Bloated Disclosures: Can ChatGPT Help Investors Process Information?

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

First draft: April 20, 2023
This draft: January 30, 2024

Abstract

Generative AI tools such as ChatGPT can fundamentally change the way investors process information. We probe the economic usefulness of these tools in summarizing complex corporate disclosures using the stock market as a laboratory. The unconstrained summaries are remarkably shorter compared to the originals, whereas their information content is amplified. When a document has a positive (negative) sentiment, its summary becomes more positive (negative). Importantly, the summaries are more effective at explaining stock market reactions to the disclosed information. Motivated by these findings, we propose a measure of information "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 constructing targeted summaries that identify firms' (non-)financial performance. Collectively, our results indicate that generative AI adds considerable value for investors with information processing constraints.

Keywords: ChatGPT, GPT, LLM, generative AI, informativeness, information processing, MD&A, conference calls, summarization, disclosure, information asymmetry.

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

CAVEAT NOTICE:
The findings presented here are preliminary and subject to revision. We are currently reevaluating our data, methods, and conclusions.

*The University of Chicago, Booth School of Business, alex.kim@chicagobooth.edu
†The University of Chicago, Booth School of Business, maximilian.muhn@chicagobooth.edu
‡The University of Chicago, Booth School of Business, valeri.nikolaev@chicagobooth.edu

We benefited from helpful comments by Bok Baik, Ryan Ball, Tom Barry, Elizabeth Blankespoor (discussant), Lin William Cong, Yiwei Dou, Zhiguo He, Gerard Hoberg, Jiekun Huang (discussant), Jing Huang, Raffi Indjejikian, Antonis Kartapanis, Eunjee Kim (discussant), Jinhwan Kim, Douglas Laporte, Reuven Lehavy, Christian Leuz, Bradford Levy, Gregor Matvos, Greg Miller, Martin Nienhaus, Joseph Piotroski, Doron Reichmann, Kristi Rennekamp, Laurence Van Lent, Paul Zarowin and workshop participants at Stanford University, University of Bochum, University of Chicago, University of Michigan, and conference participants at the 2023 GSU-RFS FinTech Conference, 2023 Korean International Accounting Conference, 2023 NYU Big Apple Accounting Conference, and the 2023 Texas A&M University Conference. Tom Barry provided excellent research assistance. The authors gratefully acknowledge financial support from the University of Chicago Research Support Center, the Fama-Miller Center for Research in Finance, and the Stevens Doctoral Program at the University of Chicago Booth School of Business.

I Introduction

The exponential growth of textual data in accounting and finance presents both opportunities and challenges for efficient information processing (e.g., Goldstein et al., 2021; Bae et al., 2023; Bochkay et al., 2023). The sheer volume and inherent unstructured nature of such data can quickly overwhelm economic agents. Large language models (LLM), epitomized by tools such as ChatGPT, offer innovative ways to analyze and interpret large volumes of text, potentially allowing investors to make more informed decisions. While the popularity of these tools has skyrocketed since November 2022, their economic usefulness in information processing remains unclear. Anecdotally, financial firms are showing a marked interest in such language models. In a recent interview, Citadel's 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 the usefulness of large language models in distilling 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 less relevant or potentially redundant textual information in corporate disclosures.

Corporate disclosures provide an ideal platform for understanding the value 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; Guay et al., 2016; Dyer et al., 2017). Indeed, the median length of a firm's 10-K increased by more than 100% between 1996 and 2013 (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 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 continue to grow. Financial statements have been increasing in length over the past 20 years, and recent evidence indicates that investors are

inattentive to their content (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 humanly understandable manner. The model is pre-trained on a vast language corpus and then fine-tuned for specific tasks. 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 retain the most relevant information when generating summaries. It is known to outperform other existing models in summarization tasks (Bhaskar et al., 2022; Goyal et al., 2022).[1]

We exploit this emerging technology to address the following questions. How effective are generative language models, and GPT-3.5 in particular, at distilling the information contained in complex corporate disclosures? How does the information content of a summary 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, is the model effective at constructing targeted summaries when investors are interested in a specific topic, such as financial or ESG performance?

To answer these questions, we focus on two primary types of corporate disclosures: Management Discussion and Analysis (MD&A) and earnings conference calls. The MD&A section 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 about the informativeness of MD&A disclosures, e.g., in relation to boilerplate language or excessive complexity (e.g. Brown and Tucker, 2011). Conference calls are voluntary events held by companies to help investors process reported performance and answer questions from analysts. They enable analysts to ask for additional information or clarification when disclosure is opaque. They are also more spontaneous, which possibly makes them harder to summarize. Both disclosure types carry useful information and are complementary.

We begin our analysis by constructing a random sample that constitutes about 20% of the pop-

[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).

ulation of MD&As and conference call transcripts from 2009 to 2020.[2] 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 subsequent analysis, we narrow down the prompt to retrieve information related to financial performance or ESG-related activities only.

We find that the average length of unconstrained summaries of MD&As (conference call transcripts) is about 25% (30%) compared to the average original document's length. This result points to potentially large gains in information processing. The key question that arises is whether the summaries, which are bound to omit many details, are still informative. In fact, we observe a slight reduction in the readability of the summarized document compared to the original. We next examine the summaries' information content.

From the perspective of an unboundedly rational investor with unconstrained information processing capacity, the original document is at least as good as the summary. Such an investor will filter out noise like a machine does and undo any possible biases in the disclosed information. The length, complexity, and information overload become irrelevant. For this reason, we focus on a scenario where an investor reading the document has limited attention (or information processing constraints). The investor's objective is to understand the primary implications of the disclosed information for a firm's future prospects. To this end, when dealing with a lengthy and complex document, it is plausible that the investor will largely focus attention on identifying positive vs. negative content. This approach is most closely captured by Loughran and McDonald (2011) measure of sentiment, which relies on positive vs. negative word counts in financial documents and which serves as a basis for our primary tests.[3] However, in what follows, we also discuss and analyze other dimensions of the summarized content.

We find that the sentiment of the summarized document is more pronounced relative to that of the original. In particular, when the original's sentiment is positive (negative), the summary is

[2]Due to initial data availability of the respective transcripts, we focus on a sample period which overlaps with the training phase of the GPT-3.5 turbo model, which was trained on data up until September 2021. However, the GPT model is designed and trained to prevent over-fitting using a number of techniques, including regularization, dropout, and the use of a large, diverse corpus. The model is known to generalize well on unseen data (Brown et al., 2020). Note, too, that it is difficult to explain our results by in-sample over-fitting of text summaries. Nevertheless, we extend our analysis by using a clean out-of-sample test reported in Appendix D. We find equally strong results despite the relatively small size of the dataset both for MD&As and conference call transcripts.

[3]We specifically do not use the state-of-the-art machine-based sentiment scores (e.g., Frankel et al., 2022; Huang et al., 2023), like BERT-based sentiment, because the objectives of such measures (trained on large amounts of data and often using similar architecture) is precisely to learn to filter out redundant or irrelevant information, which is what human readers would find prohibitively costly to do.

even more positive (negative). This can happen, for example, if companies (executives) "hedge" their views by using precautionary statements that do not contain significant amounts of information or are largely boilerplate.

While the summary could lose essential information, as the above discussion suggests, to the extent the model is able to distill the most relevant information, we expect summaries to capture true underlying sentiment more accurately.[4] To evaluate the information content of the summary vs. the original, we compare their ability to explain stock market reactions to the disclosed information. The assumption underlying this test is that the stock market as a whole is effective at processing publicly disclosed information, even though individual investors are subject to information processing constraints, and thus could be used as a benchmark to evaluate information content.

Motivated by the above discussion, we regress short-window abnormal stock returns surrounding the disclosure days on the sentiment scores of the summaries vs. the original documents. We find strong evidence that the summary sentiment is considerably more informative than the original's sentiment in explaining market reactions to the disclosed information. The summary sentiment also has a significantly greater explanatory power, both statistically and economically. We also find a sharper increase in the informativeness for the MD&A summaries compared to conference calls, which is intuitive as conference calls are less subject to concerns related to obfuscation or boilerplate language. Importantly, we also show that these results generalize when we analyze the post-2021 sample, which lies outside the GPT training period.

Our findings reveal a remarkable ability of the language model to condense disclosed information while maintaining and, in fact, enhancing the information content. The results also point to potential 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 triggers 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*.

To illustrate the measure and assess its face validity, consider the example of Disney's infa-

[4]The loss of information can happen, for example, when the original comingles positive and negative words or the same information is repeated multiple times.

mous conference call held in November of 2022. The call garnered extensive media coverage as Disney's CEO was accused of sidestepping and diminishing the company's bad performance, a controversy that led to his dismissal two weeks later. The call also spent considerable time discussing less relevant topics, such as the 100th-anniversary celebration. As we show, our *Bloat* measure spikes sharply for this conference call and, in fact, reaches the maximum level for Disney in our sample. An in-depth examination of the call and its summary shows that GPT successfully excises irrelevant details (e.g., the CEO's family visit to Disneyland) while retaining most critical information (e.g., operational losses in Disney's direct-to-consumer business).[5] We also perform a more systematic analysis of the excised content of other conference calls to reinforce this conclusion.

Utilizing the insights gleaned from this example, we examine the extent to which *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% of the variation, whereas 70% of the variation in *Bloat* is firm-specific. Firm fixed effects explain around 35-45% 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, though it also exhibits persistence. 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. 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 analysis, we explore the usefulness of generative LLMs to produce targeted sum-

[5]The 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 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 repeat the summarization process for Disney's earnings calls 50 times each. The variation in the lengths of the summaries is small, with a standard deviation of 0.010 relative to a mean of 0.702. Across all trials, we find that Disney's 2022-Q4 call has the highest *Bloat*. Furthermore, the Intraclass Correlation Coefficient (ICC) – a measure of reliability – is close to 1 (0.97).

maries. In particular, investors may be interested in understanding information concerning a specific topic (dimension), such as environmental impact or regulatory uncertainty. To explore this, in our first set of tests, we create prompt-based summaries that retain information about (1) financial performance or (2) "ESG" activities based on the original summaries. We find an increasing time trend in ESG-related content of conference calls. We also find that both types of summaries are incrementally informative in explaining market reactions and thus capture different dimensions in firms' communications. As one would expect, the sentiment from ESG-specific summaries becomes 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; Li et al., 2023). Overall, we conclude that generative LLMs show promise in extracting useful information along specified dimensions from lengthy and hard-to-read financial disclosures.

We contribute to the literature in four ways. First, we provide evidence on the economic usefulness of generative AI in analyzing textual data. As investors face significant information processing costs (e.g., Sims, 2003; Blankespoor et al., 2020; Cohen et al., 2020), AI summaries can become an important input into decision-making. We show that these summaries are concise and effective at retaining relevant information, thus allowing investors to focus attention on the information that matters. This suggests that generative AI has the potential to advance financial reporting technology in a way comparable in importance to, for example, the introduction of EDGAR (e.g., Gao and Huang, 2019; Chang et al., 2022; Goldstein et al., 2023) or XBRL (e.g., Blankespoor, 2019) by the SEC. Our study is related to Cardinaels et al. (2019), who study the summaries of earnings releases based on LexRank algorithm, which ranks and extracts key sentences from a text.[6] The research questions and methodology in our study are different.

Second, we offer a novel measure of the degree to which textual information contains redundancies and excessive details. Most prior studies on textual corporate disclosures focus on readability or linguistic complexity (e.g., Li, 2008; Loughran and McDonald, 2014a; Bonsall et al., 2017). In contrast, disclosure "bloat" is an intuitive but distinct construct with a substantial

[6]LexRank extracts the whole sentences from the text without transforming or connecting them. The algorithm focuses on sentence similarity based on bag-of-words representations, i.e., without understanding their content. As such, extracted sentences are disjoint and do not present a cohesive text. In contrast, generative AI has a deeper understanding of the substance and context of a given text. They find that LexRank summaries are more neutral than the original documents, which then lowers the perceived firm value for participants in a mTurk lab experiment.

firm-level variation. Further, Loughran and McDonald (2014a) point out that linguistic complexity commingles both textual and content complexity. *Bloat* is less subject to this concern because GPT is trained to summarize complex content while avoiding linguistic complexity. The measure can be easily applied in various corporate contexts and is of interest to investors and regulators. Along these lines, our study adds to the literature on transparency and its economic consequences (e.g., Leuz and Verrecchia, 2000; Lang et al., 2012). 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.

Third, and more broadly, we contribute to the literature on machine learning and AI in financial markets (e.g., Costello et al., 2020; Gu et al., 2020; Li et al., 2021).[7] More recently, contemporaneous work leverages advanced large language models (e.g., Huang et al., 2023; Kim and Nikolaev, 2023; Bernard et al., 2023; Chen et al., 2023; de Kok, 2023; Kim et al., 2023; Jha et al., 2023; Lopez-Lira and Tang, 2023) for various tasks, such as sentiment analysis or assessing firm's risk profile. In particular, Kim and Nikolaev (2023) use BERT to measure the value of contextual information in interpreting accounting numbers. Bernard et al. (2023) train a GPT-based LLM and use the model's confidence in predicting iXBRL tags to generate a modular measure of firm's business complexity.[8] Our study adds to this literature by exploring the value of generative AI in processing complex corporate disclosures and addressing information overload.

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; Li et al., 2023). For example, measuring a company's environmental activities is a highly complex task, and large language models show substantial promise in helping investors with this task.

[7] A related stream of research shows that machine learning techniques are powerful tools for various prediction tasks. For example, machine learning techniques have been shown to be useful in measuring disclosure sentiment (Frankel et al., 2022), predicting future earning surprises (Chen et al., 2022) or detecting accounting misstatements (Bertomeu et al., 2021).

[8] Subsequent work uses GPT-based LLMs to construct investment scores from earnings calls (Jha et al., 2023) as well as to measure the information surprise in firms' earnings calls (Bai et al., 2023) or in corporate filings (Costello et al., 2023). GPT has also shown promise in systematically detecting corporate events from boilerplate language on a large scale, such as tax audits (Choi and Kim, 2023; Armstrong, 2023). Another stream of literature focuses on the firm-level consequences of ChatGPT technology, typically by analyzing the market reactions to its release (Eisfeldt et al., 2023) or to its ban (Bertomeu et al., 2023).

II Generating Summaries with GPT

GPT is a large language model trained on a vast corpus of text data. Its objective is to predict the next word in a sentence conditional on the preceding words and a broader context. In this section, we describe how GPT generates summaries. We then motivate and illustrate 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 (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, $X = (x_1, x_2, \cdots, x_n)'$, where the number of rows corresponds to the number of words (tokens), and the number of columns is the dimension of 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 $(x_i^1, x_i^2, \cdots, x_i^{100})$ $(1 \leq i \leq 19)$, this sentence is represented by a 19×100 matrix.

The central component of the Transformer architecture is the so-called self-attention mechanism, which tells the model what to focus attention on when performing summaries and other tasks. Specifically, the attention mechanism enables the model to learn each word's relevance (relative importance) in an input text by considering its positional and contextual relationships with other words. To capture the relationships among 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 against which we want to compare the current query token. Lastly, each row in the 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 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 $n \times n$ 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

[9]It is very natural that *Our* receives a high self-attention score. *Our* is associated with the main verb *dedicate* and

"*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 while omitting redundancies, boilerplate content, and excessive details. Simultaneously, this example also effectively illustrates that any summary inherently encompasses a subset of the information found in the original text (e.g., it omits the phrase "compared to our competitors"). We discuss a more detailed real-world example in subsection C.

B Limited Attention and Underpinning of Disclosure Bloat

In this subsection, we analyze the value of language modeling to an investor with limited attention and motivate our measure of informational "bloat." An investor with limited resources cannot attend to all available information without incurring a substantial cost. Ideally, such an investor wants to know what information to focus attention on when making decisions. Intuitively, this is what a generative language model can enable investors to do. It ranks the words contained in the input text based on their self-attention scores and uses the ones with high scores to generate the summary.

To formalize this idea, consider an investor who needs to process a document with n messages (words), some of which are more decision-relevant than others. The most (least) relevant message deserves the highest attention and thus carries the highest (lowest) attention score normalized to one (zero). Accordingly, we can define the attention score associated with i-th message $s(i)$: $R \to [0, 1]$.

The investor does not know what to focus attention on and thus tries to process all messages in the document at the same time (albeit imperfectly). The marginal benefit of processing a message with a lower attention score decreases, whereas the marginal cost increases. Intuitively,

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.

the cost can be manifested as an additional time requirement or deteriorated decision quality. This can be captured in the following formulation of a user's *ex-ante* (i.e., before the actual information content is known) utility from processing textual information:

$$U(n) = \underbrace{B\Big(\sum_{i=1}^{n} s(i)\Big)}_{benefit} - \underbrace{C\Big(\sum_{i=1}^{n}(1-s(i))\Big)}_{cost} \tag{2}$$

where $B(.)$ is an increasing concave "benefit" function and $C(.)$ is an increasing convex "cost" function. The user's ex ante utility increases with the relevance of each message, $s(i)$, and decreases with the "noise", $1 - s(i)$.

The process of constructing an effective summary thus involves a trade-off between including more information (signal) and dealing with unnecessary details or redundancies (noise). GPT's self-attention mechanism enables this trade-off by identifying attention scores corresponding to specific words and including the most relevant words into a summary. For example, one can think of the arithmetic average of the elements q_{ij} in i-th row of the attention matrix $\text{Score}(Q,K) = \text{softmax}\left(\frac{Q\cdot K'}{\sqrt{\dim K}}\right)$ as relevance score $s(i)$ associated with the i-th word:

$$s(i) = \frac{1}{n}\sum_{j=1}^{n} q_{ij} \in [0,1] \tag{3}$$

With the help of GPT's attention mechanisms, an investor could increase her utility by focusing on content with high attention scores $s(i)$ and omitting less relevant content. To see this, one can rank the elements $s(i)$ in descending order: $\{s(1),\cdots,s(n)\} \to \{\bar{s}(1),\cdots,\bar{s}(n)\}$ such that $\bar{s}(i) \geq \bar{s}(i+1)$, and formulate the problem as:

$$\max_{0\leq k\leq n} U(k) = B\Big(\sum_{i=1}^{k}\bar{s}(i)\Big) - C\Big(\sum_{i=1}^{k}(1-\bar{s}(i))\Big) \tag{4}$$

The solution to this discrete program is the optimal summary length n^* that is approximated by the first-order condition:

$$B'(\cdot)\bar{s}(n^*) = C'(\cdot)(1-\bar{s}(n^*)) \tag{5}$$

For example, assume that $B(\cdot)$ and $C(\cdot)$ are linear with a slope of one to ease interpretation. The marginal benefit of including k-th message is an increase in the cumulative "signal" by $\bar{s}(k)$, while the marginal cost is an increase in "noise" by $1 - \bar{s}(k)$. An optimal length would include words as long as the marginal benefit exceeds the marginal cost. Since $\bar{s}(k)$ is a decreasing function of k, there exists an optimal length $k = n^*$ that maximizes $U(k)$. Figure 1 visually illustrates the relationship between summary length and $U(k)$.

Note that the optimal length n^* depends on the functional form of $\bar{s}(.)$. For example, if $\bar{s}(.)$ is convex, the optimal summarization point n^* is smaller than $\frac{1}{2}n$. Intuitively, convex $\bar{s}(.)$ implies that there are relatively fewer important messages, 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 words are needed to convey the relevant information, the model generates a longer summary.

Based on the above, we define the informational *Bloat* measure as follows:

$$Bloat = \left(\frac{n - n^*}{n}\right) \tag{6}$$

The spirit of this measure is to capture how far the document deviates from its optimal summary length n^*. Higher *Bloat* implies that the original disclosure text contains relatively more uninformative, less relevant, or repetitive content. Intuitively, bloated disclosure has a high frequency of tokens with less important information content, thereby leading to a convex $\bar{s}(.)$.

It is important to note that the optimal summary length can vary depending on user objectives. Some investors prefer a concise summary, while others can ask GPT to generate a more detailed summary. Users can further specify the types or dimensions of information they consider to be most relevant for their objective by instructing the model using prompts to pay attention to, e.g., information related to corporate social responsibilities, regulatory risks, etc. Thus, generating "optimal" summaries generally requires fine-tuning the model or developing a set of prompts to meet the objectives of a given user. Because we use a generic prompt to generate summaries, our empirical results should be viewed as a lower bound of the value added by the summaries.

In sum, the above discussion illustrates the usefulness of language modeling to an investor with limited attention and offers justification for our measure of informational "bloat." We next

illustrate the measure using a real-world example, before discussing the implementation of the model in more detail.

C Illustration of Disclosure Bloat

To evaluate the face validity of the newly introduced *Bloat* metric, we conducted an in-depth analysis of Disney's 2022-Q4 earnings call. Just before the call, Disney disclosed a disappointing quarterly result that led to about a 10% decline in its stock price. Compounding this negative performance was former CEO Bob Chapek's approach to the earnings discussion, which triggered a barrage of criticism. Media outlets described the earnings call as "inconceivable, "delusional," and "bizarre." Chapek particularly drew flak for his evasion of the company's poor results, choosing instead to diverge into discussions on largely irrelevant subjects.[11] Less than two weeks following the contentious earnings call, Bob Chapek was dismissed, prompting the return of his predecessor, Bob Iger.

Given contemporaneous media reports criticizing Chapek's behavior during the earnings call, it's intriguing to examine whether our measure is capable of detecting the apparent bloat in Disney's call. To do so, we use (Chat)GPT-3.5 to summarize each of Disney's earnings call transcript going back to 2015 and present the calculated bloat graphically in Figure 2(a). The figure reveals that disclosure bloat is relatively steady over time and spikes sharply during the infamous quarter. This observation suggests, even at first glance, that the generative AI possesses the capability to discern clutter and effectively sift through it.

We proceeded to examine the actual content of both the original and the summary of the 2022-Q4 earnings call. Appendix B displays the beginning of Chapek's presentation during the earnings call. In the text's first section (Panel A), Chapek portrays a highly optimistic view of Disney's standing. Notably, while the summary maintains the original's upbeat tone, it does not fail to mention Disney's operating losses and their plans to address them. More interestingly, the summary becomes much shorter for the second part (Panel B), during which Bob Chapek shifted his presentation to the discussion of largely irrelevant subjects. Specifically, nearly two

[11] For example, entertainment lawyer and journalist Matthew Belloni reported that *"[i]t was one of the most bizarre things I have ever seen. Because he is sitting there and talking about all the great things at Disney and the Oogie Boogie Bash at Halloween and how many tickets they sold for that. And the stock was going down, down, down."* or Jim Cramer on CNBC *"I had faith, but there is just no doubt that he has to go. That was just unconscionable. And the quarter itself, the way he handled it ... He made it sound like it was just a four-star quarter. Delusional."*

paragraphs concerning unrelated celebratory events, such as Chapek's family visit to Disneyland for Halloween or the success of the "Oogie Boogie Bash" are summarized by GPT into a single short sentence. This in-depth analysis provides useful preliminary evidence illustrating the ability of GPT's self-attention mechanism to focus on the information that matters. We next discuss the implementation of our model in more detail.

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; Cao et al., 2023). To construct our sample, we obtain machine-readable MD&A sections of annual reports and earnings conference call transcripts for all US non-financial public firms from fiscal years 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 40,362 conference calls.[12]

Due to GPT processing constraints, we then draw a random sample of about 20% of unique firms relative to the population described above. This results in 1,790 MD&As reported by 339 firms and 8,537 conference calls held by 360 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, 2023). 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

[12]Considering the need to control for earnings surprise in our analysis, we require that at least one analyst covers a given firm. This prerequisite tends to favor the inclusion of approximately 1,750 larger firms. Although it narrows our focus, we believe this sample more accurately reflects the segment of the market that is of greatest interest to investors.

algorithm. Our retrieval rate is comparable to other studies (e.g. Cohen et al., 2020). We delete figures, tables, html script, and XBRL tags.[13]

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 obtain conference call transcripts from S&P Global Capital IQ Pro Transcripts and use the entire 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.[14]

Other Data. We use the CRSP database for daily stock and market returns. We rely on the 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 the GPT-3.5-Turbo API provided by the OpenAI ChatCompletion endpoint to construct summaries.[15] 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 this limit and process each chunk separately. The token limit counts both input and output texts. Therefore, we divide the input text into chunks with a maximum of 2,048 tokens, leaving another 2,048 tokens for the summaries.[16] We then concatenate the generated summaries of each chunk and create a single summary for the entire document.[17]

We use a prompt that instructs the model to summarize the input text using only the information included in the text and to not restrict the length of the summary. We deliberately opted for such a straightforward prompt to establish a foundational baseline for the effects. Nonetheless, employing more intricate prompts could facilitate the extraction of more tailored information

[13]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.

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

[15]GPT-3.5 underpinned the initial version of ChatGPT from November 2022. A more recent version for subscribers from March 2023 relies on GPT-4.0. However, due to unknown system prompts and model calibrations in ChatGPT (e.g., browser or request-specific word limit), the API needs to be used for consistent summarization.

[16]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.

[17]The upcoming GPT-4.0 Turbo API is expected to have a substantially increased token limit, enabling the processing of financial documents without chunking.

(e.g., see Section VI).[18]

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 (1,801) unique firms in the MD&A sample (conference call sample) and then randomly choose 20% of these unique firms (339 and 360 firms, respectively). We then retain all documents associated with these firms for GPT processing.[19]

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 call samples do not differ materially from each other. For example, the mean *Log_ME* (the natural logarithm of the market value of equity) of the MD&A sample is 7.713, while the mean *Log_ME* of the conference calls sample is 7.798.

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 commonly studied in prior literature: length, sentiment, and readability (Fog index and plain English measure).[20] 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

[18]The prompts are provided in Online Appendix A. 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. Our prompt only exploits the user role of the API and does not specify a system or assistant role. Furthermore, due to the probabilistic nature of the model, the results might slightly vary for each trial (e.g. de Kok, 2023). Therefore, we assess the robustness of our results by returning to our Disney example and summarizing each transcript 50 times. We find a low standard deviation in bloat of 0.010 relative to a mean of 0.702. Across all trials, 2022-Q4 shows the highest disclosure bloat (see Figure 2(b).

[19]Compared to randomization on the document level, this cluster randomization allows us to track changes in the summaries within-firm across time.

[20]Another commonly used proxy, the Bog index, is only readily available for full 10-Ks via Bonsall 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.

McDonald (2011):

$$Sentiment = \frac{\sum_{x\in\mathcal{D}} \mathbf{1}[x \in \mathcal{P}] - \sum_{x\in\mathcal{D}} \mathbf{1}[x \in \mathcal{N}]}{\sum_{x\in\mathcal{D}} \mathbf{1}[x \in \mathcal{P}] + \sum_{x\in\mathcal{D}} \mathbf{1}[x \in \mathcal{N}]} \tag{7}$$

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, 1952), but adjusted for financial terms (Kim et al., 2019). It measures the percentage of complex words per sentence as a measure of readability (e.g. Li, 2008; Loughran and McDonald, 2014a). A higher *Fog* indicates less readable disclosures. We supplement the Fog index by the "plain English" measure (*Plain_Eng*) calculated in accordance with the narrative disclosure guidelines set forth by the SEC.[21] Analogous to *Fog*, higher *Plain_Eng* indicates lower readability.

Linguistic Attributes. We calculate the linguistic attributes for both raw and summarized documents and present them in Table 2 and Figure 3.

Panels A of Table 2 and Figure 3(a) focus on the MD&A sample. The levels statistics indicate that the GPT model reduces the document length of MD&As by nearly 80%, on average. The average (median) length of the original documents is 17,901 (14,254) words, whereas it goes down to 3,779 (3,433) 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 (-0.360) and that it becomes slightly more negative in the summarized documents (-0.366). More importantly, the standard deviation of sentiment increases from 0.202 (raw) to 0.316. This widening of the distribution of sentiment is indicative of sentiment becoming more clear-cut.

In terms of textual complexity, the average *Fog* in raw documents is 10.025 and the average plain English measure is 0.289. Based on common rules-of-thumb, both scores indicate high textual complexity in the raw documents. The summarized documents become slightly more complex in general, with an average Fog index of 10.175 and a plain English measure of 0.290.

[21] Specifically, following Loughran and McDonald (2014b), we calculate the following five 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 the word "respectively". We then standardize (iii), (iv), and (v) with document length and add all five elements to obtain *Plain_Eng*.

This phenomenon likely arises because summaries inevitably need to include financial jargon in a relatively shorter document. The changes statistics in Panel A 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 B of Table 2, as well as Figure 3(b). One notable difference is that conference call summaries tend to be relatively longer, i.e., close to 30% of the original transcript, on average. This result is intuitive, given their less structured nature. Unlike MD&As, the call transcripts have, on average, a positive sentiment. Overall, however, the inferences remain qualitatively similar. The summarized sentiment continues to have a higher standard deviation than the 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.

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, on average, 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 summarized sentiment is significantly more negative than the original sentiment. Figure 3(a) helps to visualize these findings. As illustrated in the figure, GPT summarization makes a relatively positive document more positive and a relatively negative document more negative. These results support the notion that GPT summaries amplify the sentiment of the raw documents.[22]

In terms of readability measures, for the above-median sentiment documents, the summaries exhibit a higher Fog index and plain English measure (10.310 and 0.289, respectively) compared to the raw documents (9.690 and 0.284, respectively). Interestingly, for the below-median doc-

[22] A widening of the sentiment distribution alone is neither necessary nor sufficient to ensure that the summaries contain more relevant information. However, as we show in subsequent analyses, the summarized sentiment also better explains stock market reactions to the disclosed information.

uments, the summaries are now less complex in terms of Fog index (10.040), which is smaller compared to raw documents (10.362). Comparing the two MD&A partitions yields some insights into this finding. In line with the managerial obfuscation hypothesis (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 3(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 summary sentiment of the above-median group is 0.424, while the average raw sentiment of the same group is 0.398 (with the difference being statistically significant at the 1% level). In contrast, the average summary sentiment of the below-median group (0.082) is slightly smaller than the corresponding raw sentiment (0.092). Overall, we observe the same asymmetric effect of summarization on sentiment in the conference call sample.

Similar to 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 (2,474 for above-median and 2,125 for below-median) and readability (Fog index of 10.163 versus 10.192 and plain English measure of 0.261 versus 0.261).

Taken together, our evidence suggests that summarization appears to amplify textual sentiment. As in prior studies (e.g. Li, 2008), narrative disclosures 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. To shed further light on the informational value of the summarized disclosures, the following subsection examines their information content more directly.

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 the standpoint of a user with a limited information processing capacity (e.g., Hirshleifer et al., 2004; Blankespoor et al., 2020; Lu, 2022). We take the perspective of an average investor who reads MD&As and conference call transcripts with the objective of learning key takeaways. We proxy for the primary takeaway by the document's sentiment.[23] However, we also explore textual uncertainty in the online appendix (see Table OA-3). In the subsequent sections, we also analyze the dimensions of removed content (disclosure bloat) and explore the informativeness of targeted summaries focused on financial or ESG dimensions of disclosed information.

Test Design. Our primary test compares raw document sentiment vs. summary sentiment in explaining stock price reactions to disclosure. Intuitively, if the model is effective at summarizing the most relevant information (key takeaways), the sentiment associated with such information should be more descriptive of stock price movements as compared to the sentiment of the entire document. Note that we make an assumption that stock prices are generally efficient at aggregating publicly disclosed information even when individual investors are subject to information processing constraints. Specifically, we estimate the following regression:

$$Abn\_Ret_{[0,1]it} = \beta Sentiment^{j}_{it} + \gamma \mathbf{X}_{it} + \delta_t + \theta_i + \varepsilon_{it} \tag{8}$$

where $Abn\_Ret_{[0,1]it}$ is firm i's cumulative abnormal return over the window of two trading days (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. $Sentiment^{j}_{it}$ is either $Sentiment^{Raw}_{it}$ (raw document sentiment) or $Sentiment^{Sum}_{it}$ (summary-based sentiment) of firm i at time t. $\mathbf{X}_{it}$ is a vector of firm-level control variables: 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 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.

[23] Although information content is a multi-dimensional concept, understanding a document's sentiment is arguably the primary interest of academics and practitioners (e.g., Henry and Leone, 2016; Chen et al., 2023).

Results. Table 4 presents our findings. We start by discussing the MD&A sample (Panel A). Columns (1)-(4) examine the informativeness of raw document sentiment (raw sentiment). Columns (1)-(2) explore different fixed effects structures, whereas Columns (3)-(4) condition the sample on whether the sentiment is above (Pos) or below (Neg) the median, provided our prior findings that the two subsamples behave somewhat differently. 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 the price before its release. The following analysis indicates that the former is a more appropriate explanation.[24]

Columns (5)-(8) use the summary-based sentiment (summary sentiment) as our explanatory variable. We find the striking result that summary sentiment exhibits highly statistically and economically significant associations with abnormal stock returns. This result holds across all models. For example, the coefficient on $Sentiment^{Sum}$ is 0.025 (standard error = 0.004) for the model with year and industry fixed effects and 0.051 (standard error = 0.009) when we include year and firm fixed effects. The results show considerable economic magnitudes. Based on estimates in column (6), a one standard deviation increase in $Sentiment^{Sum}$ translates to a 0.336 standard deviation increase in abnormal returns (or 161 basis points). We also find that summaries are more informative for a more negative sub-sample of MD&A sentiment (column (8)) in line with prior evidence that GPT successfully navigates through disclosure fog (recall that negative sentiment is associated with more complex or less readable disclosures). The corresponding adjusted R^2 also increases considerably when we use summarized sentiment.[25]

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 sam-

[24]Since the summary-based sentiment 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.

[25]Interestingly, Blankespoor et al. (2023) find that managerial tone on roadshows, but not the tone in S-1 filings, explains future accounting performance. As managers should similarly emphasize the most pertinent information in their shorter presentations, this finding mirrors the transformer architecture's method of summarizing financial filings.

ple, raw sentiment exhibits significant positive associations with stock price movement around the conference call date (consistent with prior literature such as Henry and Leone (2016)). One exception is that, when we partition the sample based on the median of the 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 0.050 (raw sentiment) to 0.097 (summary sentiment), and the statistical significance also increases with the calculated t-statistics increasing from 7.14 to 12.13) respectively. The economic magnitude of the summary sentiment implies that a one standard deviation increase in $Sentiment^{Sum}$ is associated with a 0.32 standard deviation increase in abnormal returns. Furthermore, there is a significant increase in adjusted R-squared from 7.3% (raw sentiment) to 14.1% (summary sentiment) for the same model.[26]

It is helpful to visualize these findings semi-parametrically by plotting the average cumulative abnormal returns across sentiment quintiles in Figure 4. For each period, we partition the sample into quintiles based on the value of raw or summary-based sentiment. We then calculate the average cumulative abnormal returns for each quintile and plot the results. The dotted line in Figure 4(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 upward trend in the summary-based sentiment. We repeat this exercise with the conference calls sample (Figure 4(b)). Here, we observe a positive slope for both lines, consistent with our regression results. However, the solid line is steeper, implying a stronger positive association between the summary sentiment and stock market reactions.[27]

We supplement our main results based on sentiment by analyzing the informativeness of

[26]To mitigate the concern that our finding is due to overreactions, we explore whether our returns experience a reversal for longer windows. Instead of focusing on a two-day window after the filing (conference call) dates, we also report 5-, 15-, and 30-day window returns in Online Appendix Table OA-1. Across different horizons, we observe similar positive coefficients for $Sentiment^{Sum}$

[27]Since GPT is trained on a wide variety of publicly available information, it is possible that GPT has seen more information about large companies during its training phase and, hence, generates better summaries for such companies. To test this, we interact firm size with the summarized sentiment in equation (8). As reported in Online Appendix Table OA-2, we observe a negative coefficient on the interaction term for both MD&As and conference calls. This result indicates that summaries are arguably even more helpful for small firms, which mitigates the above concerns. This finding is also consistent with our results in Section V.B in which we show that small firms are more likely to have *bloated* financial disclosures.

summaries with respect to the second moment of communicated information (e.g. Blankespoor et al., 2023). Specifically, we measure the level of textual uncertainty based on Loughran and McDonald (2014a)'s uncertainty dictionary and compute the level of uncertainty for both raw and summarized documents. In line with prior tests, we regress absolute abnormal returns of filing (conference call) dates and post-filing volatility on uncertainty indices and report the results in the Online Appendix, Table OA-3. Our findings indicate that the summary-based uncertainty measure has a superior ability to explain absolute abnormal returns and stock price volatility. This result is consistent with summaries being more informative not only along the first dimension (sentiment) but also along the other dimensions (uncertainty) of the communicated signal.

Alternative test samples. Even though GPT is neither trained to specifically summarize corporate disclosures nor to predict stock performance, one may be concerned that our sample period overlaps with the training period of the language model. This is unlikely to be an issue for our results because the model's objective, which is to predict the following word in a sentence, has nothing in common with detecting sentiment or explaining stock market movements. To mitigate this concern further, we explicitly direct the model to summarize only the information included in the input text and also use an independent methodology to measure sentiment (Loughran and McDonald, 2011). Nevertheless, we also directly address this concern by performing the analysis using data that originated outside of GPT's training period.

Specifically, we use our original sample of firms but restrict it to MD&As and conference call transcripts originating in the calendar year 2022. Since GPT stopped training in September 2021, any information generated in 2022 cannot be seen during the training phase. This design presents a clean "out-of-sample" test of the superior information content of the summarized documents. We present the results in Appendix D. We find that the results are similar and, in fact, even more pronounced compared the the results discussed above. We consistently find that the summaries' sentiment has a substantially greater explanatory power with respect to contemporaneous stock market reactions, as evidenced by higher statistical significance and higher adjusted R-squared values.

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.

D Understanding Excised Content

The transformer architecture allows GPT to focus on the most relevant portion of the text and omit less important content (see Section II.B). Following this structure, redundant statements and boilerplate language, which contain little incremental information, should be deleted first. Excessive or less relevant details, although they might be useful to some users (e.g., when performing an in-depth analysis), can also be removed or condensed. Our Disney example provided in Appendix B illustrates that GPT is capable of effectively removing or reducing such extraneous content.

To provide more systematic evidence on the deleted content, we perform three additional tests. First, we analyze the reduction in redundant or boilerplate information by measuring the cosine similarity scores of bag-of-words representations of adjacent period pairs. That is, for firm i, we compute the cosine similarity between time t and $t+1$ raw documents. Similarly, we compute the similarity scores for time t and $t+1$ summaries. We tabulate the results in Appendix C, Panel A. We find that the mean cosine similarity score for two adjacent (raw) MD&As is 85.94, whereas the mean score of the adjacent summaries is 73.51. In the case of conference calls, the mean cosine similarity between two adjacent raw call transcripts is 79.45, while that of the adjacent summaries is 70.75.[28] A considerable reduction in cosine similarity scores is indicative of the reduction in repeated or boilerplate language, suggesting that GPT summaries effectively remove uninformative contents (Brown and Tucker, 2011; Cohen et al., 2020).

Second, we calculate the percentage of boilerplate words in the raw and summarized documents following Lang and Stice-Lawrence (2015) and Dyer et al. (2017). Specifically, we obtain the tetragrams (four-word combinations) included in more than 75% of the filings (calls) each year and identify the sentences that contain such common tetragrams. We then calculate the number of words that are included in the boilerplate sentences scaled by the total number of words in each document. The results are in Appendix C Panel A. The boilerplate language use in (raw) MD&A documents is 5.45% in our sample. However, it is reduced to 2.66% in the corre-

[28] A reduction of approximately 10 in pairwise cosine similarity is sizable given that, for example, MD&A reports within the same industry typically have an LDA of around 55 (e.g. Peterson et al., 2015).

sponding summaries. The difference between these two numbers is statistically significant at the 1% level.[29]

Last but not least, we corroborate our quantitative analysis by utilizing GPT to identify the reasons for excised contents. We randomly sample 20 MD&As and 20 conference calls and then instruct GPT to compile the list of reasons for deleted content. For this task, we use GPT-4.0 to identify the difference between the raw and summarized documents.[30] After obtaining the differences for each document-summary pair, we then prompt GPT to explain why each component is deleted.[31] For MD&As, boilerplate language (32.50%) and repeated statements (18.24%) are the most common reasons for deletion, which is consistent with our findings in Panel A. Other important reasons behind deletions include historical recaps (15.5%) and excessive details (12.9%). For conference calls, excessive details (39.80%) and repeated statements (20.90%) comprise the most common reasons for removing the content, only then followed by boilerplate language (15.93%). Overall, these results indicate that the model is successful at removing less relevant information from the viewpoint of an investor with information processing constraints.

Taken together, our results support the conclusion that summaries tend to distill redundant or excessively detailed information from disclosures.

V Measuring Disclosure Bloat

The flip side of GPT's notable 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* is 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.

[29] In line with call transcripts being less subject to boilerplate language, the boilerplate use in raw conference call transcripts is only 1.12% and 1.02% in the corresponding summaries. The difference between the two numbers is not statistically significant.

[30] GPT-4.0 avoids the issue that the total input length of the original document plus the summary is a binding constraint. However, it is an order of magnitude more costly, so we do not perform a large-scale investigation

[31] See Online Appendix A for details on the procedure. Additionally, Appendix C provides the full statistics on the deletion reasons.

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 calls) sample is 0.754 (0.685), i.e., 75% (69%) of the original. In other words, GPT considers approximately 70-75% of the disclosed content to be less relevant when trading off its "signal" and "noise" components.[32] We also observe a meaningful variation in *Bloat*, which is 0.081 (0.126) for the MD&A (conference calls) sample. Overall, we observe a higher *Bloat* for the MD&A than for conference call samples.[33]

Figures 5(a) and 5(b) plot the average *Bloat* for the MD&A and conference calls samples, respectively. We observe several interesting patterns. First, *Bloat* seems to increase slightly over time yet with sizable fluctuations for both 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 are correlated over time; the Spearman correlation coefficient between annual numbers is 0.70. Lastly, in 2020, we note a steep decline in MD&A *Bloat* and an increase in conference call *Bloat*. We also observe a steep increase in length for both samples. This phenomenon is likely driven by corporate disclosures highlighting the impact of COVID-19.[34]

Variance Decomposition of *Bloat*. 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 R-squared after adding different sets of fixed effects. For the MD&A (conference calls) sample, year fixed effects explain only 1.64% (5.19%) of the total variation in *Bloat*. The inclusion of industry fixed effects increases the explained variation by 12.16% (3.36%). A more sizable portion of variance is explained by the interactions between industry and time fixed effects, which account for an additional 16.27% (19.95%) of the total variance. The remaining 69.93% (71.50%) of the total variance is thus attributable to firm-level factors. We zoom in on this firm-level variation at

[32] Note that this does not necessarily indicate that 70% of the disclosure is pure noise. This information can contain details that are generally less relevant from the perspective of a user with information processing costs.

[33] Based on the out-of-sample test in Appendix D, the average *Bloat* is 0.774 for MD&As and 0.731 for conference calls, in line with the estimates reported in Table 5.

[34] 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. Since COVID-19 often had a material impact on firms, GPT is likely to find such information relevant and include it in its summaries, evidenced by a lower *Bloat* in MD&A disclosures. However, our keyword search reveals that more than one analyst usually asks questions about the impact of the pandemic in approximately 40% of the conference calls. GPT likely condenses such repeated contents, resulting in a higher *Bloat*.

the bottom of Panel B by showing that firm-fixed effects, i.e., time-invariant firm characteristics, explain about 46.23% (35.99%) of firm-level variance. This means that less than half of firm-level variation in *Bloat* is time-varying.[35]

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 covers conference calls. The i-th row and j-th column intersection (c_{ij}) report 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.

Both Panels C and D indicate that approximately 35% 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, but it also exhibits some persistence.

B Determinants of *Bloat*

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

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

where $\mathbf{X}_{it}$ is a vector of firm-level determinants. Following Li (2008), we first include several variables that proxy for investment opportunities and incentives to obfuscate disclosed information. Specifically, 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. Next, we include variables that capture firms' performance and its variability: return-

[35]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 8.03% for MD&A *Bloat* and 4.62% for conference call *Bloat*.

on-assets (*ROA*), the textual sentiment of raw disclosure ($Sentiment^{raw}$), an indicator for whether a firm reports negative earnings (*Loss*), and the standard deviation of seasonally adjusted earnings (*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 make disclosures more opaque when things are not going well.

We also include several textual attributes that measure the readability or complexity of a firm: *Fog* and *Plain_Eng* proxy for readability, whereas *Log_Length* and *Complexity* proxies for firm complexity. We follow Loughran and McDonald (2023) to compute *Complexity*.[36] 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*, *Inst_Own*, *Earn_Vol*, *ROA*, *Loss*, *Sentiment*, *Log_Length*, and *Complexity*. A negative coefficient on *Log_ME* implies that larger firms have less bloated disclosures on average. We further find a positive coefficient on *Loss*, a positive coefficient on *Earn_Vol*, a negative coefficient on *ROA*, and a negative coefficient on *Sentiment*. Collectively, these relations suggest that *Bloat* increases as performance declines or becomes less stable. However, there are two plausible explanations for this finding: (1) *Bloat* may increase if negative performance is more difficult to explain, or (2) managers may try to obfuscate the bad performance (e.g. Li, 2008; Loughran and McDonald, 2014a). Although both explanations are plausible, the Disney example in Section II.C is more in line with the second explanation. Next, we observe that document length, *Log_Length*, and firm complexity, *Complexity*, have a positive and significant relation with *Bloat*. However, economically, the effect is modest. A 10% increase in length corresponds to a 0.008 (or 0.010 standard deviations) increase in *Bloat*. Jointly, the above variables explain about 64.4% of the variation in the MD&A *Bloat*. In column (2), we replace industry fixed effects with firm fixed effects. While the coefficient estimates are generally consistent regardless of the fixed effect structure, the adjusted R-squared increases to 77.1%.[37]

[36] We use this bigram-based measure of complexity as other measures (e.g. Hoitash and Hoitash, 2018; Bernard et al., 2023) are not available for our entire sample period.

[37] The inclusion of document length is one of the reasons for relatively high R^2. *Bloat* tends to increase with

For the conference call analysis, tabulated in columns (3) and (4), the results are generally similar, except that we find a positive and statistically significant coefficient on *Plain_Eng*, which is likely to reflect disclosure readability. Compared to the MD&A sample, the determinants of conference calls' *Bloat* jointly explain a somewhat lower portion of its variation (about 36.5% without and 55.6% 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, a reduction in disclosure quality leads to lower liquidity and higher cost of capital (e.g., Leuz and Verrecchia, 2000; Lambert et al., 2007). Specifically to our measure, the presence of excessively detailed, irrelevant, or redundant information is likely to slow down price discovery and give rise to informational asymmetries among investors trading the stock, e.g., to the extent some investors have better information processing capacity.

Test Design. We use three proxies that jointly capture the degree of price informativeness and information asymmetry. First, we calculate the probability of informed trading *PIN* (Easley et al., 1996) by following the algorithm in Brown and Hillegeist (2007). We use annual PINs for the MD&A and quarterly PINs for conference call samples.[38]

Second, we use abnormal daily bid-ask spreads measured on the announcement day, following Corwin and Schultz (2012). The daily spreads are calculated using intraday indicators, and we use simple-averaged spreads. Subsequently, we measure abnormal bid-ask spreads by taking the difference between the spread on the disclosure date and the average daily spread over the two-week period preceding the disclosure date. This design alleviates potential concerns that

document length, which is intuitive. By dropping *Log_Length*, the adjusted R^2s of all specifications decrease by about 30%. Other estimates remain qualitatively similar, however.

[38] 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.

firm-year (quarter) characteristics confound our tests.

Third, we use the post-filing volatility of daily stock returns (*Post_Vol*) as a measure of uncertainty in the informational environment created by firms' poor-quality disclosures (Loughran and McDonald, 2014a). Following their methodology, we estimate the market model over the period starting six days after the disclosure date and ending 28 days after the disclosure date. We then compute the root-mean-squared error of the residuals to measure post-filing volatility.

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

$$Info\ Friction_{it} = \beta Bloat_{it} + \gamma \mathbf{X}_{it} + \delta_t + \theta + \varepsilon_{it} \tag{10}$$

where $Info\ Friction_{it}$ is either PIN_{it}, Abn_Spread_{it} (abnormal bid-ask spread), or $Post\_Vol_{it}$ (post-filing volatility), $\mathbf{X}_{it}$ 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 Table 6 of Section IV.B and also include the absolute value of earnings surprise (*abs_SUE*), Friday indicator (*Friday*), and disclosure date returns (*One_Day_Ret*) to control for the news component of the announcement. Time fixed effects help us to eliminate common shocks and firm fixed effects to remove "persistent bloat" that is less likely to confuse market participants (Breuer and deHaan (2023)). 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. We use industry fixed effects in columns (1), (3), and (5), and firm fixed effects in columns (2), (4), and (6). In line with our expectations, we find that *Bloat* exhibits positive associations with the probability of informed trade, abnormal bid-ask spread, and post-filing volatility. In terms of economic magnitudes, a one standard deviation increase in *Bloat* is associated with a 0.8% point (=0.099 $\times$ 0.081) increase in the probability of informed trading, a 4.3% (=0.494 $\times$ 0.081) increase in abnormal bid-ask spread, and a 12.2% (=$\frac{0.027 \times 0.081}{0.018}$) increase in post-filing volatility from the mean (these values computed using coefficients in columns (1), (3), and (5), respectively).[39] These are considerable

[39]In additional tests, we include pre-filing volatility, which is measured between 257 days before the filing and 6 days before the filing, as a control variable when we use post-filing volatility as the dependent variable. Furthermore, in Online Appendix Table OA-4, we estimate abnormal volatility by computing the change between pre- and post-

economic magnitudes, which point to an economically important role of disclosure bloat in capital markets. The results remain similar when we focus on within-firm variation in columns (2), (4), and (6).

In Panel B, we repeat the same analysis for the conference calls sample. The inferences are qualitatively similar. A one standard deviation increase in *Bloat* is associated with a 0.4% point (=0.034 $\times$ 0.126) increase in the probability of informed trading, a 2.2% (=0.172 $\times$ 0.126) higher abnormal bid-ask spread, and a 4.8% (=$\frac{0.008 \times 0.126}{0.021}$) increase in post-filing volatility (computed using columns (1), (3), and (5), respectively).[40]

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

VI Analysis of Theme-Specific Summaries

Our primary test relies on the analysis of the *overall sentiment* conveyed by the document, which is admittedly a single dimension along which the information content can be evaluated (though we also explore the uncertainty dimension in the Online Appendix). In this section, we provide preliminary evidence on the usefulness of generative AI in the extraction of more specific dimensions of the information contained in the original document. Specifically, we design a query that gives a set of instructions to a machine to prepare summaries specifically related to financial and non-financial performance. We then examine their informativeness.

filing volatility (i.e., (post - pre)/pre) and use it as our dependent variable. In the same set of tests, we also use five-day abnormal bid-ask spreads and 29-day volatility, measured from the filing (conference call) dates to 28 days after the filing (call), as alternative outcome variables. Throughout these additional robustness tests, we find consistent results.

[40] In addition to the tabulated proxies, we also 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:

$$IPT_{[0,5]} = \sum_{i=0}^{4} \left(\frac{Abn\_Ret_{[0,i]}}{Abn\_Ret_{[0,5]}} \right) + \frac{1}{2} \tag{11}$$

where $Abn\_Ret_{[0,i]}$ denotes cumulative abnormal returns from day 0 to day i. We use market-adjusted abnormal returns in the calculation of $Abn\_Ret_{[0,i]}$. Intuitively, a higher *IPT* indicates faster price discovery after a release of certain information. In untabulated tests, we find that *Bloat* is negatively associated with IPT, implying that high *Bloat* is associated with slower price discovery, consistent with Cohen et al. (2020).

A Construction of Theme-Specific Summaries

Our targeted summaries aim to distinguish between financial and ESG performance. This analysis focuses on conference call transcripts because ESG discussions are not as common within MD&As, whereas conference call transcripts typically have a broader scope and feature discussions of environmental and social issues (Hassan et al., 2019; Sautner et al., 2023).

We use the unconstrained summaries performed by GPT as a starting point and further instruct the model to extract information about specific aspects of a firm's disclosed performance.[41] Our prompts ask the model to summarize information about ESG or financial performance without any additional explanations of these concepts. Although a more complex prompt might yield even stronger results, we aim to establish a baseline effect for targeted summaries. We discuss additional implementation details in Appendix E.

To assess the face validity of the summaries, Appendix F provides several snippets taken from ESG- and financial-performance-related summaries. In the case of ESG, we observe that firms discuss greenhouse gas emissions, environmental sustainability, renewable energy, etc. In the case of financial performance, the discussion is mainly related to operations, earnings, cash flows, fluctuations in revenue, etc. Based on these examples, the contents of theme-specific summaries seem to be aligned with our objectives.

B Attributes of Financial- vs. ESG-related Summaries

We start by examining time trends in performance summary attributes. Following a sharp rise in ESG-related attention among investors, we expect the proportion of summaries that contain ESG-related information to increase over time. In contrast, because financial performance constitutes the basis of firm valuation, we expect financial performance to be featured in most summaries.

Table 8, Panel A, provides summary statistics illustrating the information content of the financial and ESG-related summaries. *%ESG* represents the percentage of non-empty ESG-

[41] We perform summaries on summaries to economize on estimation time and cost, both of which are proportional to the number of summaries and number of input tokens. The original summaries are optimized to contain the most relevant information only. Thus, if the algorithm considers ESG or financial performance as less important or noisy, it should have discarded it in the initial summarization step. Therefore, besides relaxing the processing constraints of GPT, extracting theme-specific information from the pre-processed summaries guarantees that we identify highly important information only.

related summaries (an empty summary is generated when GPT determines there is no relevant information to summarize), whereas *%Fin* represents the percentage of non-empty financial-performance-related summaries. Recall that we perform summaries on summaries, which ensures that only the most salient information is retained. *lenESG* and *lenFin* are the length of ESG-performance-related and financial-performance-related summaries, respectively, scaled by the length of the original summarized document. We find that summaries of financial performance are almost always non-empty. At the same time, the frequency of non-empty ESG summaries varies between 15% and 38% over time. Only 20.37% of the original summaries contained ESG-related information in 2009. This number gradually increased to 38.20% in 2020. Similarly, the length of ESG summaries has increased gradually from 1.29% of the original summary in 2009 to 5.00% of the original summary in 2020. In contrast, the length of financial summaries does not exhibit a systematic trend.

We visualize these results in Figure 6(a). The left-hand-side depicts the time trend for *%ESG* and *lenESG* and the right-hand-side depicts *%Fin* and *lenFin*. One can observe a positive time trend in ESG-related summary characteristics, while this is not the case for financial summaries.

Informativeness. To examine the informativeness of theme-specific summaries to investors, we use them to generate ESG-performance-related sentiment ($Sentiment^{ESG}$) and financial-performance-related sentiment ($Sentiment^{Fin}$), using the same definition of sentiment as in Section IV.A. We then run the following ordinary least squares regression by year:

$$Abn\_Ret_{[0,1]it} = \beta_1 Sentiment^{j}_{it} + \gamma \mathbf{X}_{it} + \delta_t + \theta_j + \varepsilon_{it} \tag{12}$$

where $Abn\_Ret_{[0,1]it}$ is the cumulative abnormal return around the conference call date, $Sentiment^{j}_{it}$ is either $Sentiment^{ESG}_{it}$ (ESG sentiment) or $Sentiment^{Fin}_{it}$ (financial sentiment), $\mathbf{X}_{it}$ is the same set of firm-level controls as used in Table 4, 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*-values for $Sentiment^{ESG}$.[42] From 2009 to 2014, we find insignificant coefficients with the

[42]In this section only, we report *t*-values instead of standard errors. This change is implemented to facilitate the

exception of 2012 (coefficient = 0.035, t-value = 2.99). Several coefficients are even negative (although statistically indistinguishable from zero). However, starting in 2015, the coefficients related to ESG sentiment turn positive and gradually increase in both magnitude and statistical significance. In 2020, the coefficient reaches 0.070 with a t-value of 4.06. At the bottom of the panel, we report the estimates from the pooled (full sample) estimation. We observe a positive coefficient of 0.022 (t-value = 3.49) 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 2009 to 2020.[43]

Columns (3) and (4) report the estimates and t-values for $Sentiment^{Fin}$. Unlike for ESG sentiment, we observe positive and statistically significant coefficients on $Sentiment^{Fin}$ throughout the sample period (one exception is 2009, where the coefficient loses statistical significance but remains economically important). In the full sample analysis, we also find a positive and statistically significant coefficient (0.026) on $Sentiment^{Fin}$ (t-value = 10.41). The t-statistic is higher than that of $Sentiment^{ESG}$. Interestingly, we also observe a positive and statistically significant time trend (0.209 with a t-value of 2.94) for financial sentiment, although it is not as pronounced as for ESG sentiment.

Figure 6(b) helps to visualize these findings. The left-hand side of the figure shows a time trend in the coefficients over time. The solid line, which represents yearly coefficients on $Sentiment^{ESG}$, shows a strong increasing time trend. Similarly, the dotted line, which represents yearly coefficients on $Sentiment^{Fin}$, also shows some increasing time trend yet not as strong as in the ESG sentiment. On the right-hand-side, we observe that the t-values related to $Sentiment^{ESG}$ and $Sentiment^{Fin}$ exhibit similar increasing time trends.

Overall, we find that GPT's targeted summaries successfully extract specific dimensions of information relevant to a broader stakeholder base. Thus, generative AI shows promise in providing standardized insights on important firm-level topics.

interpretation of our t-value time trend analysis.

[43] We estimate the following model: $t\text{-}value_t = \gamma_0 + \gamma_1 Year_t + \epsilon_t$ and report $\hat{\gamma}_1$ accompanied by its heteroskedasticity-robust t-statistics.

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 70-75%, on average. Importantly, the obtained summaries appear to provide the most relevant insights as compared to the original documents. Specifically, we show that the summary-based sentiment better explains stock market reactions to disclosed information than the original's sentiment. Building on this insight, we construct a novel and easy-to-implement measure of the degree of "bloat" in corporate 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 higher price efficiency and higher information asymmetry, thus implying negative capital market consequences. Finally, we show that GPT summaries are useful to investors interested in targeting specific dimensions of communicated information, such as firms' ESG 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 in a timely manner (e.g. Blankespoor et al., 2020; Cohen et al., 2020). Our findings suggest that generative language models show promise in dealing with this information overload. Summaries generated by GPT are significantly shorter while they retain and amplify the main takeaways. 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 infrastructure to provide on-demand summaries to facilitate the informational efficiency of capital markets.[44] This presents an alternative to a two-tier reporting system featuring financial summaries that the SEC considered in the 1980s and 1990s to deal with informational overload (SEC, 1995; FASB, 1995; Bushman et al., 1996).

Finally, by relying on recent advancements in generative AI, we develop a simple and intuitive measure of the degree to which textual information contains redundancies and irrelevant or

[44]This information provision by an intermediary is different than the voluntary guidance provided by managers themselves (e.g. Guay et al., 2016; Dyer et al., 2016)

excessive details. Due to its straightforward nature, our methodology can be easily implemented for any other type of corporate communication (e.g., press releases, job postings, and websites). Additionally, as textual data has become ubiquitous in many disciplines (e.g. Gentzkow et al., 2019), our approach can prove useful in non-corporate settings (e.g., news outlets).

In closing, we note that our results are subject to two caveats. First, generative output may contain inaccuracies or nonsensical information ("hallucinations"). This issue has been raised in various generative AI applications and generally arises due to the probabilistic nature of the model (e.g. de Kok, 2023). Compared to other tasks, however, summarization relies on the specific context (e.g., the MD&A section) and, therefore, is less subject to this concern (e.g. Pu et al., 2023). Nevertheless, while the summaries are informative based on statistical criteria, information users should view summaries only as input and need to verify critical information before making decisions.[45] Second, while our results suggest that GPT produces high-quality summaries, it is unclear whether less sophisticated investors will actually use the information provided in the summaries. Drawing on the findings of Blankespoor et al. (2019), unsophisticated investors might anchor on the wrong information due to behavioral biases even when provided with relevant information. In addition to tailoring summaries for less sophisticated investors, the impact of GPT summaries on users' information uptake can be directly assessed in lab and field experiments, which we leave to future research.

[45] A convenient way to address this issue is to instruct GPT to provide references and hyperlinks for key points in the summary.

References

Armstrong, D.M., 2023. Measuring tax enforcement with generative AI. UNC Research Note.

Bae, J., Yu Hung, C., van Lent, L., 2023. Mobilizing text as data. European Accounting Review 5, 1–22.

Bai, J.J., Boyson, N.M., Cao, Y., Liu, M., Wan, C., 2023. Executives vs. chatbots: Unmasking insights through human-AI differences in earnings conference Q&A. Boston College Research Paper.

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

Bertomeu, J., Cheynel, E., Floyd, E., Pan, W., 2021. Using machine learning to detect misstatements. Review of Accounting Studies 26, 468–519.

Bertomeu, J., Lin, Y., Liu, Y., Ni, Z., 2023. Capital market consequences of generative AI: Early evidence from the ban of ChatGPT in italy. Washington St. Louis Working Paper.

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

Blankespoor, E., 2019. The impact of information processing costs on firm disclosure choice: Evidence from the XBRL mandate. Journal of Accounting Research 57, 919–967.

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.

Blankespoor, E., Dehaan, E., Wertz, J., Zhu, C., 2019. Why do individual investors disregard accounting information? the roles of information awareness and acquisition costs. Journal of Accounting Research 57, 53–84.

Blankespoor, E., Hendricks, B.E., Miller, G.S., 2023. The pitch: Managers' disclosure choice during initial public offering roadshows. The Accounting Review 98, 1–29.

Bochkay, K., Brown, S.V., Leone, A.J., Tucker, J.W., 2023. Textual analysis in accounting: What's next? Contemporary accounting research 40, 765–805.

Bonsall, S.B.I., 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.

Breuer, M., deHaan, E., 2023. Using and interpreting fixed effects models. Stanford Working Paper.

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

Brown, S.V., Tucker, J.W., 2011. Large-sample evidence on firms' year-over-year md&a modifications. Journal of Accounting Research 49, 309–346.

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.

Bushman, R.M., Gigler, F., Indjejikian, R.J., 1996. A model of two-tiered financial reporting. Journal of Accounting Research 34, 51–74.

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.

Cao, S., Jiang, W., Yang, B., Zhang, A.L., 2023. How to Talk When a Machine Is Listening: Corporate Disclosure in the Age of AI. The Review of Financial Studies 36, 3603–3642.

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.
Chen, X., Cho, Y.H., Dou, Y., Lev, B., 2022. Predicting future earnings changes using machine learning and detailed financial data. Journal of Accounting Research 60, 467–515.
Chen, Y., Kelly, B.T., Xiu, D., 2023. Expected returns and large language models. Chicago Booth Research Paper.
Choi, G.Y., Kim, A.G., 2023. Economic footprints of tax audits: A generative ai-driven approach. Chicago Booth Research Paper .
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.
Costello, A., Levy, B., Nikolaev, V., 2023. Uncovering information. Working Paper.
Costello, A.M., Down, A.K., Mehta, M.N., 2020. Machine+ man: A field experiment on the role of discretion in augmenting ai-based lending models. Journal of Accounting and Economics 70, 101360.
Doherty, K., Marques, F., 2023. Citadel negotiating enterprise-wide ChatGPT license, Griffin says. Bloomberg .
Dyer, T., Lang, M., Stice-Lawrence, L., 2016. Do managers really guide through the fog? on the challenges in assessing the causes of voluntary disclosure. Journal of Accounting and Economics 62, 270–276.
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.
Eisfeldt, A.L., Schubert, G., Zhang, M.B., 2023. Generative AI and firm values. NBER Working Paper.
FASB, 1995. Prospectus: Disclosure Effectiveness. FASB, Stamford, Connecticut.
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.
Goldstein, I., Yang, S., Zuo, L., 2023. The real effects of modern information technologies: Evidence from the EDGAR implementation. Journal of Accounting Research Forthcoming.
Goyal, T., Li, J.J., Durrett, G., 2022. News summarization and evaluation in the era of GPT-3. arXiv preprint arXiv:2209.12356 .
Gu, S., Kelly, B., Xiu, D., 2020. Empirical asset pricing via machine learning. The Review of Financial Studies 33, 2223–2273.
Guay, W., Samuels, D., Taylor, D., 2016. Guiding through the fog: Financial statement complexity and voluntary disclosure. Journal of Accounting and Economics 62, 234–269.
Gunning, R., 1952. Technique of clear writing. McGraw-Hill, New York.
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.

Henry, E., Leone, A.J., 2016. Measuring qualitative information in capital markets research: Comparison of alternative methodologies to measure disclosure tone. The Accounting Review 91, 153–178.

Hirshleifer, D., Kewei Hou, Teoh, S.H., Yinglei Zhang, 2004. Do investors overvalue firms with bloated balance sheets? Journal of Accounting and Economics 38, 297–331.

Hoitash, R., Hoitash, U., 2018. Measuring Accounting Reporting Complexity with XBRL. The Accounting Review 93, 259–287.

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

Jha, M., Qian, J., Weber, M., Yang, B., 2023. ChatGPT and corporate policies. Chicago Booth Research Paper.

Kim, A.G., Muhn, M., Nikolaev, V.V., 2023. Can generative AI uncover corporate risks? Fama-Miller Center Research Paper.

Kim, A.G., Nikolaev, V.V., 2023. Context-based interpretation of financial information. University of Chicago Working Paper.

Kim, C., Wang, K., Zhang, L., 2019. Readability of 10-k reports and stock price crash risk. Contemporary accounting research 36, 1184–1216.

de Kok, T., 2023. Generative LLMs and textual analysis in accounting:(chat)GPT as a research assistant? University of Washington 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.

Lang, M., Lins, K.V., Maffett, M., 2012. Transparency, liquidity, and valuation: International evidence on when transparency matters most. Journal of Accounting Research 50, 729–774.

Lang, M., Stice-Lawrence, L., 2015. Textual analysis and international financial reporting: Large sample evidence. Journal of Accounting and Economics 60, 110–135.

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 Supplement, 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.

Li, Q., Shan, H., Tang, Y., Yao, V., 2023. Corporate climate risk: Measurements and responses. Review of Financial Studies Forthcoming.

Lopez-Lira, A., Tang, Y., 2023. Can ChatGPT forecast stock price movements? return predictability and large language models. ArXiv preprint arXiv:2304.07619.

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., 2023. Measuring firm complexity. Journal of Financial and Quantitative Analysis Forthcoming.

Lu, J., 2022. Limited attention: Implications for financial reporting. Journal of Accounting Re-

search 60, 1991–2027.

Peterson, K., Schmardebeck, R., Wilks, T.J., 2015. The Earnings Quality and Information Processing Effects of Accounting Consistency. The Accounting Review 90, 2483–2514.

Pu, X., Gao, M., Wan, X., 2023. Summarization is (almost) dead. ArXiv preprint arXiv:2309.09558.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al., 2018. Improving language understanding by generative pre-training. arXiv preprint arXiv:2302.08081 .

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. Springer. Natural Language Processing Using Very Large Corpora, pp. 157–176.

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

SEC, 1995. Use of Abbreviated Financial Statements in Documents Delivered to Investors Pursuant to the Securities Act of 1933 and Securities Exchange Act of 1934.

SEC, 2013. Report on review of disclosure requirements in regulation S-K.

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.

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

Appendix A. Variable Descriptions

Name	Description
Bloat	The difference between the length of the original document and summarized documents, 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 provided by Loughran and McDonald (2011).
$Sentiment^{Sum}$	Textual sentiment of a summarized document calculated using financial keyword dictionaries provided by Loughran and McDonald (2011).
Fog	Fog readability index calculated based on Gunning (1952), adjusted for financial terms, with higher values indicating less readable text.
Plain_Eng	Plain English readability measure based on the methodology in Loughran and McDonald (2014b) with higher values indicating less readable text. Specifically, we add the following components: (i) the average number of words per sentence, (ii) the average number of characters per word, (iii) the number of passive voice verbs scaled by the document length, (iv) number of pronouns scaled by the document length, and (v) number of the word "respectively" scaled by the document length.
Complexity	A measure of firm complexity as used in Loughran and McDonald (2023).
$Abn\_Ret_{[0,1]}$	Market adjusted cumulative abnormal returns accumulated over two days starting from the filing date to a day after the filing date.
PIN	Probability of informed trading 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.
Abn_Spread	The difference in the bid-ask spread on the disclosure date and the average bid ask spread calculated over a two-week window preceding the disclosure date.
Post_Vol	The root mean squared error of residuals (abnormal retunrs) from the market model (Loughran and McDonald, 2014a). We estimate the RMSE using daily returns data starting from day 6 and ending on day 28 following the disclosure date.
Log_ME	The natural logarithm of the market capitalization.
Log_BE_ME	The natural logarithm of the book-to-market ratio.
N_Analyst	The number of analysts following the firm.
Inst_Own	Institutional ownership based on 13-F filings. Missing data is set to zero.
Earn_Vol	Earnings volatility calculated as the standard deviation of earnings (net income) scaled by averaged total assets. Earnings are measured annually over the preceding five years. For quarterly data, we use earnings for the same quarter in preceding five years.
Loss	An indicator variable that equals one when a firm reports negative earnings.
ROA	Return (net income) scaled by total assets.
One_Day_Ret	Raw return on the filing date (conference call date).
Price	The price of a stock measured as of the fiscal period end date.
SUE	Earnings surprise calculated as the difference between reported earnings-per-share and the analyst consensus forecast from the end of the prior quarter, scaled by the stock price as of the fiscal period end date. We require at least three distinct analyst estimates.
Abs_SUE	Absolute value of earnings surprise, *SUE*.
Friday	An indicator variable that equals one when the reporting date is Friday.
%Fin	The percentage of summaries that contain financial-performance-related information.
%ESG	The percentage of summaries that contain ESG-performance-related information.
lenFin	Length of the financial summary scaled by the total length of the summary.
lenESG	Length of ESG summary scaled by the total length of the summary.
$Sentiment^{Fin}$	Sentiment of financial-performance-related summaries.
$Sentiment^{ESG}$	Sentiment of ESG-performance-related summaries.

Appendix B. 2022-Q4 Disney Transcript

This figure depicts the process of summarizing conference call transcripts with GPT, specifically using Disney's 2022-Q4 earnings call as a case study. Panel A shows the initial segment of Bob Chapek's presentation. It has the same upbeat messaging, but retains the information about operating losses in the summary. Panel B then shows the summarization of the more "bloated" second part of his speech. As highlighted with the red box, more than two paragraphs about extraneous celebratory events are absorbed into a single short sentence.

Panel A. Beginning of the CEO's Presentation

[...] Thank you, Alexia and good afternoon everyone. Fiscal 2022 was a strong year for our company as we continued our journey of telling incredible Disney stories, utilizing groundbreaking technology in order to further develop our brands and franchises while customizing and personalizing experiences to make magical memories that last a lifetime. Those efforts resulted in truly phenomenal storytelling, record annual results at our Parks, Experiences and Products segment, and outstanding growth at our Direct-to-Consumer services, which added nearly 57 million subscriptions this year to reach a total of more than 235 million.

We are particularly pleased with growth in the fourth quarter, which saw the addition of 14.6 million subscriptions across our suite of services including 12 million Disney+ subscriptions, over 9 million of which were core Disney+. It has taken just 3 short years for Disney+ to transform from a nascent business to an industry leader. That transformation is the direct result of the strategic decision we made at launch to heavily invest in our Direct-to-Consumer offering, a decision made knowing that achieving rapid growth would result in short-term losses.

Building a streaming powerhouse has required significant investment. And now with its scale, incredible content pipeline and global reach, Disney+ is well situated to leverage our position for long-term profitability and success. Our financial results this quarter represent a turning point as we reached peak DTC operating losses, which we expect to decline going forward. That expectation is based on 3 factors.

First, the benefit of both price increases and the launch of a Disney+ ad tier next month. Second, a realignment of our cost, including meaningful rationalization of our marketing spend. And third, leveraging our learnings and experience in direct-to-consumer to optimize our content slate and distribution approach to deliver a steady state of high-impact releases that efficiently drive engagement and subscriber acquisition.

With these factors, we believe we are on a path to profitable streaming business that generates shareholder value long into the future. And assuming we do not see a meaningful shift in the economic climate, we still expect Disney+ to achieve profitability in fiscal 2024 as losses begin to shrink in the first quarter of fiscal 2023. Christine will go into more detail on the drivers of our direct-to-consumer operating performance and provide more insight into our expectations going forward and some of our key assumptions. But first, I'd like to share a few highlights from the quarter. [...]

[...] One of the noteworthy highlights of the fiscal year 2022 was the outstanding performance of the Parks, Experiences, and Products division. Nevertheless, the most remarkable achievement was the substantial growth in Disney's Direct-to-Consumer services. The company reported a significant increase of almost 57 million subscriptions, pushing the total to over 235 million. This surge in subscriptions, including a notable gain of 14.6 million in the fourth quarter, with 12 million attributed to Disney+, can be attributed to their strategic investments in the Direct-to-Consumer segment. Achieving such substantial growth did require some short-term sacrifices, but Disney remained committed. They made substantial investments in Disney+ to transform it into a formidable streaming platform with an extensive content library and global accessibility. Despite recent operating losses for Disney+, there is optimism that these losses will decrease due to price adjustments, the introduction of ad-supported tiers, cost optimization, and strategic content releases. However, fiscal 2022 wasn't solely about streaming success. [...]

Panel B. Continuation of the CEO's Presentation

[...] Q4 was another strong period for Parks, Experiences and Products, which continue to deliver phenomenal results despite the impact of Hurricane Ian. I want to thank the amazing cast members in Orlando who went above and beyond to help keep our guests safe and entertained during the storm. From protecting the many animals at Disney Animal Kingdom to packing thousands of meal kits, to donating and delivering emergency supplies to the community, I am so proud of how our team came together to support our guests, our neighbors and each other.

Our Parks team is laser focused on enhancing the guest experience in creating those magical memories I mentioned at the outset. This focus, along with the investments we made to bringing new attractions and experiences to our domestic parks, is generating consistently strong demand, which, on many days, exceeds our current capacity. And we continue to manage attendance levels with a focus on providing guests with the highest quality experience in enhancing our Park's overall financial performance.

One of the things our guests love most is the opportunity to celebrate at our parks, as evidenced by the post-pandemic return and sell-out of special ticketed events like Boogie Boogie Bash and Mickey's Not So Scary Halloween party. I visited Disneyland with my family just before Halloween and the celebration was phenomenal. Tickets for Mickey's Very Merry Christmas party at Walt Disney World has now officially gone on sale and over half of all dates have already sold out.

As you know, we are about to embark on the company's 100th anniversary celebration. The fun kicks off in our parks on January 27 at Disneyland. where we'll unveil new platinum infused decor, Premier II, all new nighttime spectaculars and opened a highly anticipated Mickey & Minnie's Runaway Railway attraction. This is only one part of what will be the largest cross-company celebration in Disney's history with activations around the world, and we're so excited for fans and families to join us.

At our international parks, Disneyland Paris is enjoying a great resurgence. Our fantastic new Marvel Avengers Campus opened on July 20, and guests love a highly immersive and dynamic environment of the first-ever Marvel-themed land in Europe. Prior to the recent closure of Shanghai Disney Resort, we are seeing positive momentum there and at Hong Kong Disneyland. We are hopeful that the situation will improve and are thinking of all of our employees there as we manage through the challenging COVID environment. Our Disney Cruise line is showing strong signs of recovery. The new Disney Wish is in high demand, and we've seen a ramp-up in bookings for our base fleet. [...]

[...] Despite the challenges posed by Hurricane Ian, the Parks, Experiences, and Products segment exhibited resilience. The Orlando team demonstrated remarkable dedication by ensuring guest safety, protecting animals, and providing community assistance during the storm. Their focus on enhancing guest experiences remained a top priority, leading to investments in new attractions and experiences, even as they managed overwhelming demand. Celebratory events also gained immense popularity. Disney's international ventures also thrived. Disneyland Paris benefited from the Marvel Avengers Campus, and the Asian parks performed well before unexpected closures. Disney Cruise Line, especially the new Disney Wish, experienced high demand. [...]

Appendix C. Excised Content

Panel A compares the degree of boilerplate language usage between the raw and summarized documents. Cosine Similarity$_{t,\,t+1}$ denotes the cosine similarity scores of firm i's year t and $t+1$ filings (conference call transcripts), averaged over all is and ts. Boilerplate % is the average percentage of words included in boilerplate sentences as defined in Dyer et al. (2017). ***, **, and * denote statistical significance at 1%, 5%, and 10% level, respectively. Panel B illustrates why some contents are deleted from the raw documents to generate summaries. We obtain these percentages by instructing GPT to classify the reasons for differences between the raw document and its summary (see Online Appendix A for details). Panel B1 shows the reasons for deletion in MD&As and Panel B2 shows the reasons for deletion in conference calls.

Panel A. Boilerplate			
A1. MD&As			
Metrics	Raw Document	Summarized Document	Difference
Cosine Similarity$_{t,t+1}$	85.94	73.51	12.43***
Boilerplate %	5.45%	2.66%	2.79%***
A2. Conference Calls			
Metrics	Raw Document	Summarized Document	Difference
Cosine Similarity$_{t,t+1}$	79.45	70.75	8.70***
Boilerplate %	1.12%	1.02%	0.10%
Panel B. Deleted Reasons			
B1. MD&As			
Reason			Count
Excessive Details			16.80%
Repeated Statements			18.24%
General Market Overview			3.15%
Boilerplate Languages			32.50%
Historical Recaps			15.20%
Other Reasons			14.11%
Total			100.00%
B2. Conference Calls			
Reason			Count
Excessive Details			39.80%
Repeated Statements			20.90%
General Market Overview			5.47%
Boilerplate Languages			15.92%
Historical Recaps			9.95%
Other Reasons			7.96%
Total			100.00%

Appendix D. Out-of-Sample Tests

This table reports the association between the textual sentiment and two-day cumulative abnormal returns for samples in calendar year 2022. In columns (1) and (2), we use random samples chosen from the MD&A disclosures. In columns (3) and (4), we use random samples chosen from conference call transcripts. As control variables, we include *Log_ME*, *Log_BE_ME* , *Inst_Own*, and *SUE*. Industry definition is based on two-digit SIC codes. Standard errors are clustered at the industry level and are reported within parentheses. ***, **, 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 = $Abn\_Ret_{[0,1]}$				
	MD&A		Conference Call	
	Raw	Summarized	Raw	Summarized
	(1)	(2)	(3)	(4)
$Sentiment^{Raw}$	0.011		0.073***	
	(0.023)		(0.020)	
$Sentiment^{Sum}$		0.082***		0.144***
		(0.027)		(0.018)
Controls	Yes	Yes	Yes	Yes
Time FE	No	No	Yes	Yes
Firm FE	No	No	No	No
Industry FE	Yes	Yes	Yes	Yes
Cluster	Ind	Ind	Ind	Ind
N	221	221	552	552
Adjusted R^2	0.008	0.064	0.124	0.216

Appendix E. Theme-Specific Summaries

For our query-based summaries, we add a prompt to prevent the model from including other topics if it is not able to identify the requested information (e.g., ESG-related information). We explicitly instruct the model to print out NA when there is no information available and provide a related example (one-shot training), which in our experience improves the quality of theme-specific summaries. The full prompts for the targeted summaries are provided in Online Appendix A. Note that these prompts do not specify the concept of ESG or financial performance. Thus, to validate GPT's understanding of ESG and financial performance concepts, we ask it to write a paragraph in a hypothetical conference call transcript: "Imagine that you are a CEO of a company. You are holding an earnings conference call. Write a paragraph of the 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 F. Sample Summaries

Panel A. Snippets from a sample summary (ESG-performance-specific)
[. . .] The company's success can be attributed to its focus on renewals and the introduction of innovative products, including mobile solutions and privacy technology. Norton has initiated its journey towards Digital Safety by introducing new offerings that focus on ESG, such as Wi-Fi privacy and Norton Core, a home IT security and next-generation parental control solution. [. . .] [. . .] The shift to FIFO is expected to provide greater transparency and accuracy in inventory valuation, which will enhance the company's ability to manage its supply chain and reduce waste. [. . .] [. . .]The technology delivery company is committed to improving its environmental capabilities to assist clients in achieving their emission reduction objectives. The company remained committed to advancing clean energy technologies to help their clients reduce their carbon footprint. The company is committed to reducing its greenhouse gas emissions and has implemented various measures to achieve this goal. These measures include facility consolidations, employee remote work options, and increased recycling efforts. [. . .] [. . .]The team behind the project has developed a new set of risk governance models and tools with the aim of assisting members in managing risks while also maximizing the business value of information. CEB researchers made a significant impact at the Society for Industrial Organizational Psychology event in Houston by leading 63 of the panels. This event provided a platform for the company to showcase its expertise in the field of organizational psychology. [. . .] [. . .] The Kroger Co. Foundation and the Zero Hunger — Zero Waste Foundation will each receive an equal share of the funds. This decision reflects the company's commitment to addressing issues such as hunger and waste, and its dedication to making a positive impact in the communities it serves. The company allocated $5 million to a fund aimed at promoting racial equity and justice. The fund is expected to support initiatives that address systemic discrimination and inequality in society. [. . .]
Panel B. Snippets from a sample summary (Financial-performance-specific)
[. . .] The company had a strong quarter with 17% year-over-year revenue growth to 47.7 million, non-GAAP gross margin of 70.2% and non-GAAP net income of $0.02 per share. There is no further information on financial performance. In Q2, the company's total revenue increased 17% year-over-year to $47.7 million, with product revenue increasing 14% year-over-year to $38.9 million and professional services and support revenue increasing 53% year-over-year to $8.5 million. [. . .] [. . .] The company had cost-reduction initiatives and lower input costs that helped offset the impact of reduced volumes. Net interest expense was down from last year due to lower borrowings and a drop in interest rates. The company expects to recover most of the additional cash taxes paid in the next two quarters in 2010 due to a recent tax law change allowing a five-year carryback of operating losses. [. . .] [. . .] The company expects the conditions seen in the first quarter to continue into the second quarter and will be taking some actions to reorganize some of its operations in the segments, with total costs to be incurred related to this to be in the area of $6 million. These charges will be recorded in the quarter in which they are recognizable for accounting purposes. Annual savings from these actions are projected to be in the area of $3 million pretax, with most of the savings expected to be realized in the beginning of next year. [. . .]

Figure 1. Visual Illustration of Disclosure Bloat

This figure illustrates the conceptual underpinning 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 the utility of an investor who processes a document of length k.

Figure 1. Visual Illustration of Disclosure Bloat

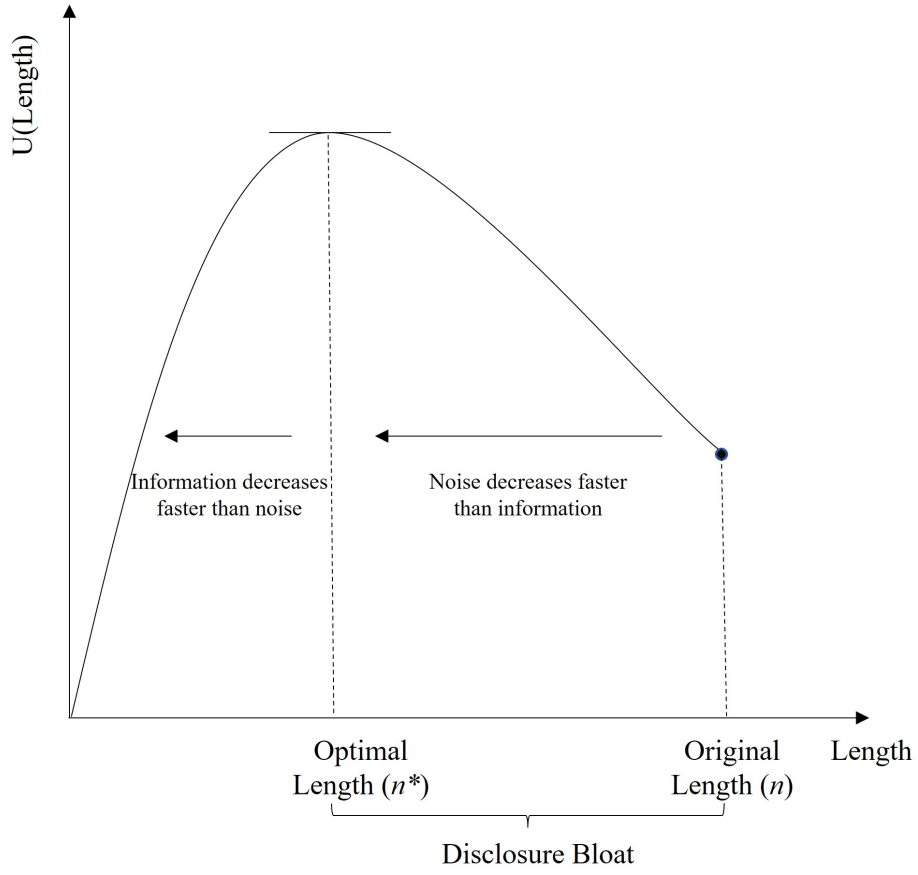

Figure 2. Disney's Disclosure Bloat over Time

This figure illustrates the time series of Disney's *Bloat* from 2015 until 2023-Q2. Figure 2(a) shows the disclosure bloat based on a representative draw of our summarizaiton algorithm. Figure 2(b) shows the same information, but also reports 95% confidence intervals based on 50 repetitions of our summarization algorithm.

Figure 2(a). Disney's Disclosure Bloat over Time

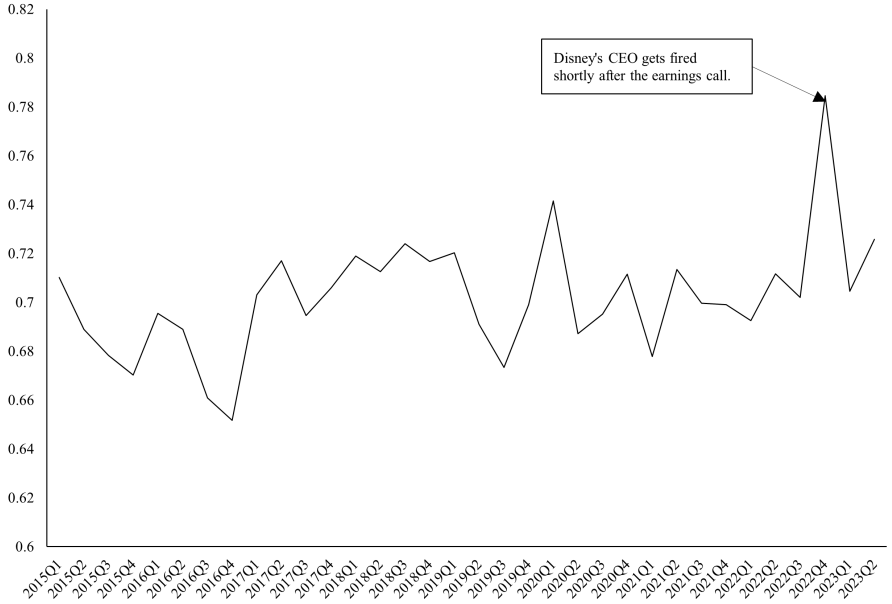

Figure 2(b). Disney's Disclosure Bloat over Time with 95% Confidence Intervals

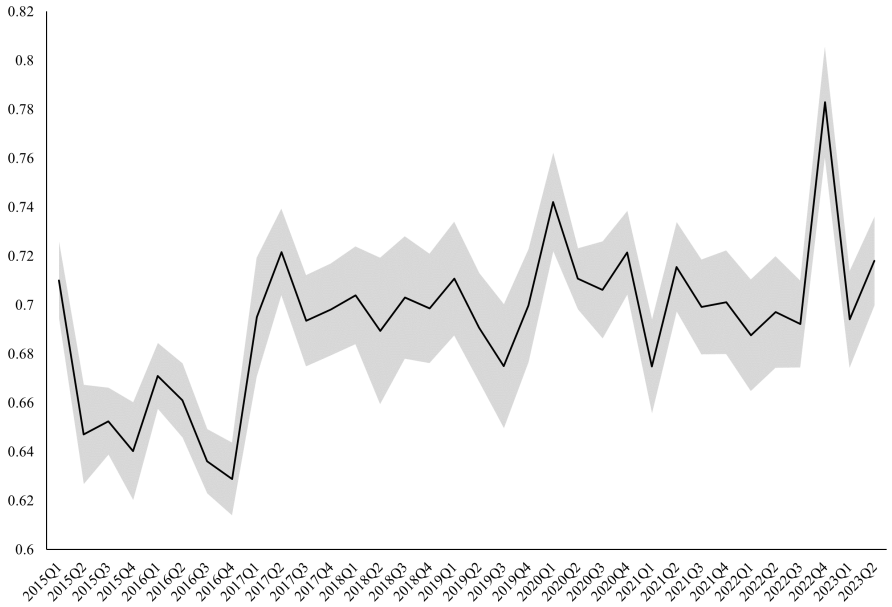

Figure 3. Changes in Average Length and Sentiment

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 (e.g., AboveMed length is the average length of the documents whose sentiment is larger than the median and AboveMed sentiment is the average sentiment of the documents whose sentiment is larger than the median). Bar charts represent the length and line graphs represent the sentiment of the documents. Figure 3(a) uses the MD&A sample and Figure 3(b) uses the conference call sample. Refer to Appendix A for variable descriptions.

Figure 3(a). Changes in Average Length and Sentiment (MD&A Sample)

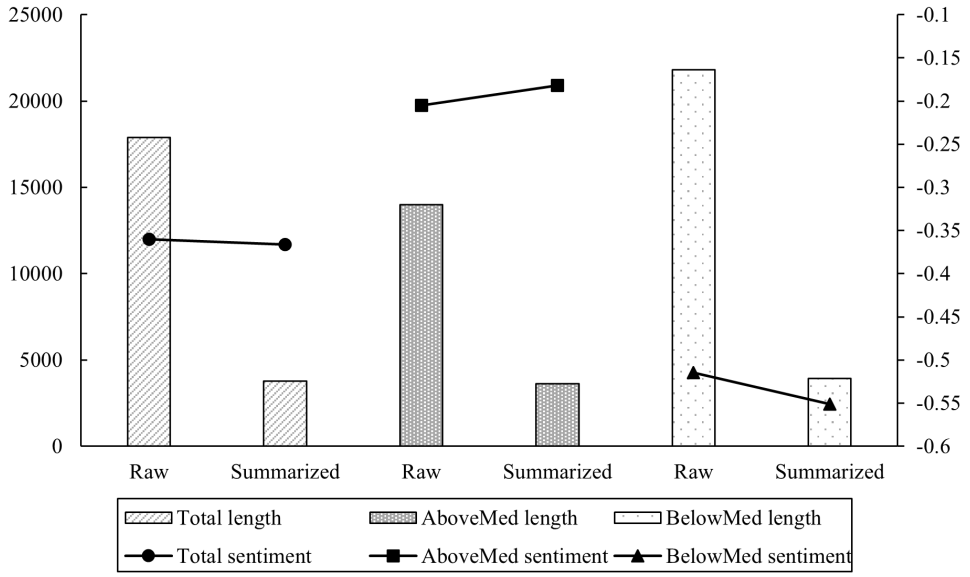

Figure 3(b). Changes in Average Length and Sentiment (Conference Call Sample)

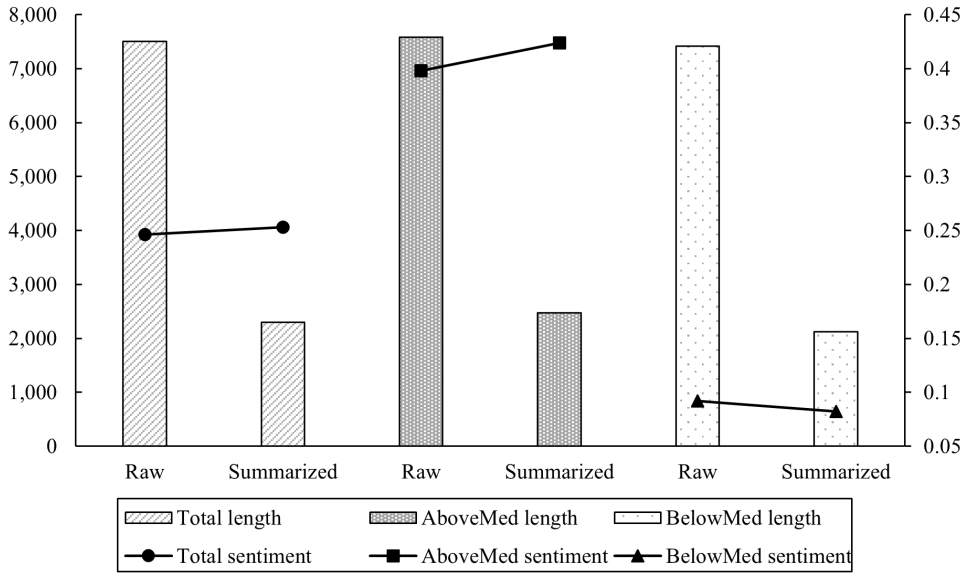

Figure 4. Abnormal Returns and Sentiment

This figure illustrates the averaged cumulative abnormal returns over two days (filing date and one day after) across sentiment quintiles. For each period, we partition the sample into 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 Figures represents raw sentiment, and the solid line represents the summary sentiment. Figure 4(a) uses the MD&A sample and Figure 4(b) uses the conference call sample. Refer to Appendix A for variable descriptions.

Figure 4(a). Abnormal Returns and Sentiment Quintiles (MD&A Sample)

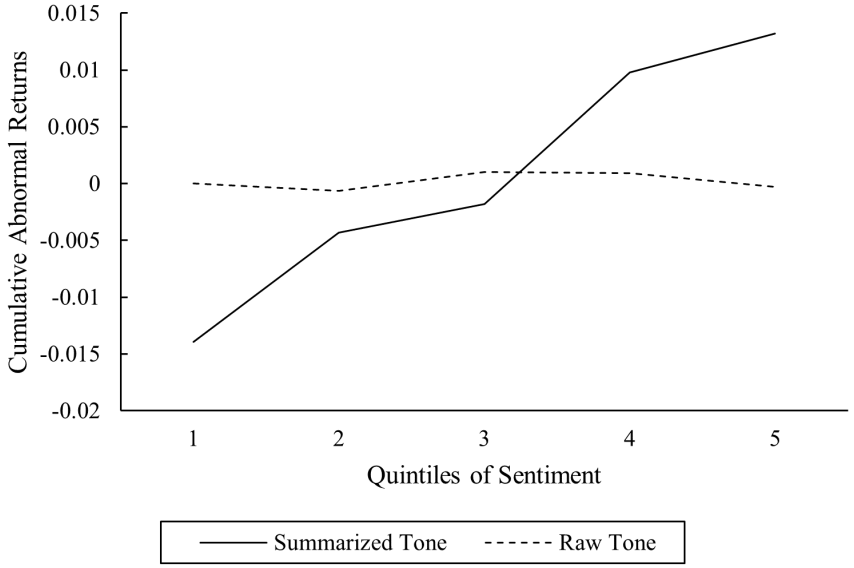

Figure 4(b). Abnormal Returns and Sentiment Quintiles (Conference Call Sample)

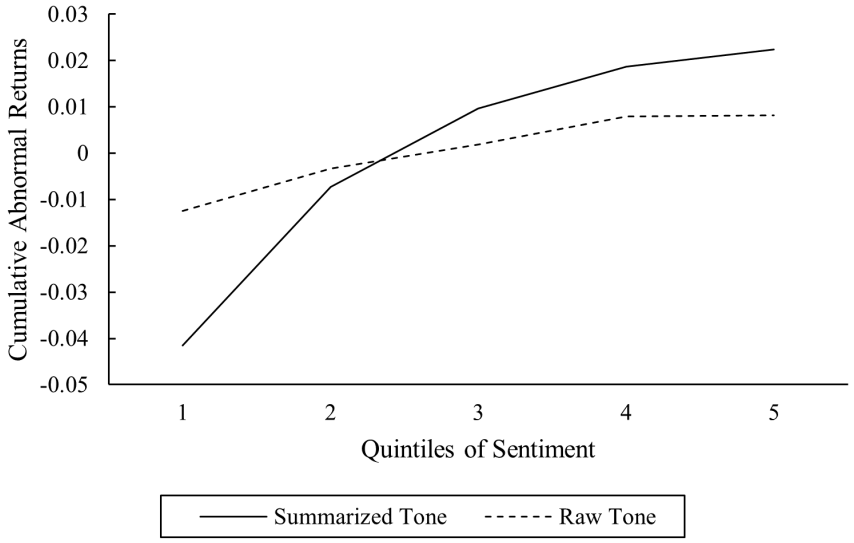

Figure 5. Average Length and Disclosure Bloat

Figures 5(a) and 5(b) plot the average annual length and *Bloat* for the MD&A and conference calls samples, respectively. Figure 5(a) uses the MD&A sample and Figure 5(b) relies on the conference call sample. Refer to Appendix A for variable descriptions.

Figure 5(a). Average Length and Disclosure Bloat (MD&A Sample)

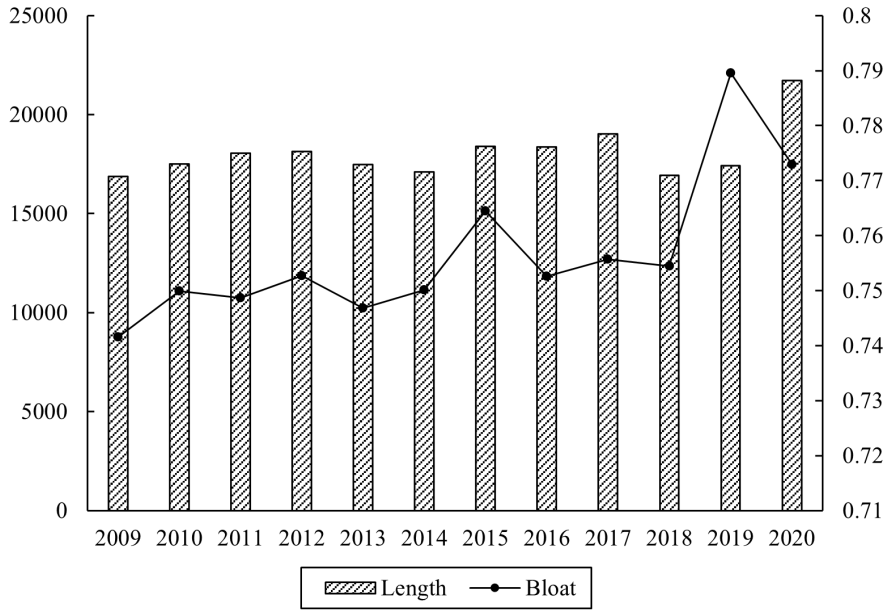

Figure 5(b). Average Length and Disclosure Bloat (Conference Call Sample)

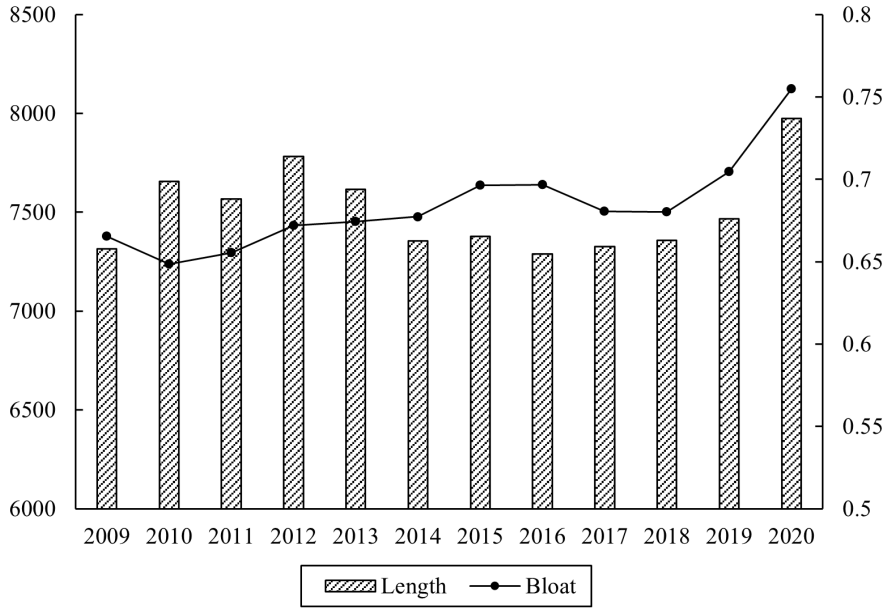

Figure 6. Theme-Specific Summaries

Figure 6(a) shows the time trend of ESG-related summaries (left) and financial-performance-related summaries (right). The left-hand-side depicts the time trend of *%ESG* and *lenESG* and the right-hand-side depicts *%Fin* and *lenFin*. The left-hand side of the Figure 6(b) shows the time trend in *t*-statistics over time. The solid line represents yearly *t*-values of $Sentiment^{ESG}$ and, in contrast, the dotted line represents yearly *t*-values of $Sentiment^{Fin}$. On the right-hand-side, we plot the coefficient values of both $Sentiment^{ESG}$ and $Sentiment^{Fin}$. Refer to Appendix A for detailed variable descriptions.

Figure 6(a). Time Trend (left: ESG, right: financial performance)

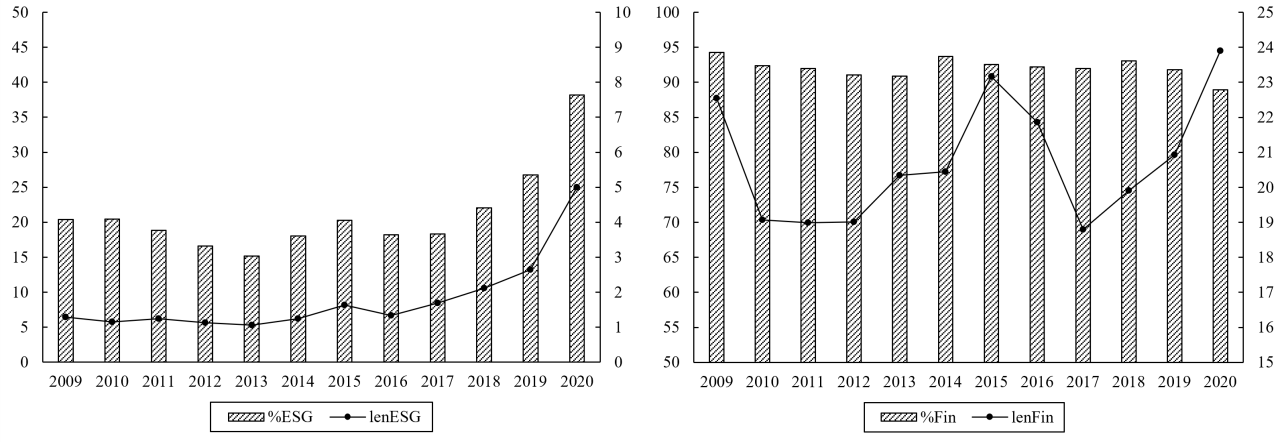

Figure 6(b). Time Trend of Informativeness (left: *t*-values, right: coefficient magnitudes)

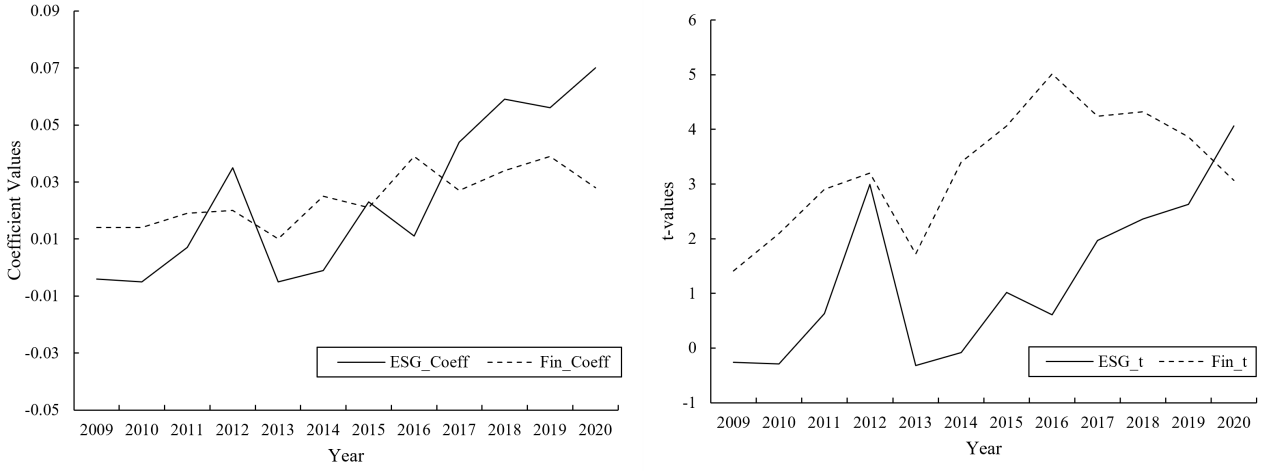

Table 1. Descriptive Statistics

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

Panel A. MD&A Sample

	Universe				Sample				
	N	Mean	Median	Std	N	Mean	Median	Std	Diff.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(2) – (6)
Log_ME	8,699	7.801	7.653	1.698	1,790	7.713	7.648	1.706	-0.088
									(-1.03)
Log_BE_ME	8,699	-1.106	-1.031	0.880	1,790	-1.100	-1.022	0.869	0.006
									(0.13)
N_Analyst	8,699	5.170	3.000	5.282	1,790	4.915	3.000	5.344	-0.255
									(-0.99)

Panel B. Conference Calls Sample

	Universe				Sample				
	N	Mean	Median	Std	N	Mean	Median	Std	Diff.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(2) – (6)
Log_ME	40,362	7.801	7.740	1.841	8,537	7.798	7.738	1.844	-0.003
									(-0.09)
Log_BE_ME	40,362	-1.088	-1.002	0.913	8,537	-1.087	-0.998	0.915	0.001
									(0.00)
N_Analyst	40,362	7.554	6.000	6.430	8,537	7.557	6.000	6.428	0.003
									(0.02)

Table 2. Linguistic Attributes of Summaries

This table presents the linguistic attributes of both raw and summarized documents. In the last column we report the difference between the mean values of the two groups with its statistical significance. In Panels A, we tabulate statistics for the MD&A sample. In Panels B, we repeat the analysis for 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

Levels:

		Raw Document			Summarized Document			
	N	Mean	Median	Std	Mean	Median	Std	Diff.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(5) – (2)
Length	1,790	17,901	14,254	13,151	3,779	3,433	1,882	-14,122***
Sentiment	1,790	-0.360	-0.371	0.202	-0.366	-0.438	0.316	-0.006
Fog	1,790	10.025	9.960	1.591	10.175	10.130	0.936	0.150***
Plain_Eng	1,790	0.289	0.286	0.039	0.290	0.290	0.011	0.000

Changes:

				Percentiles		
	N	Mean	Std	p25	p50	p75
Δ*Length*	1,790	-14,122	11,788	-16,729	-10,462	-7,093
Δ*Sentiment*	1,790	-0.006	0.188	-0.132	-0.039	0.110
Δ*Fog*	1,790	0.150	1.144	-0.600	0.190	0.920
Δ*Plain_Eng*	1,790	0.000	0.036	-0.020	0.004	0.026

Panel B. Conference Call Sample

Levels:

		Raw Document			Summarized Document			
	N	Mean	Median	Std	Mean	Median	Std	Diff.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(5) – (2)
Length	8,537	7,501	7,657	2,371	2,300	2,240	1,098	-5,201***
Sentiment	8,537	0.246	0.261	0.197	0.253	0.257	0.286	0.008***
Fog	8,537	8.970	8.950	0.852	10.177	10.040	0.952	1.207***
Plain_Eng	8,537	0.252	0.251	0.027	0.261	0.260	0.014	0.009***

Changes:

				Percentiles		
	N	Mean	Std	p25	p50	p75
Δ*Length*	8,537	-5,201	2,085	-6,533	-5,015	-3,637
Δ*Sentiment*	8,537	0.008	0.190	-0.107	0.011	0.118
Δ*Fog*	8,537	1.207	0.828	0.680	1.190	1.690
Δ*Plain_Eng*	8,537	0.009	0.026	-0.008	0.010	0.029

Table 3. Partitions Based on Raw Sentiment

This table reports the linguistic attributes of raw and summarized documents when we partition the sample based on the sentiment of the raw document. Specifically, we split the sample into two groups depending on the annual (quarterly) 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

Subsample 1: $Sentiment^{Raw} >$ Median

		Raw Document			Summarized Document			
	N	Mean	Median	Std	Mean	Median	Std	Diff.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(5) – (2)
Length	897	13,997	12,445	7,493	3,637	3,419	1,602	-10,360***
Sentiment	897	-0.205	-0.240	0.147	-0.182	-0.276	0.322	0.024***
Fog	897	9.690	9.540	1.591	10.310	10.310	0.914	0.620***
Plain_Eng	897	0.284	0.282	0.039	0.289	0.289	0.011	0.005***

Subsample 2: $Sentiment^{Raw} <$ Median

		Raw Document			Summarized Document			
	N	Mean	Median	Std	Mean	Median	Std	Diff.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(5) – (2)
Length	893	21,821	17,066	16,117	3,922	3,444	2,118	-17,899***
Sentiment	893	-0.515	-0.499	0.110	-0.551	-0.564	0.168	-0.036***
Fog	893	10.362	10.190	1.518	10.040	10.010	0.939	-0.322***
Plain_Eng	893	0.294	0.290	0.038	0.290	0.291	0.011	-0.004***

Panel B. Conference Call Sample

Subsample 1: $Sentiment^{Raw} >$ Median

		Raw Document			Summarized Document			
	N	Mean	Median	Std	Mean	Median	Std	Diff.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(5) – (2)
Length	4,281	7,579	7,799	2,388	2,474	2,423	1,062	-5,105***
Sentiment	4,281	0.398	0.386	0.107	0.424	0.448	0.208	0.025***
Fog	4,281	8.701	8.670	0.819	10.163	10.040	0.874	1.462***
Plain_Eng	4,281	0.235	0.234	0.022	0.261	0.260	0.013	0.025***

Subsample 2: $Sentiment^{Raw} <$ Median

		Raw Document			Summarized Document			
	N	Mean	Median	Std	Mean	Median	Std	Diff.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(5) – (2)
Length	4,256	7,422	7,515	2,351	2,125	2,030	1,106	-5,297***
Sentiment	4,256	0.092	0.117	0.139	0.082	0.091	0.249	-0.010***
Fog	4,256	9.241	9.220	0.797	10.192	10.050	1.024	0.951***
Plain_Eng	4,256	0.268	0.267	0.022	0.261	0.260	0.015	-0.007***

Table 4. Informativeness of Summarized Documents

This table reports OLS regressions of two-day cumulative abnormal returns on the sentiment proxies. In columns (1) - (4), we use the sentiment calculated from raw documents. In columns (5) - (8), we use the sentiment calculated based on the summaries. In columns (3) and (7), we use a sub-sample with sentiment larger than the median value (denoted as *Pos*). In columns (4) and (8), we use a sub-sample with sentiment smaller than the median value (denoted as *Neg*). As control variables, we include *Log_ME*, *Log_BE_ME* , *Inst_Own*, and *SUE*. Industry definition is based on two-digit SIC codes. 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. Standard errors are clustered at the industry level and are reported within parentheses. ***, **, 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 = $Abn\_Ret_{[0,1]}$

	Raw Documents				Summarized Documents			
	Full		Pos	Neg	Full		Pos	Neg
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$Sentiment^{Raw}$	-0.008	0.001	0.005	-0.028				
	(0.008)	(0.012)	(0.017)	(0.020)				
$Sentiment^{Sum}$					0.025***	0.051***	0.027***	0.099***
					(0.004)	(0.009)	(0.006)	(0.011)
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	897	893	1,790	1,790	897	893
Adjusted R^2	-0.009	0.041	-0.027	0.001	0.017	0.098	0.001	0.121

Panel B. Conference Call Sample

Dependent Variable = $Abn\_Ret_{[0,1]}$

	Raw Documents				Summarized Documents			
	Full		Pos	Neg	Full		Pos	Neg
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$Sentiment^{Raw}$	0.035***	0.050***	0.006	0.040***				
	(0.006)	(0.007)	(0.013)	(0.011)				
$Sentiment^{Sum}$					0.078***	0.097***	0.092***	0.105***
					(0.006)	(0.008)	(0.008)	(0.011)
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	8,537	8,537	4,281	4,256	8,537	8,537	4,281	4,256
Adjusted R^2	0.054	0.073	0.045	0.065	0.111	0.141	0.093	0.147

Table 5. Variance Decomposition of Disclosure Bloat

This table performs variance decomposition of our *Bloat* measure. We report descriptive statistics in Panel A. In Panel B, we first report incremental R^2 from adding sets of year and industry fixed effects. It then zooms in on firm-specific variation and reports R^2 when including firm fixed effects. In Panel C, presents a transition matrix accross *Bloat* quintiles. Each row corresponds to a quintile of *Bloat* in prior year. Each column, corresponds to a quintile of *Bloat* in the current year. 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. Panel D repeats reports analogous transition matrix for the conference calls sample.

Panel A. Descriptive Statistics

			Percentiles		
Sample from	Mean	Std	p25	p50	p75
MD&A	0.754	0.081	0.694	0.744	0.819
Conference Call	0.685	0.126	0.577	0.670	0.785

Panel B. Fixed Effect Structure

	MD&A (1)	Conference Call (2)
B1. Incremental R^2		
Time FE	1.64%	5.19%
Industry FE	12.16%	3.36%
Time × Industry FE	16.27%	19.95%
Implied Firm Level	69.93%	71.50%
Sum	100.00%	100.00%
B2. Fraction of Variation		
Firm FE	46.23%	35.99%
Residual	53.77%	64.01%
Sum	100.00%	100.00%

Panel C. Time-Series Variation for MD&As

	$Q[Bloat_{it}]$					
$Q[Bloat_{it-1}]$	Low	2	3	4	High	Total
Low	48.81%	25.79%	14.29%	7.94%	3.17%	100.00%
2	23.90%	24.63%	27.57%	16.54%	7.36%	100.00%
3	12.89%	22.66%	26.95%	23.83%	13.67%	100.00%
4	7.11%	14.62%	24.11%	27.67%	26.49%	100.00%
High	0.93%	4.67%	9.81%	29.44%	55.15%	100.00%

Panel D. Time-Series Variation for Conference Calls

	$Q[Bloat_{it}]$					
$Q[Bloat_{it-1}]$	Low	2	3	4	High	Total
Low	48.01%	23.13%	15.37%	8.87%	4.62%	100%
2	23.86%	30.10%	21.76%	15.12%	9.16%	100%
3	13.99%	21.01%	26.62%	23.51%	14.87%	100%
4	7.96%	15.65%	23.28%	27.87%	25.24%	100%
High	3.28%	7.79%	14.96%	25.96%	48.01%	100%

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. Industry definition is based on two-digit SIC codes. Standard errors are reported within parentheses and 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%.

Dependent Variable =	*Bloat*			
Sample =	MD&A		Conference Call	
	(1)	(2)	(3)	(4)
Log_ME	-0.012***	-0.004	-0.006**	-0.009***
	(0.001)	(0.004)	(0.002)	(0.003)
Log_BE_ME	-0.005*	-0.000	0.004	0.010***
	(0.003)	(0.003)	(0.004)	(0.003)
N_Analyst	0.000	-0.000	0.000	0.000
	(0.000)	(0.000)	(0.001)	(0.000)
Inst_Own	-0.008**	-0.001	-0.006	-0.000
	(0.003)	(0.005)	(0.005)	(0.003)
Earn_Vol	0.026***	-0.005	0.117*	0.063
	(0.009)	(0.014)	(0.061)	(0.045)
ROA	-0.036*	-0.056**	-0.028	-0.008
	(0.021)	(0.026)	(0.053)	(0.043)
Loss	0.008*	0.002	0.019***	0.010***
	(0.004)	(0.004)	(0.005)	(0.003)
Sentiment	-0.073***	-0.026*	-0.030*	-0.111***
	(0.010)	(0.014)	(0.016)	(0.012)
Fog	-0.001	0.002	-0.004	0.019***
	(0.002)	(0.002)	(0.006)	(0.007)
PlainEng	0.017	-0.062	2.021***	1.058***
	(0.084)	(0.064)	(0.172)	(0.190)
Log_Length	0.084***	0.108***	0.085***	0.089***
	(0.006)	(0.012)	(0.009)	(0.008)
Complexity	0.033**	0.041***	0.076***	0.066***
	(0.013)	(0.012)	(0.025)	(0.014)
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	1,790	8,537	8,537
Adjusted R^2	0.643	0.771	0.365	0.556

Table 7. Capital Market Consequences of Bloat

This table reports the estimates from an OLS regression of proxies for capital markets' informational frictions on our *Bloat* measure. We use three proxies: probability of informed trading (*PIN*), abnormal bid-ask spreads (*Abn_Spread*), and post-filing volatility (*Post_Vol*). respectively. As control variables, we include all variables used in the determinant tests in Table 6 and several market-related variables (*Friday*, *One_Day_Ret*, *Price*, and *abs_SUE*). Industry definition is based on two-digit SIC codes. Panel A is based on a random sample of the MD&A disclosures. Panel B is based on a random sample of conference call transcripts. Avg.Dep. means the mean value of the dependent variable. Standard errors are reported within parentheses and 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 =	*PIN*		*Abn_Spread*		*Post_Vol*	
	(1)	(2)	(3)	(4)	(5)	(6)
Bloat	0.099***	0.062***	0.494***	0.544**	0.027***	0.037***
	(0.014)	(0.015)	(0.125)	(0.214)	(0.007)	(0.007)
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.	0.009	0.009	0.043	0.043	0.018	0.018
Cluster	Industry	Industry	Industry	Industry	Industry	Industry
N	1,790	1,790	1,790	1,790	1,790	1,790
Adjusted R^2	0.146	0.452	0.076	0.310	0.476	0.569

Panel B. Conference Calls Sample

Dep Var =	*PIN*		*Abn_Spread*		*Post_Vol*	
	(1)	(2)	(3)	(4)	(5)	(6)
Bloat	0.034***	0.037***	0.172**	0.226***	0.008***	0.007***
	(0.007)	(0.005)	(0.068)	(0.052)	(0.003)	(0.001)
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.	0.023	0.023	0.279	0.279	0.021	0.021
Cluster	Industry	Industry	Industry	Industry	Industry	Industry
N	8,537	8,537	8,537	8,537	8,537	8,537
Adjusted R^2	0.284	0.477	0.174	0.368	0.393	0.498

Table 8. Theme-Specific Summaries

This table reports descriptive statistics for theme-specific summaries and evaluates their informativeness. In Panel A, we report the annual averages of *%ESG*, *%Fin*, *lenESG*, and *lenFin*. $Sentiment^{ESG}$ is the sentiment of ESG summaries. $Sentiment^{Fin}$ is the sentiment of financial summaries. We run regressions by year and report the coefficient for each sentiment variable using the same set of controls as in Table 4. Standard errors are clustered at the industry level. Industry definition is based on two-digit SIC codes. We report the *t*-statistics associated with each regression coefficient. *t*-value time trend is obtained by regressing *t*-values on years. We use heteroscedasticity-robust standard errors to obtain the significance of time-series regression coefficients. Refer to Appendix A for detailed variable descriptions.

Panel A. Time Trends

	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
%ESG	20.37	20.48	18.84	16.61	15.20	18.06	20.28	18.21	18.32	22.08	26.75	38.20
%Fin	94.24	92.38	91.99	91.05	90.88	93.70	92.57	92.18	92.00	93.07	91.80	88.95
lenESG	1.29	1.15	1.24	1.13	1.06	1.25	1.64	1.34	1.69	2.12	2.65	5.00
lenFin	22.55	19.07	18.99	19.01	20.34	20.45	23.16	21.86	18.80	19.91	20.93	23.90

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

Dependent Variable	$= Abn\_Ret_{[0,1]}$			
	$Sentiment^{ESG}$		$Sentiment^{Fin}$	
	Coefficient	*t*-value	Coefficient	*t*-value
Year	(1)	(2)	(3)	(4)
2009	-0.004	-0.26	0.014	1.41
2010	-0.005	-0.29	0.014	2.09
2011	0.007	0.63	0.019	2.90
2012	0.035	2.99	0.020	3.20
2013	-0.005	-0.32	0.010	1.72
2014	-0.001	-0.08	0.025	3.40
2015	0.023	1.02	0.021	4.06
2016	0.011	0.61	0.039	5.01
2017	0.044	1.97	0.027	4.24
2018	0.059	2.36	0.034	4.32
2019	0.056	2.63	0.039	3.86
2020	0.070	4.06	0.028	3.07
Full Sample	0.022	3.49	0.026	10.41
t-value Time Trend	0.296	3.27	0.209	2.94