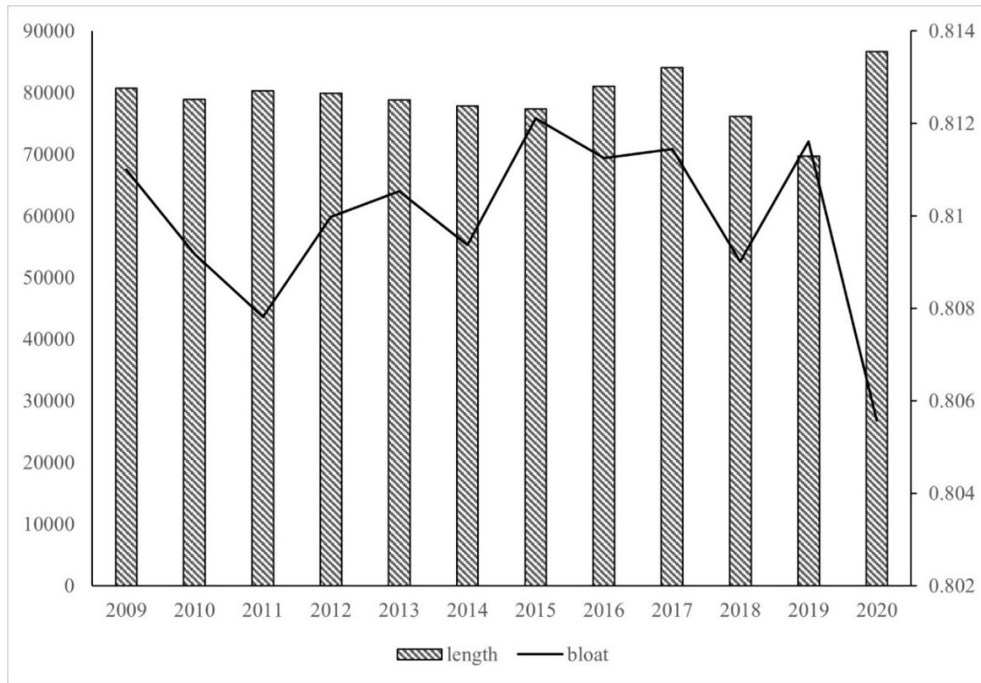
Figure 4(b). Average Length and Disclosure Bloat (conference call sample)
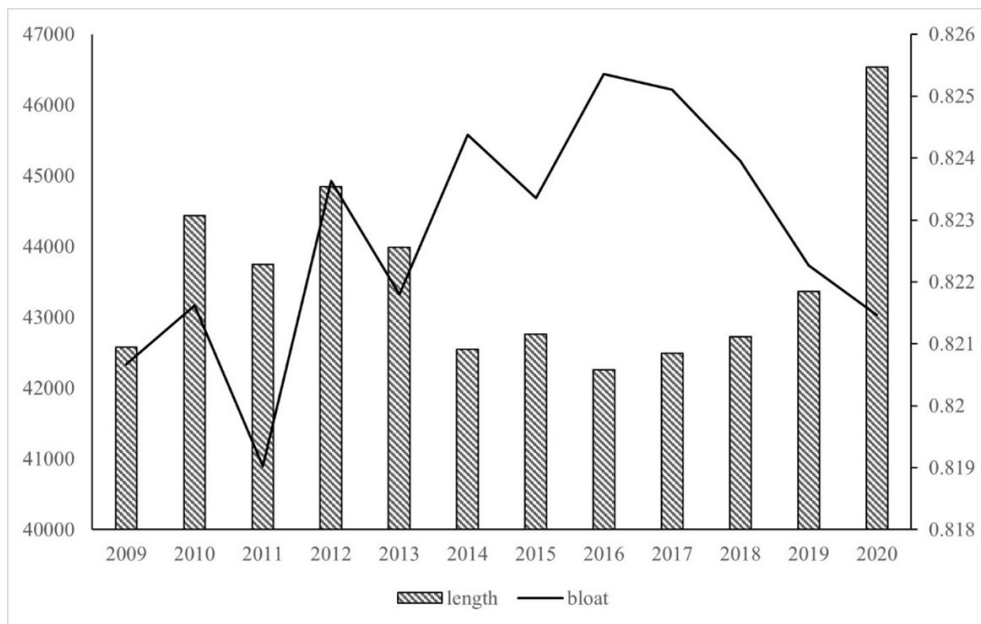
Figure 5. Theme-Specific Summaries

This figure illustrates the time-series trend of theme-specific summaries extracted from conference call summaries and their marked-based informativeness. Figure 5(a) shows the time trend of ESG-related summaries (left) and financial-performance-related summaries (right). %ESG (%Fin) denotes the percentage of summaries that contain ESG-related (financial-performance-related) contents. *len*ESG (*len*Fin) denotes the length of ESG-specific (financial-specific) summaries scaled by the length of the original summaries. Figure 5(b) shows the trend in the absolute values of *t*-statistics (left) and coefficients (right) when we regress abnormal returns on sentiments of ESG-specific summaries ($Sentiment^{ESG}$) and financial-performance-specific summaries ($Sentiment^{Fin}$).

Figure 5(a). Time Trend (left: ESG, right: financial performance)
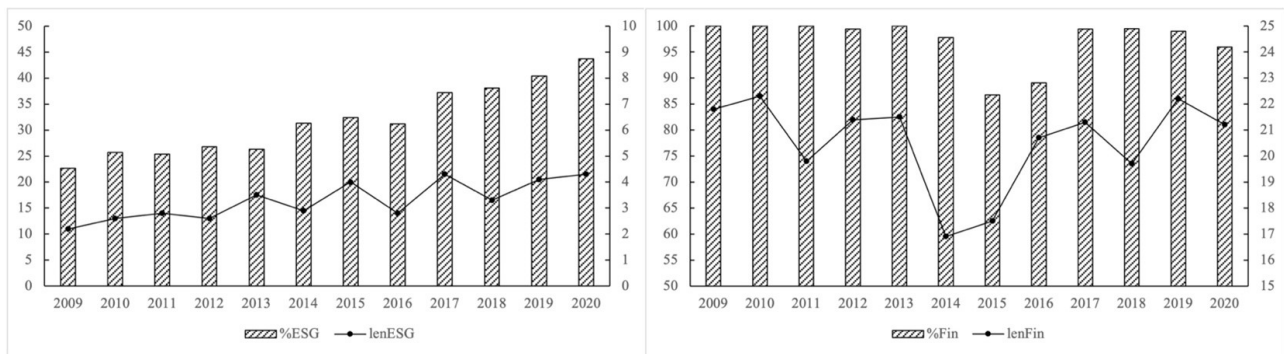
Figure 5(b). Time Trend of Informativeness (left: *t*-values, right: coefficient magnitudes)
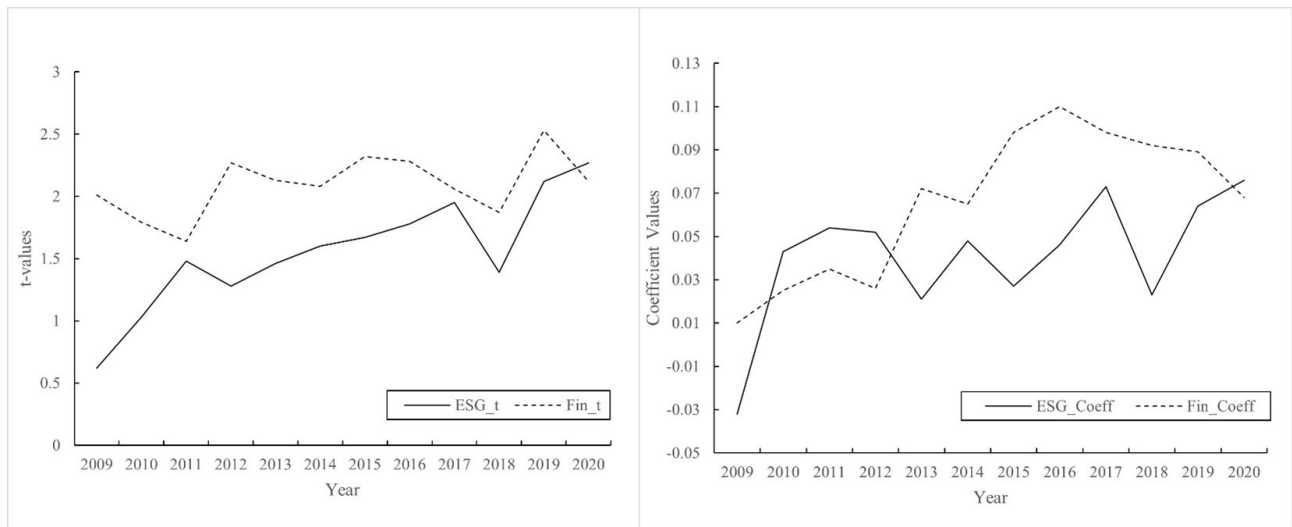
Table 1. Descriptive Statistics

This table reports the descriptive statistics of the entire and the randomly selected sample. We report the natural logarithm of market capitalization, the natural logarithm of book-to-market ratio, the number of analysts following, and institutional ownership. In the last column, we report the difference of mean values between the universe and the random sample. In Panel A, we report descriptive statistics of MD&A samples and in Panel B, we report descriptive statistics of conference call samples. Refer to Appendix A for detailed variable descriptions. Standard two-sided t-tests were performed to calculate the statistical significance of the differences. t-values are reported in parentheses.

Panel A. MD&A

	Universe				Sample				Diff.
	N	Mean	Median	Std	N	Mean	Median	Std	(2) − (6)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(2) − (6)
log(ME)	8,699	14.58	14.52	1.69	1,790	14.62	14.55	1.72	-0.04
									(-0.27)
log(BE/ME)	8,699	-7.95	-7.87	0.88	1,790	-7.97	-7.92	0.89	0.02
									(-0.24)
$N\_Analyst$	8,699	5.17	3.00	5.28	1,790	4.91	3.00	5.34	0.26
									(1.01)
$Inst\_Own$	8,699	0.62	0.76	0.37	1,790	0.61	0.72	0.37	0.01
									(0.01)

Panel B. Conference Calls

	Universe				Sample				Diff.
	N	Mean	Median	Std	N	Mean	Median	Std	(2) − (6)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(2) − (6)
log(ME)	42,238	14.68	14.64	1.78	8,907	14.72	14.67	1.85	-0.04
									(-1.10)
log(BE/ME)	42,238	-7.95	-7.85	0.89	8,907	-7.99	-7.90	0.93	0.04
									(1.03)
$N\_Analyst$	42,238	7.49	6.00	6.36	8,907	7.60	6.00	6.44	-0.11
									(-0.84)
$Inst\_Own$	42,238	0.68	0.70	0.36	8,907	0.71	0.73	0.36	-0.03
									(-0.31)

Table 2. Linguistic Attributes of Summaries

This table reports the linguistic attributes of raw documents and summarized documents. Length is the number of words in the document. Sentiment is the tone of the document calculated using Loughran and McDonald (2011) keyword dictionaries. Fog is based on Gunning et al. (1952) and Plain_Eng is defined following Loughran and McDonald (2014b). In the last column we report the difference between the mean values of the two groups with its statistical significance in parentheses. In Panel A, we use the randomly chosen MD&A samples. In Panel B, we tabulate the change in each variable. In Panels C and D, we repeat the analyses for conference call transcripts. Refer to Appendix A for detailed variable descriptions. ***, **, and * denote statistical significance at 1%, 5%, and 10% level, respectively.

Panel A. Random Sample of MD&As

		Raw Document			Summarized Document			
	N	Mean	Median	Std	Mean	Median	Std	Diff.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(5) − (2)
Length	1,790	79,310	71,264	41,791	14,974	13,543	7,693	-64,336***
Sentiment	1,790	-0.249	-0.261	0.203	-0.300	-0.316	0.265	-0.051***
Fog	1,790	10.026	9.870	1.563	10.504	10.490	1.071	0.478***
Plain_Eng	1,790	-0.303	-0.311	1.152	-0.214	-0.229	0.853	0.089***

Panel B. Change in Variables (MD&A Sample)

				Percentiles		
	N	Mean	Std	p25	p50	p75
ΔLength	1,790	-64,336	5,836	-70,215	-65,500	-58,203
ΔSentiment	1,790	-0.051	0.065	-0.126	-0.056	-0.030
ΔFog	1,790	0.478	1.133	-1.521	0.367	2.532
ΔPlain_Eng	1,790	0.089	0.050	0.038	0.093	0.155

Panel C. Random Sample of Conference Calls

		Raw Document			Summarized Document			
	N	Mean	Median	Std	Mean	Median	Std	Diff.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(5) − (2)
Length	8,907	43,511	44,488	13,495	7,661	7,804	2,606	-35,850***
Sentiment	8,907	0.248	0.264	0.202	0.347	0.381	0.298	0.099***
Fog	8,907	8.982	8.940	0.803	11.398	11.390	0.943	2.416***
Plain_Eng	8,907	-0.863	-0.853	1.033	-0.708	-0.711	0.951	0.155***

Panel D. Change in Variables (Conference Call Sample)

				Percentiles		
	N	Mean	Std	p25	p50	p75
ΔLength	8,907	-35,850	11,263	-38,825	-35,056	-28,023
ΔSentiment	8,907	0.099	0.185	-0.028	0.105	0.207
ΔFog	8,907	2.416	0.867	0.658	2.442	3.850
ΔPlain_Eng	8,907	0.155	0.107	0.013	0.167	0.315

Table 3. Partitions Based on Raw Sentiment

This table reports the linguistic attributes of raw documents and summarized documents based on raw sentiment. Specifically, we split the sample into two groups depending on the median value of raw sentiment. In Panel A, we use the MD&A sample and in Panel B, we use the conference call sample. Refer to Appendix A for detailed variable descriptions. ***, **, and * denote statistical significance at 1%, 5%, and 10% level, respectively.

Panel A. MD&A Sample

A1. When $Sentiment^{Raw} >$ Median

| | N | Raw Document | | | Summarized Document | | | Diff. |
| | | Mean | Median | Std | Mean | Median | Std | |
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(5) − (2)
Length	896	73,388	69,214	32,189	14,078	13,305	6,182	-59,310***
Sentiment	896	-0.191	-0.129	0.144	-0.162	-0.155	0.231	0.029***
Fog	896	8.193	10.030	1.595	10.610	11.610	1.085	2.417***
Plain_Eng	896	-0.359	-0.366	0.732	-0.231	-0.225	0.864	0.128***

A2. When $Sentiment^{Raw} <$ Median

| | N | Raw Document | | | Summarized Document | | | Diff. |
| | | Mean | Median | Std | Mean | Median | Std | |
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(5) − (2)
Length	894	85,239	73,935	48,879	15,871	13,889	8,866	-69,368***
Sentiment	894	-0.308	-0.389	0.105	-0.438	-0.470	0.175	-0.130***
Fog	894	11.860	9.690	1.514	10.397	11.370	1.046	-1.463***
Plain_Eng	894	-0.246	-0.239	0.697	-0.197	-0.183	0.866	0.049*

Panel B. Conference Call Sample

B1. When $Sentiment^{Raw} >$ Median

| | N | Raw Document | | | Summarized Document | | | Diff. |
| | | Mean | Median | Std | Mean | Median | Std | |
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(5) − (2)
Length	4,456	44,134	45,615	13,576	7,805	8,069	2,431	-36,329***
Sentiment	4,456	0.329	0.393	0.102	0.554	0.538	0.196	0.225***
Fog	4,456	7.062	8.040	0.504	11.445	11.440	0.952	4.303***
Plain_Eng	4,456	-0.995	-1.034	0.834	-0.693	-0.702	0.936	0.302***

B2. When $Sentiment^{Raw} <$ Median

| | N | Raw Document | | | Summarized Document | | | Diff. |
| | | Mean | Median | Std | Mean | Median | Std | |
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(5) − (2)
Length	4,451	42,887	43,249	13,389	7,517	7,658	2,438	-35,370***
Sentiment	4,451	0.166	0.165	0.136	0.140	0.143	0.189	-0.026***
Fog	4,451	10.902	10.860	0.581	11.351	11.340	0.933	0.449***
Plain_Eng	4,451	-0.731	-0.749	0.765	-0.723	-0.737	0.940	0.008

Table 4. Informativeness of Summarized Documents

This table reports the association between the textual sentiment and two-day cumulative abnormal returns. In columns (1) - (4), we use the sentiment calculated from raw documents. In columns (5) - (8), we use sentiment calculated from summarized documents. In columns (3) and (7), we focus on samples with positive sentiment. In columns (4) and (8), we focus on samples with negative sentiment. As control variables, we include log(*ME*), log(*BE/ME*), *Inst_Own*, and *SUE*. In Panel A, we use random samples chosen from the MD&A disclosures. In Panel B, we use random samples chosen from conference call transcripts. *t*-values are reported in parentheses and standard errors are clustered at the industry level. ***, **, and * denote statistical significance at 1%, 5%, and 10% level, respectively. Refer to Appendix A for detailed variable descriptions. Continuous variables are winsorized at 1% and 99%.

Panel A. MD&A Sample

Dependent Variable = $CAR_{[0,1]}$

	Raw Documents				Summarized Documents			
	Full		Pos	Neg	Full		Pos	Neg
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$Sentiment^{Raw}$	-0.007	-0.008	0.136*	-0.017				
	(-0.71)	(-1.05)	(2.00)	(-1.68)				
$Sentiment^{Sum}$					0.030***	0.027***	0.059*	0.035***
					(3.69)	(3.99)	(1.93)	(2.82)
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	No	No	Yes	No	No
Industry FE	Yes	No	Yes	Yes	Yes	No	Yes	Yes
Cluster	Ind	Ind	Ind	Ind	Ind	Ind	Ind	Ind
N	1,790	1,790	187	1,590	1,790	1,790	216	1,547
Adjusted R^2	0.031	0.035	0.195	0.075	0.045	0.054	0.207	0.082

Panel B. Conference Call Sample

Dependent Variable = $CAR_{[0,1]}$

	Raw Documents				Summarized Documents			
	Full		Pos	Neg	Full		Pos	Neg
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$Sentiment^{Raw}$	0.055***	0.065***	0.058***	0.046				
	(8.65)	(9.51)	(6.23)	(1.37)				
$Sentiment^{Sum}$					0.072***	0.077***	0.076***	0.084***
					(9.62)	(14.35)	(7.71)	(3.03)
Controls	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	No	No	Yes	No	No
Industry FE	Yes	No	Yes	Yes	Yes	No	Yes	Yes
Cluster	Ind	Ind	Ind	Ind	Ind	Ind	Ind	Ind
N	8,907	8,907	7,809	1,050	8,907	8,907	7,665	1,107
Adjusted R^2	0.037	0.042	0.086	0.167	0.094	0.129	0.111	0.188

Table 5. Variance Decomposition of Disclosure Bloat

This table provides a time-series variation and variance decomposition of our *Bloat* measure. We report descriptive statistics in Panel A. In Panel B, we first report incremental R^2 from adding different sets of fixed effects. In Panel C, the vertical axis represents the quintile of bloat in previous year and the horizontal axis corresponds to the quintile of bloat in the current year. Each cell denotes the transition probability from the past quintile to the current quintile. Panel D repeats the same exercise with the conference call sample.

Panel A. Descriptive Statistics

Sample from	Mean	Std	Percentiles		
			p25	p50	p75
MD&A	0.811	0.088	0.741	0.823	0.885
Conference Call	0.824	0.102	0.645	0.801	0.887

Panel B. Fixed Effect Structure

	MD&A	Conference Call
	(1)	(2)
B1. Incremental R^2		
Time FE	0.56%	1.32%
Industry FE	7.90%	3.93%
Time × Industry FE	26.34%	20.59%
Implied Firm Level	68.20%	74.16%
Sum	100.00%	100.00%
B2. Fraction of Variation		
Firm FE	32.33%	26.04%
Residual	67.67%	73.96%
Sum	100.00%	100.00%

Panel C. Time-Series Variation (MD&A)

$Q[Bloat_{it-1}]$	$Q[Bloat_{it}]$					
	Low	2	3	4	High	Total
Low	30.40%	25.28%	19.60%	13.07%	11.65%	100.00%
2	23.20%	20.17%	24.03%	20.17%	12.43%	100.00%
3	20.06%	23.12%	29.50%	20.89%	6.43%	100.00%
4	11.60%	17.13%	21.82%	26.52%	22.93%	100.00%
High	13.35%	15.19%	15.04%	19.67%	36.75%	100.00%

Panel D. Time-Series Variation (Conference Call)

$Q[Bloat_{it-1}]$	$Q[Bloat_{it}]$					
	Low	2	3	4	High	Total
Low	32.78%	24.20%	17.67%	13.69%	11.66%	100.00%
2	23.46%	23.30%	22.01%	17.26%	13.97%	100.00%
3	17.67%	19.51%	22.07%	21.68%	19.07%	100.00%
4	12.97%	18.05%	21.19%	23.20%	24.59%	100.00%
High	12.11%	15.46%	17.72%	24.60%	30.19%	100.00%

Table 6. Determinants of Disclosure Bloat

This table reports the determinants of our *Bloat* measure. In columns (1) and (2), we focus on the MD&A sample. In columns (3) and (4), we use the conference call sample. Standard errors are reported in parentheses and clustered at the industry level. ***, **, and * denote statistical significance at 1%, 5%, and 10% level, respectively. Refer to Appendix A for detailed variable descriptions. Continuous variables are winsorized at 1% and 99%.

Dependent Variable =		*Bloat*		
Sample =	MD&A		Conference Call	
	(1)	(2)	(3)	(4)
log(*ME*)	-0.071**	-0.063	-0.094***	-0.145
	(-2.32)	(-1.53)	(-2.54)	(-1.35)
log(*BE/ME*)	-0.007	-0.057	-0.009*	-0.017*
	(-1.41)	(-0.74)	(-1.96)	(-1.82)
N_Analyst	0.022**	0.037***	0.015**	0.026
	(2.52)	(3.31)	(2.63)	(1.38)
Inst_Own	-0.042	-0.236	-0.282*	-0.162
	(-0.36)	(-1.32)	(-1.74)	(-1.24)
Report_Lag	0.004**	0.001	0.007	0.003
	(2.32)	(0.88)	(1.36)	(0.73)
Earn_Vol	0.405	0.401	0.143**	0.226**
	(0.90)	(1.18)	(2.14)	(2.08)
Loss	0.105**	0.113**	0.141*	0.005
	(2.23)	(2.52)	(1.82)	(1.54)
One_Day_Ret	-0.022***	-0.022***	-0.018***	-0.018***
	(-3.62)	(-3.98)	(-4.73)	(-5.41)
Sentiment	-0.837***	-0.013**	-0.620***	-0.075**
	(-3.36)	(-2.29)	(-3.27)	(-2.39)
Fog	0.006	0.007	0.001	0.005
	(0.37)	(0.48)	(1.65)	(0.57)
Plain_Eng	0.037***	0.031***	0.020***	0.017***
	(7.64)	(4.37)	(8.19)	(5.80)
Log_Length	0.026***	0.028***	0.022***	0.025***
	(12.97)	(7.78)	(11.29)	(8.59)
Time FE	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	Yes
Industry FE	Yes	No	Yes	No
Cluster	Industry	Industry	Industry	Industry
N	1,790	1,787	8,907	8,875
Adjusted R^2	0.166	0.445	0.117	0.222

Table 7. Bloat in Narrative Disclosures

This table reports the association between our bloat measure and information asymmetry. We use three proxies for information asymmetry. We use intraperiod timeliness (*IPT*), probability of informed trade (*PIN*), and bid-ask spreads, respectively. As control variables, we include log(*ME*), log(*BE/ME*), *Inst_Own*, *Friday*, *N_Analysts*, *OneDayRet*, *Report_Lag*, *Price*, *Loss*, *abs_SUE*, *Fog*, *Plain_Eng* and *Earn_Vol*. In Panel A, we use random samples chosen from the MD&A disclosures. In Panel B, we use random samples chosen from conference call transcripts. Avg.Dep. means the mean value of the dependent variable. *t*-values are reported in parentheses and standard errors are clustered at the industry level. ***, **, and * denote statistical significance at 1%, 5%, and 10% level, respectively. Refer to Appendix A for detailed variable descriptions. Continuous variables are winsorized at 1% and 99%.

Panel A. MD&A Sample

Dep Var =	$IPT_{[0,5]}$		*PIN*		*Spread*	
	(1)	(2)	(3)	(4)	(5)	(6)
Bloat	-3.131**	-2.936**	1.863***	1.593**	0.026***	0.103***
	(-2.14)	(-2.03)	(2.71)	(2.29)	(5.21)	(4.39)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	Yes	No	Yes
Industry FE	Yes	No	Yes	No	Yes	No
Avg. Dep.	3.132	3.132	9.751	9.751	0.013	0.013
Cluster	Industry	Industry	Industry	Industry	Industry	Industry
N	1,790	1,787	1,790	1,787	1,790	1,787
Adjusted R^2	0.136	0.152	0.122	0.168	0.151	0.186

Panel B. Conference Call Sample

Dep Var =	$IPT_{[0,5]}$		*PIN*		*Spread*	
	(1)	(2)	(3)	(4)	(5)	(6)
Bloat	-1.755***	-1.411***	1.723***	2.265***	0.067***	0.145***
	(-3.20)	(-4.16)	(3.11)	(3.96)	(6.50)	(9.24)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Time FE	Yes	Yes	Yes	Yes	Yes	Yes
Firm FE	No	Yes	No	Yes	No	Yes
Industry FE	Yes	No	Yes	No	Yes	No
Avg. Dep.	3.565	3.613	10.203	10.035	0.015	0.014
Cluster	Industry	Industry	Industry	Industry	Industry	Industry
N	8,907	8,875	8,907	8,875	8,907	8,875
Adjusted R^2	0.105	0.198	0.112	0.165	0.099	0.178

Table 8. Theme-specific Summaries

This table reports descriptive statistics of theme-specific summaries and their informativeness. In Panel A, we report the annual averages of %Fin, %ESG, *len*ESG, and *len*Fin. %Fin is the percentage of summaries that contain financial-related contents each year. %ESG is the percentage of summaries that contain ESG-related contents each year. *len*ESG is the length of ESG-specific summary scaled by the length of entire summary. *len*Fin is the length of financial-performance-specific summary scaled by the length of entire summary. In Panel B, we regress two-day market-adjusted abnormal returns on theme-specific sentiment. $Sentiment^{Fin}$ is the sentiment of financial-performance-related summaries. $Sentiment^{ESG}$ is the sentiment of ESG-related summaries. We run the regressions each year and report the coefficient on each sentiment variable along with its statistical significance. *t*-value time trend is obtained by regressing *t*-values on years.

Panel A. Time Trends

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
%ESG	22.7	25.7	25.4	26.8	26.3	31.3	32.4	31.2	37.2	38.1	40.4	43.7
%Fin	100	100	100	99.4	100	97.8	86.8	89.1	99.4	99.5	99.0	96.0
*len*ESG	2.2	2.6	2.8	2.6	3.5	2.9	4.0	2.8	4.3	3.3	4.1	4.3
*len*Fin	21.8	22.3	19.8	21.4	21.5	16.9	17.5	20.7	21.3	19.7	22.2	21.2

Panel B. Theme-Specific Sentiment and Stock-Market Reactions

Dependent Variable	$= CAR_{[0,1]}$			
	$Sentiment^{ESG}$		$Sentiment^{Fin}$	
	Coefficient	*t*-value	Coefficient	*t*-value
Year	(1)	(2)	(3)	(4)
2009	-0.032	-0.62	0.010	2.01
2010	0.043	1.03	0.025	1.79
2011	0.054	1.48	0.035	1.64
2012	0.052	1.28	0.026	2.27
2013	0.021	1.46	0.072	2.13
2014	0.048	1.60	0.065	2.08
2015	0.027	1.67	0.098	2.32
2016	0.046	1.78	0.110	2.28
2017	0.073	1.95	0.098	2.06
2018	0.023	1.39	0.092	1.87
2019	0.064	2.12	0.089	2.53
2020	0.076	2.27	0.068	2.12
Full Sample	0.048	1.88	0.063	2.10
t-value Time Trend	0.159	2.80	0.031	1.63