Executives vs Chatbots: Unmasking Insights through Human-AI Differences in Earnings Conference Q&A

November 2023

ABSTRACT

A significant portion of information shared in earnings calls is conveyed through verbal communication by corporate managers. However, quantifying the extent of new information provided by managers poses challenges due to the unstructured nature of human language and the difficulty in gauging the market's existing knowledge. In this study, we introduce a novel measure of information content (Human-AI Differences, HAID) by exploiting the discrepancy between answers to questions at earnings calls provided by corporate executives and those given by several context-preserving Large Language Models (LLM) such as ChatGPT, Google Bard, and an open source LLM. HAID strongly predicts stock liquidity, abnormal returns, analyst forecast accuracy following these calls, and propensity of managers to provide management guidance, consistent with HAID capturing new information conveyed by managers. Overall, our results highlight the importance of using LLM as a tool to help investors unveil the veiled – penetrating the information layers and unearthing hidden insights.

Keywords: ChatGPT; Bard; Large Language Model; AI; Conference Call; Chatbot; Information Content

JEL Classifications: C45, C88, D80, G3; G11, G12, G14, M41

1. Introduction

Prior research has underscored the significant role of earnings calls, specifically managers' responses during the Q&A sessions, as a pivotal source of value-relevant information (Matsumoto et al., 2011). The value attributed to a manager's response hinges on its capacity to offer novel insights that surpass the knowledge of a well-informed investor derived from prior sources, including the presentation session and the manager's preceding responses. However, quantifying the extent of new information embedded in the manager's response faces two challenges: the dynamic nature of soft information conveyed during earnings calls in the form of unstructured human language and the unobservable nature of an investor's prior knowledge. In this paper, we leverage the advancements offered by Large Language Models (LLMs) to address these challenges, reconstruct the pre-existing knowledge of an informed investor, and unveil the novel information conveyed by managers.

Previous studies have employed varied metrics, including tonal variations (Price et al., 2012), scripted language usage (Lee, 2016), and linguistic complexity (Bushee et al., 2018), aiming to capture distinct linguistic facets. However, while these metrics prove useful in gauging different dimensions of linguistic features like tone and complexity, they fall short of directly extracting the nuanced information content inherent in the language. In contrast, our approach employs LLMs, such as ChatGPT, to emulate an informed investor in real-time during a conference call. Specifically, before receiving a manager's response to a question, we provide all the information presented in the conference call to ChatGPT, equipping the model with real-time context and background knowledge akin to what an informed participant in the call would possess. Subsequently, we pose to ChatGPT the same question directed at the manager and evaluate the semantic disparities between responses to identical questions provided by the manager and those given by ChatGPT. In our approach, ChatGPT's answer to the identical question serves as a "benchmark," representing what an informed investor would expect a manager to respond before hearing the actual response. If the manager's actual response significantly differs from ChatGPT's response, it suggests the presence of new information provided by the manager. We conduct extensive tests to validate the effectiveness of our approach.

ChatGPT is an artificial intelligence (AI) model that uses deep learning techniques to produce natural-sounding text in response to user prompt.[2] The efficacy of our methodology relies on two distinctive features of ChatGPT that set it apart from previous AI models. Firstly, the model's knowledge base stems from pre-training on extensive textual data, providing the model a general knowledge background in macroeconomics, industry trends, and firm conditions that an informed investor is likely to possess.[3] Secondly, the " chatting " feature of ChatGPT exhibits a remarkable ability to generate responses by integrating information from prior discussions into ongoing discourse, effectively maintaining context throughout a conversation. When applied in our conference call setting, this contextual awareness empowers the model to produce responses not only contextually fitting but also enriched with insights gleaned from previous interactions between managers and investors during the call. Essentially, through the iterative process of supplying all the preceding information presented in the call before each manager's response to ChatGPT, we treat the model as a "mimic" of an informed investor's real-time knowledge base.

There are at least three primary benefits for our innovative context-preserving approach. First, ChatGPT can generate responses that are context-specific because these

[2] For an interesting exposition of GPT, please see https://www.newyorker.com/magazine/2019/10/14/can-a-machine-learn-to-write-for-the-new-yorker.
[3] We set the knowledge cutoff date of each conference call experience as the date of the conference call held. Specifically, we add the following phrase into the ChatGPT's system role instruction: "Knowledge cutoff: {<date>}", where <date> is the date of the conference call. Nevertheless, it is unclear how closely the LLM adheres to this specific request.

responses directly incorporate the newly provided conference call information into ChatGPT's knowledge base. Due to this iterative process, the model can use the information to better understand the context of investor questions and provide more tailored answers. Second, conference calls often have a distinct tone and style of communication established by the management team. By incorporating timely conference call information, we enable the model to adopt a similar tone and style in its responses, enhancing the coherence and consistency with the overall conversation. Finally, in instances where context is lacking, ChatGPT may respond to questions with apparent conviction but inaccuracy. The inclusion of additional information from the conference call serves to enhance ChatGPT's response accuracy, diminishing the likelihood of speculative or incorrect answers. By incorporating context shared by the CEO/CFO during the call, the model can refine its responses, improving the overall accuracy.[4]

The final step in constructing our novel measure of information content, is to use natural language processing techniques to gauge the semantic similarity between the actual answers provided by senior executives and those generated by ChatGPT for the same questions. We take the difference between 1 and this constructed similarity (i.e., 1-similarity) and term this the Human AI Difference (HAID hereafter). Intuitively, this measure provides the additional information that is ***not*** captured by what one can learn from analyzing all publicly available patterns and information as well as the specific information provided up to the moment an investor asks a question during the call.

Our empirical analysis begins with an initial sample of 190,538 earnings conference calls obtained from Capital IQ. We then merge stock price and return information (from CRSP) as well as analysts' forecasts (from I/B/E/S) before and after the earnings calls. We

[4] In Section 4 and Appendix A, we provide more technical details on how to accomplish this context preservation.

also obtain and successfully match firms' financial information (from Compustat) for a final sample of 104,932 earnings conference calls between 2004 and 2020 corresponding to 5,570 unique firms.

If the HAID measure, which quantifies semantic differences between human and ChatGPT generated answers, successfully captures new information content, earnings calls with higher HAID will generate more-informed trading activity from investors and analysts that should help reduce information asymmetry and enhance market liquidity. To investigate this idea, we correlate HAID with various outcome variables, including absolute cumulative abnormal stock returns and abnormal trading volume during and after the earnings call, the dispersion of analyst forecast revisions after the call, and analyst forecast accuracy after the call. Importantly, our empirical specifications account for the major component of the news during the conference call which is the earnings surprises, and other attributes such as the number of questions and sentences during the call. Following prior literature of conference calls (e.g., Bushee et al., 2018), we further control for time-varying firm characteristics such as firm size, market-to-book ratio, leverage, R&D, ROA, stock price volatility, analyst coverage, and the special items. To address unobserved heterogeneity and common temporal shocks, we include firm fixed-effects and year-quarter fixed effects.

Our initial analysis focuses on the impact of HAID on investors' response, reflected in the absolute cumulative abnormal return (CAR) and abnormal trading volume around the earnings call. The absolute CAR and abnormal trading volume capture the market's overall perception of the information content conveyed through the earnings call. If HAID truly captures hidden information that is not captured by previously known attributes associated with the earnings call, it should possess additional predictive power for the absolute CAR and abnormal trading volume even after controlling for these known characteristics. We find that this is indeed the case: HAID is strongly positively associated with the absolute CAR and abnormal trading volume in the three-day window before and after the conference call date event. In terms of economic magnitude, an increase in HAID from its 10th to 90th percentiles is associated with a 0.15 higher abnormal trading volume (8.2% of the sample mean) and a 2.7% higher absolute CAR (63.8% of the sample mean, or one half of the sample standard deviation).[5]

Next, we study HAID's relation to analyst behavior after earnings conference calls. Seminal work (Kothari, 2001; Michaely and Womack, 2005; and Frankel, Kothari, and Weber, 2006) finds that sell-side analysts' recommendations significantly influence market outcomes. Because nearly all financial valuation models directly or indirectly rely on earnings forecasts, earnings forecasts often trigger substantial movements in equity prices and returns. We focus on two specific elements: analyst forecast errors and analyst forecast dispersion. If earnings calls are more informative, as captured by a higher HAID, we expect that analysts will make smaller forecast errors and exhibit less cross-analyst disagreement after earnings calls. Our findings are consistent with this hypothesis. In terms of economic magnitude, an increase in HAID from its 10th to 90th percentiles is associated with a -.013 lower forecast errors (63% of the sample mean) and a -0.006 smaller absolute dispersion (4.2% of the sample mean).

Disclosure theory strongly posits a positive correlation between disclosure quality and liquidity (Verrecchia, 2001; Leuz and Wysocki, 2016; Goldstein and Yang, 2017). Consequently, if a higher HAID captures more value-relevant information disclosed during

---

[5] Absolute cumulative abnormal return is calculated on a three-day window around the conference call (one day before to one day after the call). The abnormal return is calculated as the raw return minus the return of the firms within the same size decile of the index following Kimbrough (2005). Absolute trading volume is computed as the mean daily trading volume during the event period (-1, 3) excluding the event day, minus the daily trading volume during the non-filing period (-49, -5), deflated by the standard deviation of daily trading volume during the non-filing period (-49, 5) following Miller (2010).

conference calls, we should anticipate an association where higher HAID is linked with enhanced liquidity. We measure illiquidity by the bid-ask spread and Amihud Ratio (both measures of stock illiquidity), and find that HAID is significantly and negatively related to both illiquidity measures, implying that HAID contains value-relevant information, thus improving price discovery and reducing information asymmetry. An increase in HAID from its 10^{th} to 90^{th} percentiles is associated with a 4.0% lower bid-ask spread and a 5.7% lower Amihud ratio.

To further substantiate our hypotheses regarding HAID capturing new information content, we reinforce our main findings with two supplementary tests. Firstly, we investigate the relationship between HAID and managers' propensity to issue guidance. If HAID genuinely reflects managers' readiness and capacity to provide private information, we expect these managers to be more inclined to offer guidance. Our results align with this expectation: an elevation in HAID from its 10th to 90th percentiles correlates with a 12.1% higher probability of issuing management guidance relative to the sample mean. In our second supplementary analysis, we build on the notion that if HAID truly captures new information content, its impact should be more pronounced when firms are inherently more complex, leading to a greater degree of information asymmetry between corporate insiders and investors. We employ four distinct proxies for firm complexity and information asymmetry utilized in prior research: R&D intensity, the number of business segments, insider trading intensity, and the number of managers participating in the conference call. In cross-sectional tests, we consistently find evidence that the association between HAID and abnormal trading volume, absolute abnormal return, and managers' likelihood to provide guidance is more pronounced when firms are more complex and when a greater degree of information asymmetry exists between firms and their investors. These results not only shed further light on the underlying mechanism but also provide suggestive yet prescriptive insights into when managers' non-machine-like information conveyance might be most beneficial.

Various LLMs are trained on distinct datasets and employ different underlying model structures. ChatGPT, for instance, utilizes the GPT-3.5 language model, while Google Bard is trained on Google's internal model, Palm 2. To ensure the robustness of our findings across diverse training datasets and language models, we examine our findings on two alternative LLMs, Google Bard and an open-source LLM. The results obtained from these alternative models are qualitatively similar, providing additional assurance that our documented effects are not driven by the specific choice of any particular LLM, such as ChatGPT. This substantiates the generalizability of our results to broader settings.

We view our paper's uniqueness as actually "chatting" with a Chatbot and allowing it to enhance its situational awareness to its fullest potential, thereby preserving the context. In a way, we are simulating an informed investor, who is physically dialing into the earnings conference call to absorb and digest all disclosed information. By constructing such a context-aware benchmark, our paper attempts to uncover the incremental information given by the executive *after* accounting all publicly available information up to the very question posed by the financial analyst. In our final set of analyses, we scrutinize the ChatGPT-generated HAID measure from various angles, striving for a deeper understanding of the significance of the context-preserving "chatting" feature inherent in our approach.

First, we explore the factors influencing the HAID measure, examining various firm and conference characteristics. We find that larger firms, those with more participating managers, higher instances of special items, and those covered by more analysts tend to exhibit higher HAID. This suggests that such firms, likely due to their complexity and heightened information demand from investors, tend to have more nuanced and detailed discussions. Conversely, firms with higher profitability and R&D expenditure demonstrate lower HAID, possibly indicating managerial caution in disclosing proprietary business information. Additionally, we observe that more complex questions and responses, as measured by the Fog index, are associated with lower HAID, implying that less information is conveyed when the language used is complex. Next, we formally test the value of the “chatting” feature of ChatGPT in our approach. To do so, we pose identical questions to ChatGPT as before, but withholding any contextual information, including the presentation session and preceding questions and responses. The outcomes reveal that, in the absence of contextual information, our contextless HAID measure exhibits no discernible associations with stock returns, trading volume, analyst forecast errors and dispersion, or the likelihood of managers providing guidance. This outcome underscores the inherent value of the context-preserving "chatting" feature of ChatGPT in constructing our human-AI disparity measure.

Our study primarily contributes to the voluntary disclosure literature, especially research on conference calls, by proposing a new methodology to uncover private information managers voluntarily disclose during earnings calls. Previous studies analyze linguistic features of language such as tone (Price et al., 2012), linguistic complexity (Bushee et al., 2018), disclosure horizon (Brochet et al., 2015), spontaneity (Lee, 2016), engagement (Rennekamp et al., 2022), and vocal cues (Mayew and Venkatachalam, 2012; Mayew et al., 2020). We contribute to existing literature by introducing a novel measure, HAID, leveraging advanced LLMs to quantify the semantic disparity between managers' actual responses and the "expected" responses envisioned by well-informed investors. A notable strength of our methodology lies in its ability to dynamically incorporate investor knowledge, encompassing both the corpus of public information and details revealed during the conference call. This dynamic integration provides a distinct advantage in identifying nuanced information content embedded within human language. Given the diverse array of topics covered in conference call conversations—ranging from revenue and costs to competitive landscape, innovation strategy, marketing efforts, and investment plans, among others—our methodology offers a practical and comprehensive measure of voluntary disclosure. In contrast, common measures of voluntary disclosure, such as the availability and accuracy of management forecasts, cannot fully capture the breadth of information shared during conference calls.

Our paper is also related to a burgeoning literature on how the advent of LLMs in processing and creating textual information. Concurrent research focuses on LLMs' superior ability to summarize or uncover hidden information in texts. Kim et al. (2023a) demonstrate that ChatGPT significantly reduces the length of corporate disclosures, enhancing their explanatory power for stock market reactions. Kim et al. (2023b) use the GPT 3.5 model to generate risk summaries from conference call transcripts, demonstrating that GPT-based measures offer significant information content, outperforming existing risk measures in predicting firm-level volatility and strategic choices. Bernard et al. (2023) utilize a fine-tuned GPT model to measure business complexity, linking it to delayed market price adjustments, filing delays, and increased regulatory scrutiny. Li et al. (2023) use ChatGPT to analyze analyst reports, extracting components related to corporate culture. They identify major drivers of cultural changes and find that firms with a strong culture experience positive outcome in market share, growth, profitability, and innovation. Bertomeu et al. (2023) leverage the March 2023 ban on ChatGPT by the Italian data protection authority as a natural experiment and find that the ban negatively affects the productivity of firms with higher exposure to generative AI and hampers analysts' information production. In contrast to concurrent studies, our distinctive methodology capitalizes on the "chatting" feature of ChatGPT, simulating an informed investor actively engaging in the earnings conference call, absorbing, and digesting all disclosed information in real time.

Finally, our contribution extends to the broader literature on AI-human interaction in financial markets. While existing studies often focus on AI-human competition and comparison (e.g., Costello et al., 2020; Cao et al., 2003) or how humans react to machine adoption (Du et al., 2003), we innovate by employing AI as a simulated investor to discern information already known to an informed investor in a dynamic setting. In this regard, our approach is also related to a recent literature that uses machine learning models as benchmarks to evaluate human decision errors (Kleinberg et al., 2018; Liu 2022), with our unique focus on using AI to evaluate human language rather than real decisions.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature, institutional background, and develops the testable hypotheses. Section 3 discusses the LLMs' theoretical usefulness as a context preserving benchmark. Section 4 demonstrates our sample construction and empirical methodology. Section 5 reports the empirical findings and we conclude in Section 6.

2. Related Literature, Institutional Background, and Hypothesis Development

2.1. HAID as Management's Private Information

Prior research explores the role of earnings conference calls as a means of voluntary disclosure and their impact on capital markets (Frankel et al., 1999; Bowen et al., 2002; Kimbrough, 2005). Matsumoto et al. (2011) finds that both management presentations and subsequent Q&A sessions with analysts contribute additional information beyond earnings press releases. Importantly, Q&A segments provide more useful information than formal presentations. Previous studies examine the origin and characteristics of this information, including managers' tone (Price et al., 2012), linguistic complexity (Bushee et al., 2018), disclosure horizon (Brochet et al., 2015), spontaneity (Lee, 2016), engagement (Rennekamp et al., 2022), and vocal cues (Mayew and Venkatachalam, 2012; Mayew et al., 2020).

Augmenting this literature, we leverage newly-developed Large Language Models (LLM) such as ChatGPT to quantify mangers' incremental information disclosure during Q&A sessions. We develop a measure called Human AI Difference (HAID), which measures the semantic dissimilarity between senior executives' answers to questions during actual conference calls and those generated by ChatGPT for the same questions. We begin by examining whether conference calls with larger variation between human and LLM responses (higher HAID) convey more information to investors. Consequently, our first hypothesis investigates whether conference calls with higher HAID generate more intensive trading and larger price response.

H1a: Conference calls with higher HAID exhibit higher abnormal trading volume and absolute cumulative abnormal returns.

Next, we examine analyst behavior using two measures: analyst forecast error and analyst forecast dispersion. Analyst forecast error measures the deviation between analyst predictions and corporate outcomes. Lower forecast errors indicate analysts' greater ability to predict company performance accurately. Forecast dispersion measures the extent to which different analysts' forecasts diverge. Higher dispersion indicates more disagreement among analysts. If a conference call effectively conveys better information, we expect both forecast error and dispersion to be lower. We summarize this below:

H1b: Conference calls with higher HAID are correlated with lower analyst error and analyst dispersion.

A primary benefit of disclosure, strongly supported by theory, is that it improves market liquidity. Information asymmetries among investors introduce adverse selection (Glosten and Milgrom, 1985; Kyle, 1985; Easley and O'Hara, 1987; Admati and Pfleiderer, 1988). Generally, uninformed investors that trade with informed investors face the risk of unfavorable trades - an informed investor is willing to sell (buy) at the current market price because they possess information indicating that the price is too high (too low). Consequently, uninformed investors may trade less, reducing market liquidity. However, more informative managerial disclosure can alleviate this adverse-selection problem and enhance market liquidity by leveling the playing field among investors (Verrecchia, 2001).

H1c: Conference calls with higher HAID are positively associated with stock liquidity.

2.2. HAID and Managerial Guidance

Next, we investigate the relationship between HAID and managers' propensity to issue guidance. Corporate executives provide guidance as a means to communicate important information about the company's financial performance, operations, and future prospects to investors, analysts, and other stakeholders. If HAID genuinely reflects managers' readiness and capacity to provide private information, we expect these managers to be more inclined to offer guidance. This leads us to the following hypothesis:

H2: Conference calls with higher HAID are positively associated with the probability of management guidance.

2.3. HAID, information content, and firm complexity

Finally, if HAID truly captures new information content, its impact should be more pronounced when firms are inherently more complex or present a higher degree of information asymmetry between corporate insiders and investors., such as those with more R&D intensity, operate in more industries, have more insider trading. Therefore, we hypothesize that the market outcomes influenced by HAID will be more prominent for such firms:

H3: The association between HAID and abnormal trading volume, absolute abnormal return, and managers' likelihood to provide guidance is more pronounced when firms are more complex and when a greater degree of information asymmetry exists between firms and their investors.

3. LLMs' Theoretical Usefulness as a Context Preserving Benchmark

The conference call setting is dynamic, characterized by rapid information exchange between managers and the market. Real-time assessment of the incremental information introduced by a manager's response during the call necessitates understanding what investors have already learned before hearing that response, a formidable task. Our methodology tackles this challenge by harnessing the unique " chatting " feature of ChatGPT, known for its capability to generate context-preserving content. Specifically, when engaging LLMs in a conversation, these models demonstrate proficiency in crafting insightful responses by incorporating information from previous discussions into ongoing discourse. This ability arises from the LLMs' robust capacity for contextual understanding, long-range dependencies, and memory retention. In this section, we offer a non-technical overview of each of these features.

3.1. Contextual Understanding

LLMs are deep neural networks trained on extensive text corpora with the primary objective of predicting the subsequent word in a sentence within a broader textual context. These models rely on the revolutionary Transformer architecture, trained to dynamically alter the interpretation of words based on contextual factors such as their position within a sentence (Vaswani et al., 2017). A pivotal component of the Transformer is the attention mechanism, a computational mechanism directing the model's focus toward pertinent words, facilitating understanding of contextual nuances. One can intuitively think of these mechanisms as the model's way of allocating attention to specific parts of the conversation, mimicking how we attend to certain details from earlier parts of the conversation while chatting.[6] When applied in our conference call setting, this contextual awareness empowers the model to produce responses not only contextually fitting but also enriched with insights gleaned from previous interactions between managers and investors during the call.

3.2. Long-Range Dependencies

LLMs exhibit a remarkable proficiency in capturing long-range dependencies within textual data (Vaswani et al., 2017). This attribute empowers them to establish meaningful connections across extended spans of text, presenting a substantial advantage in scenarios where insights are distributed across multiple turns of a conversation. The model's ability to link information from distant segments of the conversation history contributes significantly to the depth and coherence of its response generation process. This becomes particularly valuable in our conference call setting, where a manager's response follows complex prior discussions between the manager and other conference participants, as well as a presentation session. The long-range dependencies feature of LLMs ensures that the model can successfully consider and incorporate relevant information from various parts of the conversation history during the conference in its generated responses.

3.3. Memory Retention

The memory retention capacity of LLMs refers to their ability to remember and recall information from various contexts. This is made possible by the extensive datasets on which

[6] Seminal contributions from Vaswani et al. (2017) and Devlin et al. (2019) lay the foundation for the transformer architecture and its associated attention mechanisms.

these models are trained, affording them the ability to learn and retain information from diverse contexts.[7] Just like we remember important facts or discussions from earlier chats, LLMs can do the same but on a much larger and more complex scale. Memory retention, combined with their attention to relevant details from distant segments of the conversation history, helps LLMs provide insightful and contextually relevant responses when applied in our conference call setting.

4. Data Source, Sample Construction, and Empirical Methodology

4.1 Sample Construction

We start our sample construction using the universe of 190,538 conference calls' transcripts from a period of 2004 to 2020 retrieved from the CapitalIQ database. We then merge these conference calls to Compustat quarterly data to obtain data on firms' aggregated financials, the I/B/E/S data to obtain analysts' forecasts and management guidance based on the fiscal quarter end date of the conference call. We further drop the observations with the missing firm-level control variables that are listed in Appendix C. Our main sample consists of 104,932 earnings conference calls for the period of 2004 – 2020.

4.2 Empirical Methodology

4.2.1 Context Preservation in LLMs

One of our paper's key innovations is the real-time preservation of conversation context with various LLMs. This necessitates ensuring that the LLM's answers are contextually informed up to a given time point. We achieve this by iteratively providing all information presented in the conference call to the LLM, updating the model's knowledge

[7] Key academic references supporting these aspects include the foundational work of Vaswani et al. (2017) and Devlin et al. (2019), who have laid the principles underlying how the transformer architecture and attention mechanisms create the memory retention capabilities of LLMs.

like an informed participant in the call. This involves feeding the LLM all the information in the presentation session and preceding conversations in the Q&A session until a specific question is asked and updating the information environment with the actual response to the previous question when moving to the next. For ChatGPT, we use a combination of "system" and "assistant" interaction roles for context preservation, with technical procedures detailed in Appendix A.

4.2.2 Measures of Textual Similarity and HAID

To measure the HAID, we first use three distinct methodologies to construct the similarity between the answer provided by the senior executives and that provided by a context-relevant LLM such as ChatGPT to a specific question asked by an analyst. These similarities measures are computed at the question level by comparing answers by executives and the LLM. We then take the average value of similarity measures across all the questions in a conference call. Finally, HAID – at the conference call level - is then computed as the difference between 1 and this average value of similarities (i.e., 1-Similarity) across all answers in response to analyst questions asked during a conference call. Intuitively, the higher the value of HAID, the more different are the answers provided by the LLM vs. those provided by the executives. We employ three distinct methods of computing similarities detailed below:

Semantic similarity with BERT

Our first measure of similarity utilizes the BERT (Bidirectional Encoder Representations from Transformers) model developed by Google. BERT has emerged as a groundbreaking approach in natural language processing (NLP) based on deep learning. Unlike traditional methods that rely on fixed word embeddings, BERT introduces a contextualized word representation scheme. Through its pre-training phase on large-scale corpora, BERT captures the intricacies of language by considering both left and right context in a bidirectional manner. This contextual understanding allows BERT to generate rich, high-dimensional vector representations known as BERT embeddings. Important for our purpose, these embeddings encode comprehensive semantic information, accounting for subtle syntactic and semantic nuances in text. Consequently, BERT-based similarity measures provide a robust framework for assessing the relatedness between sentences or documents. BERT-based similarity often outperforms traditional methods like cosine similarity in NLP tasks.

Cosine word similarity

Our second measure of textual similarity is the commonly used cosine similarity, which in mathematical terms, can be expressed as $(A \cdot B) / (\|A\| * \|B\|)$, where A and B represent the vector representations of two textual entities, such as sentences, documents, or embeddings; $A \cdot B$ denotes the dot product of the two vectors, which measures the similarity in their directions or textual content; $\|A\|$ and $\|B\|$ represent the magnitudes or norms of the textual vectors A and B, respectively. By calculating the dot product of the textual vectors and dividing it by the product of their magnitudes, textual cosine similarity has a value that lies between -1 and 1. A value of 1 indicates that the textual vectors are perfectly similar, 0 indicates no similarity, and -1 indicates perfect dissimilarity. Cosine similarity is a straightforward and computationally efficient method for measuring similarity.

Word-embedding similarity (Word2Vec)

Word2Vec comprises a collection of interconnected models utilized for generating word embeddings. These models are characterized by their shallow architecture, consisting of two layers, and are trained to reconstruct the linguistic contexts of words. By taking a substantial body of text as input, Word2Vec constructs a vector space, usually spanning several hundred dimensions. In this vector space, each distinct word from the text corpus is associated with a specific vector, facilitating semantic representation and analysis of words. Our implementation is based on Word2Vec, a pre-trained embedding provided by Google, which provides an efficient implementation of the continuous bag-of-words and skip-gram architectures for computing vector representations of words.[8] Word2Vec similarity measures the relatedness or similarity between words based on the distributional properties learned during the Word2Vec training process. Words with similar meanings or contexts tend to have higher cosine similarity scores, as their vectors align more closely.

[Insert Table 1 Here]

Table 1 contains the summary statistics on all the variables used in our analyses. Panel A presents the mean, standard deviation and percentile distributions, while Panel B contains the pairwise correlation matrix between these key variables. Panel A shows that the distribution of all three proxies of HAID (*HAID(Bert), HAID(Cos),* and *HAID(Word2vec)*) is close to normal. The mean and standard deviation of other key metrics are also comparable to prior studies, which provides assurance that our sample is representative.

Panel B shows that the correlation between three measures of HAID and other variables is rather low. For example, none of the absolute value of the correlations between *HAID(Bert)* and other variables exceeds 0.10, implying that there are new sources of variation that is captured by *HAID(Bert)* but not other existing variables. *HAID(Bert)*, which is based on a more advanced language model, has relatively lower correlation (0.43 and 0.47, respectively) with the *HAID(Cos)* and *HAID(Word2vec)*; while the correlation coefficient between the latter two is 0.74.

[8] The pre-trained embeddings are available at https://code.google.com/archive/p/word2vec/.

4.2.3 Examples of Chatbot vs. Executives Answers

To give a better sense of how informed the answers given by a LLM can be, we provide two examples in Appendix B. Take Case 1 for instance, the Chatbot provides an answer that is quite similar to the one given by the executive. Specifically, it discusses the progress that has been made, the current status of hitting the SG&A target, and the management's overall (optimistic) attitude towards achieving the 10% SG&A reduction goal. Case 2 demonstrates a similar capability of providing similar answers to those given by the executives. The BERT similarity between answers given by the Chatbot and executives is 0.903 and 0.928 for Case 1 and Case 2, respectively. Correspondingly, HAID(Bert), which is one minus the BERT similarity, in the two examples is rather low (0.097 and 0.072), indicating that the content of executive responses is largely expected.

The purpose of these two examples is to demonstrate that the context-preserving process significantly helps with narrowing down the scope of the answer, precisely identifying the relevant data and/or situation, and ultimately giving an informed answer that is almost indistinguishable from those given by a business professional.

4.3 Empirical Specification

For our baseline analyses, we exploit a multivariate panel regression framework by estimating the following:

$$y_{iqt} = \alpha_1 HAID_{iqt} + X_{iqt}\beta + v_i + \omega_{qt} + \varepsilon_{iqt} \qquad (1)$$

where *i, q, and t* index firm, quarter, and year, respectively. The dependent variable y_{iqt} is at the firm-quarter-year level. The regression model includes as control variables an array of conference-call related variables and firm-level characteristics. Variables related to conference calls include: logarithm of the number of questions asked by analysts in the earnings call (*Ln_Qcount*) and logarithm of the average number of sentences in executive responses (*Ln_Sent*). The firm-level controls include the number of analysts covering the firm (*Numest*), earnings surprise (*SUE*), stock return volatility (*Ret_Sd*), firm size (*Size*), market-to-book ratio (*MtoB*), leverage ratio (*Lev*), return on assets (*ROA*), R&D expenditure normalized by total assets (*RD*) and special items (*SPI*). v_i is firm fixed effects that control for time-invariant omitted firm characteristics and ensure that estimates of α_1 reflect average within-firm changes in the outcome variable over time rather than simple cross-sectional correlations. The quarter-year fixed effects account for transitory nationwide factors such as macroeconomic conditions that could affect the outcome variables and *HAID* simultaneously.

The key parameter of interest is α_1, which measures the impact of HAID on the outcome variables.

5. Empirical Findings

5.1 Main Results: Information Content in HAID

In this section, we systematically investigate the impact of HAID on various market outcomes ranging from abnormal trading volume to analysts' forecasts. By examining a range of outcomes from different market participants, these tests provide a validation for the information content that is captured by our measure of human-AI differences.

5.1.1 HAID and Abnormal Trading & Absolute Abnormal Returns

We begin our analyses by investigating the impact of HAID on abnormal trading volume and the absolute cumulative abnormal returns in the short window after the earnings conference calls. The motivations for these tests are as follows:

First, if HAID truly conveys a greater amount of (previously unknown) private information, we expect that investors will respond more actively by engaging in increased trading. We thus expect that HAID be positively correlated with greater abnormal trading volume. Second, the new information contained in managerial responses through HAID should translate into more stock market reactions on a risk-adjusted basis. Therefore, we anticipate that HAID is positively related to the absolute value of cumulative abnormal returns.

[Insert Table 2 Here]

We estimate Equation (1) and present these results in Table 2. The results show a strong, positive correlation between HAID and abnormal trading volume and the absolute value of cumulative abnormal returns in the aftermath of the earnings conference calls. These results are consistent with H1a. The economic magnitude of the estimated coefficients on HAID is also quite sizeable. An increase in HAID from its 10^{th} to 90^{th} percentiles is associated with a 0.15 higher abnormal trading volume (8.2% of the sample mean) and a 2.7% higher absolute CAR (63.8% of the sample mean, or one half of the sample standard deviation).

5.1.2 HAID and Analyst Forecast Accuracy and Dispersion

After documenting the impact of HAID on trading volume and absolute cumulative announcement return, we now turn to one of the key participants in the earnings conference calls: sell-side financial analysts. To the extent that earnings conference calls usually contain material value-relevant information that is different from analysts' prior beliefs, previous research finds that analysts are more likely to issue forecasts with higher accuracy right after the conference call and these forecasts are more agreeing to each other (Cornell and Landsman, 1989; Kimbrough, 2005).

If HAID indeed contains material value-relevant information, we expect that the analysts' forecast error and dispersion should reduce significantly after the earnings call. We thus estimate Equation (1) but replace the outcome variable with the analyst forecast error and forecast dispersion. The forecast error is calculated as the difference between the closest consensus forecast issued right after the conference call of year-quarter *t* about year-quarter *t+1* and the actual EPS of year-quarter *t+1*. The forecast dispersion is calculated as the difference between the highest and lowest forecasted EPS issued right after the conference call of year-quarter *t* about year-quarter *t+1* and the actual EPS of year-quarter *t+1*.

[Insert Table 3 Here]

The results of this exercise are presented in Table 3. Irrespective of the underlying methodology for computing HAID, we observe a negative and significant coefficient which indicates that a higher HAID is associated with lower forecast errors and forecast dispersion, which is consistent with H1b. In terms of economic magnitude, an increase in HAID from its 10^{th} to 90^{th} percentiles is associated with a -.013 lower forecast errors (63% of the sample mean) and a -0.006 smaller absolute dispersion (4.2% of the sample mean). As these forecasts are issued right after the conference call, this test lends us further confidence in our hypothesized information content captured by HAID.

5.1.3 HAID and Liquidity

We next study the impact of the HAID on stock liquidity. As discussed previously, information asymmetries among investors introduce adverse selection (Glosten and Milgrom, 1985; Kyle, 1985; Easley and O'Hara, 1987; Admati and Pfleiderer, 1988), and managerial revelation with actual value-relevant information content (i.e., high HAID) can help alleviate the adverse-selection problem and enhance market liquidity by leveling the playing field among investors (Verrecchia, 2001). We thus expect that higher HAID be negatively correlated with illiquidity measures.

[Insert Table 4 Here]

The results of this exercise are presented in Table 4. We employ two measures to capture stock illiquidity: the Amihud illiquidity measure and the bid-ask spread. The Amihud illiquidity measure quantifies the level of liquidity risk associated with a particular security by measuring the sensitivity of its price to changes in trading volume. Specifically, the Amihud illiquidity measure is calculated by dividing the absolute value of the daily return of an asset by its trading volume.[9] Our second metric, the bid-ask spread, measures the difference between the highest price a buyer is willing to pay (bid price) and the lowest price a seller is willing to accept (ask price) for a particular security or asset. The bid-ask spread reflects the immediate trading costs faced by market participants. A higher (lower) value of bid-ask spread measure can be bought or sold with (without) significantly impacting its price.

We find results consistent with our conjecture (H1c): higher HAID is negatively associated with illiquidity. Specifically, an increase in HAID from its 10^{th} to 90^{th} percentiles is associated with a 4.0% lower bid-ask spread and a 5.7% lower Amihud ratio. That is, the more dissimilar it is the executives' answer compared to that given by ChatGPT, the higher the liquidity of the stocks subsequently.

5.2 HAID and Managerial Guidance

After establishing the relation between HAID and various market outcomes on information, we also investigate how managers' propensity to issue quantitative forward-looking guidance relates to HAID. If HAID genuinely reflects managers' readiness and

[9] Amihud illiquidity measure is calculated as the daily average of (|stock return| / trading volume).

capacity to provide private information, we expect these managers to be more inclined to offer specific guidance.

[Insert Table 5 Here]

We test this hypothesis (H2) with a linear probability model and present the results in Table 5. Our results align with this expectation: an elevation in HAID from its 10th to 90th percentiles is associated with a 12.1% higher probability of issuing quantitative management guidance relative to the sample mean.

5.3 Cross Sectional Tests

If the semantic dissimilarity between the answers provided by executives and ChatGPT as captured by HAID has strong predictive power for various market outcomes, the impact of such dissimilarity should be more pronounced when the firms are more complex to begin with, resulting in a greater degree of information asymmetry between corporate insiders and investors. To capture such information wedge, we employ four distinct proxies inspired by prior literature: R&D intensity, number of business segments, insider trading, and number of managers participated in the conference call. Prior research has provided abundant evidence that insider trading activities, high R&D intensity, and across-industry operations are sources of information asymmetry between managers and investors, due to the opaque and complex nature of such activities (Aboody and Lev, 2000; Chan et al., 2001; Cohen and Lou, 2012). Consequently, firms with complex business usually have a larger number of managers participating in conference calls to address more specific issues raised by analysts (Lu et al., 2023).

Specifically, R&D is an indicator variable that is set equal to one if the firm-quarter R&D expense is greater than zero in the Compustat data and zero otherwise. Insider Trade is the net percentage insider trading during the firm-quarter. Business segments and participants are quartile ranking of the total number of business segments and the total number of managers participated in the conference call, respectively.

We then re-estimate a variant of Equation (1) by augmenting the baseline model with the above conditioning variables and their respective interactions with HAID. Our focus then, is to examine whether the coefficients on the interaction terms are economically meaningful and statistically significant. The results of this exercise are contained in Table 6. Overall, we find broad support for our conjecture, holding HAID constant, the abnormal trading volume, absolute abnormal return around the conference calls, and managers' likelihood to provide earnings guidance are higher when firms are more complex and when a greater amount of information asymmetry exists between firms and their investors. These results shed further light not only on the underlying mechanism, but also offer suggestive yet prescriptive evidence in terms of when managers' non-machine like information conveyance might be most beneficial.

5.4 Robustness Tests – Other LLMs

Our main analyses are conducted by using ChatGPT as our primary source of machine-generated responses. As mentioned previously, we manually feed into ChatGPT context-specific information related to the specific conference call and "train" the ChatGPT to be as situationally aware as possible. However, different LLMs are trained using different techniques. For instance, ChatGPT utilizes a masked language modeling (MLM) objective. It randomly masks some words in the input and learns to predict those masked words based on the surrounding context. This process helps the model acquire a general understanding of language patterns and structures. Hence, it is crucial to ensure that our documented results are not the artifact of the specific models that are employed by ChatGPT, in which case HAID would simply be capturing a bias inherent that arises because of ChatGPT's specific methodology.

[Insert Table 7 Here]

We therefore re-estimate our main regressions (hypotheses 1a through 1c) using two alternative LLMs: Google Bard and Stability AI Language Models (StableLM).[10] The results of these tests are contained in Table 7. All of our baseline regression results are robust to employing these alternative LLMs as the engine for generating responses.

5.5 The Context Preserving Feature of LLMs

Our methodology capitalizes on the "chatting" capability of Chatbot, enhancing its situational awareness and preserving context. This unique feature allows us to simulate an informed investor, like ChatGPT, actively participating in an earnings conference call to thoroughly comprehend disclosed information in real-time. In this section, we examine the context-preserving property of our HAID measure from different perspectives, aiming at better understanding the importance of the "chatting" feature in our methodology.

First, we investigate the determinants of the HAID measure, considering diverse firm and conference attributes. In Table 8, we find that larger firms, those with more participating managers, higher instances of special items, and broader analyst coverage tend to have higher HAID, indicating more detailed and nuanced discussions, possibly due to heightened investor information demand. Conversely, firms with higher profitability and R&D expenditure show lower HAID, suggesting managerial caution in divulging proprietary information. We also find that more complex language, as measured by the Fog index, is associated with lower HAID, indicating reduced information conveyance. Notably, lagged

[10] For details about the implementation of StableLM-Alpha, please refer to https://github.com/Stability-AI/StableLM.

HAID does not correlate with current HAID after accounting for various factors, implying that HAID reflects managers' spontaneous responses rather than a consistent communication style. Overall, firm and conference characteristics (alongside fixed effects) can explain 27-60% of HAID variation.

Next, we formally assess the significance of ChatGPT's "chatting" feature in our methodology. To do so, we pose identical questions to ChatGPT as before, but withholding any contextual information, including the presentation session and preceding questions and responses. The HAID measure created in this way is contextless in the sense that the Chatbot is not aware of the information environment. We revisit our baseline results using this contextless HAID and report our findings in Table 9. We observe that the contextless HAID measure lacks discernible associations with stock returns, trading volume, analyst forecast errors and dispersion, or the likelihood of managers providing guidance. This outcome emphasizes the intrinsic value of ChatGPT's context-preserving "chatting" feature in shaping our human-AI disparity measure.

Lastly, we verify that our HAID measure captures soft information content in conversations that is distinct from linguistic features. Bushee et al. (2018) demonstrate that linguistic complexity in conference calls involves two latent components: an "information" component reflecting linguistic complexity related to informative technical disclosure about the business, and an "obfuscation" component involving linguistic complexity designed to diminish the informativeness of the disclosure. The novelty of their empirical approach builds on the idea that analysts have little incentive to obfuscate, and thus complex language conveyed by managers have a larger “information” component if analysts also use similarly complex language in their questions. In contrast, complex language used by managers have a larger “obfuscation” component if analysts do not use complex language in their questions.

They measure complexity using the Fog index based on the average number of words per sentence and the percentage of complex words. Therefore, the underlying theoretical construct captured by their measures relate to linguistic features of disclosure rather than concrete information content. In contrast, our approach aims at capture the concrete information content of manager responses by leveraging the newly available LLM technology.

Finally, we validate that our HAID measure captures soft information content in conversations that is distinct from linguistic features. Seminal work of Bushee et al. (2018) decomposes linguistic complexity in conference calls into an "information" component reflecting complexity in informative technical disclosure and an "obfuscation" component involving complexity to reduce informativeness. The novelty of their empirical approach relies on the idea that analysts have little incentive to obfuscate, resulting in complex language by managers having a larger "information" component if analysts also use similarly complex language in their questions. In contrast, complex language used by managers has a larger "obfuscation" component if analysts do not use complex language in their questions. Importantly, Bushee et al. (2018) gauge complexity using the Fog index based on the average number of words per sentence and the percentage of complex words, and therefore do not account for contextual information embedded in the language.

In contrast, we aim to capture the tangible information content of manager responses by harnessing the advanced LLM technology, providing a novel perspective that goes beyond linguistic features to delve into the substance of the disclosed information. To verify that HAID genuinely captures information content distinct from linguistic features, we conduct a regression analysis of our outcome variables on HAID and the two complexity components identified by Bushee et al. (2018), alongside the list of control variables. The results, presented in Table 10, show minimal changes in the coefficients of HAID when the two variables representing linguistic complexity are introduced, compared with our previous findings in Tables 2, 3, and 5. This stability in the coefficients suggests that HAID effectively captures soft information content that is distinguishable from linguistic features alone.

6. Conclusion

With the rise of artificial intelligence and automated information gathering, firms are learning how to "talk" and communicate with the outside investing public (Cao, Jiang, Yang, and Zhang, 2023). A previously unexplored area is whether machines and large language models (LLM) based on artificial intelligence can help investors detect value-relevant information from firms' voluntary disclosure.

In this paper, we use earnings conference calls as a setting and introduce a novel measure of information content (Human Machine Differences, HAID) by exploiting the discrepancy between answers to questions at earnings conference calls provided by actual corporate CFOs and CEOs and those given by several context-preserving Large Language Models (LLM) including ChatGPT. Unique to our methodology is our systematic approach to preserve the context of the conversation. By feeding the most relevant real-time information to the LLM, our approach provides a methodical way for future studies to form the most informative language benchmark.

We validate HAID and test its impact on various market outcomes. In particular, HAID has significant predictive power for the absolute cumulative abnormal return and trading volumes around earnings call, stock liquidity, analyst forecast accuracy and dispersion, as well as management's propensity to provide quantitative guidance. Our results are robust to alternative LLM such as Google Bard and open source LLM (StableLM-Alpha).

Overall, we find that HAID provides a unique and previously unidentified source and methodology to help investors uncover new information content. Our results highlight the importance of using LLM as a tool to help investors unveil the veiled – penetrating the information layers and unearthing hidden insights, in the new era of the rise of machines.

References

Aboody, D., & Lev, B. (2000). Information asymmetry, R&D, and insider gains. *Journal of Finance,* 55, 2747–2766.

Abramova, I., J. E. Core, and A. Sutherland. 2020. Institutional investor attention and firm disclosure. *The Accounting Review* 95(6): 1–21.

Admati, A. and P. Pfleidrer. 1988. "Selling and Trading on Information in Financial Markets." *American Economic Review* 13: 96–103.

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, Lin Y, Liu Y, Ni Z. Capital Market Consequences of Generative AI: Early Evidence from the Ban of ChatGPT in Italy. Working Paper.

Blankespoor, E. 2019. The impact of information processing costs on firm disclosure choice: Evidence from the XBRL mandate. *Journal of Accounting Research* 57(4): 919–967.

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

Bowen, R.M., Davis, A.K. and Matsumoto, D.A., 2002. Do conference calls affect analysts' forecasts?. *The Accounting Review*, 77(2), pp.285-316.

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

Bushee, B. J., and Y. Huang, 2002. Do Analysts and Investors Efficiently Respond to Managerial Linguistic Complexity during Conference Calls? Working Paper

Bushee, B. J., D. Matsumoto, and G. Miller. 2004. Managerial and investor responses to disclosure regulation: The case of Reg FD and conference calls. *The Accounting Review* 79 (3): 617– 643.

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

Chan, L., Lakonishok, J., Sougiannis, T., 2001. The stock market valuation of research and development expenditures. *Journal of Finance* 56, 2431–2456.

Cohen, Lauren, and Dong Lou, 2012, Complicated firms, *Journal of Financial Economics* 104, 383–400.

Costello, A., A. Down; and M. Mehta. "Machine + Man: A Randomized Field Experiment on the Role of Private Information in Lending Markets." *Journal of Accounting and Economics* 70 (2020): 101360.

DellaVigna, S., and J. M. Pollet. 2009. Investor inattention and Friday earnings announcements. *Journal of Finance* 64(2): 709–749.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1, 4171–4186 (2019).

Du, K., Liu, M. and Wang, W., 2023. Human Information Production in the Machine Age: Evidence from Automated Information Acquisition in the Asset Management Industry. Working Paper

Dyer, T., Lang, M. and Stice-Lawrence, L., 2017. The evolution of 10-K textual disclosure: Evidence from Latent Dirichlet Allocation. *Journal of Accounting and Economics*, 64(2-3), pp.221-245.

Easley, D. and O'hara, M., 1987. Price, trade size, and information in securities markets. *Journal of Financial economics*, *19*(1), pp.69-90.

Frankel, R., Johnson, M. and Skinner, D.J., 1999. An empirical examination of conference calls as a voluntary disclosure medium. *Journal of Accounting Research*, *37*(1), pp.133-150.

Glosten, L.R. and Milgrom, P.R., 1985. Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of financial economics*, 14(1), pp.71-100.

Goldstein, I., and L. Yang "Information Disclosure in Financial Markets." *Annual Reviews of Financial Economics* 9 (2017): 101–25

Graham, J., M. Hanlon, T. Shevlin, and N. Shroff. 2017. Tax rates and corporate decision-making. Review *of Financial Studies* 30(9): 3128–3175.

Hirshleifer, D., S. S. Lim, and S. H. Teoh. 2009. Driven to distraction: Extraneous events and underreaction to earnings news. *Journal of Finance* 64(5): 2289–2325.

Hirshleifer, D., S. S. Lim, and S. H. Teoh. 2011. Limited investor attention and stock market misreactions to accounting information. *Review of Asset Pricing Studies* 1(1): 35–73.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig; and . Mullainathan, "Human Decisions and Machine Predictions," *Quarterly Journal of Economics* 133 (2018): 237–93

Kim, A.G., Muhn, M. and Nikolaev, V.V., 2023a. Bloated Disclosures: Can ChatGPT Help Investors Process Information?. Chicago Booth Research Paper, (23-07).

Kim, A.G., Muhn, M., Nikolaev, V.V., 2023b. From Transcripts to Insights: Uncovering Corporate Risks Using Generative AI Fama-Miller Center Research Paper.

Kimbrough, M. D. 2005. The effect of conference calls on analyst and market underreaction to earnings announcements. *Accounting Review* 80 (1): 189–219.

Kothari, S.P., Shu, S. and Wysocki, P.D., 2009. Do managers withhold bad news?. *Journal of Accounting research*, *47*(1), pp.241-276.

Kyle, A.S., 1985. Continuous auctions and insider trading. *Econometrica: Journal of the Econometric Society*, pp.1315-1335.

Lawrence, A., J. P. Ryans, E. Sun, and N. Laptev. 2018. Earnings announcement promotions: A Yahoo Finance field experiment. *Journal of Accounting and Economics* 66 (2–3): 399–414.

Lee, J., 2016. Can investors detect managers' lack of spontaneity? Adherence to predetermined scripts during earnings conference calls. *Account. Rev.*

Leuz, C., and P. Wysocki. "The Economics of Disclosure and Financial Reporting Regulation: Evidence and Suggestions for Future Research." *Journal of Accounting Research* 54 (2016): 525–622.

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

Li, K., Mai, F., Shen, R., Yang, C. and Zhang, T., 2023. Dissecting Corporate Culture Using Generative AI–Insights from Analyst Reports. Available at SSRN 4558295.

Liberti, J. M. and M. A. Petersen. 2018. Information: Hard and soft. *Review of Corporate Finance Studies 8(1): 1–41.*

Liu, M. 2022. Assessing human information processing in lending decisions: A machine learning approach. *Journal of Accounting Research* 60(2): 607–651.

Lu, J. Matsumoto, D., Moon, K., 2023. Straight From the Horses' Mouth: Determinants and Consequences of Managers' Conference Call Participation. Working Paper

Matsumoto, D., M. Pronk, and E. Roelofsen. 2011. What makes conference calls useful? The information content of managers' presentations and analysts' discussion sessions. *The Accounting Review* 86(4): 1383–1414.

Mayew, W. J., and M. Venkatachalam. 2012. The power of voice: Managerial affective states and future firm performance. *The Journal of Finance* 67 (1): 1–44.

Mayew, W.J., Sethuraman, M., Venkatachalam, M., 2020. Individual analysts' stock recommendations, earnings forecasts, and the informativeness of conference call question and answer sessions. *Account. Rev.* 95 (6), 311-337.

Price, S. M., J. Doran, D. Peterson, and B. Bliss. 2012. Earnings conference calls and stock returns: The incremental informativeness of textual tone. *Journal of Banking & Finance 36 (4): 992–1011.*

Rennekamp, K., M. Sethuraman, and B. Steenhoven. 2022. Engagement in Earnings Conference Calls. *Journal of Accounting and Economics.*

Simon, H.A. 1955. A behavioral model of rational choice. *Quarterly Journal of Economics* 69(1): 99–118.

Sims, C.A. 2003. Implications of rational inattention. *Journal of Monetary Economics* 50(3): 665–690.

Twedt, B., and L. Rees. 2012. Reading between the lines: An empirical examination of qualitative attributes of financial analysts' reports. *Journal of Accounting and Public Policy* 31 (1): 1–21.

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.

Verrecchia, R. "Essays on Disclosure." *Journal of Accounting & Economics* 32 (2001): 97–180.

Appendix A: In-context Learning and Dynamic Conversation Context Preservation

In this section, we detail the specific steps and technical procedures we follow to allow Large Language Models (LLMs) such as ChatGPT to learn the specific context pertaining to the earnings conference calls.

In the ChatGPT API, three roles are provided that allow for different types of interaction with the LLM: "system", "assistant", and "user". We achieve context preservation with a combination of text summarization and the full utilization of the 'assistant' role in ChatGPT API.

The "system" role refers to the core functionality of the ChatGPT API. In our setting, it contains the following two instructions: "From the perspective of a top executive, please answer the following question raised by a financial analyst during an earnings conference call." and " Knowledge cutoff: {<date>}". The latter aims to limit the knowledge of the LLM to the date of the conference call held.

The "assistant" role is used to personalize chat experience and can be configured to provide LLM with the information content pertaining to an analyst question.

It is imperative to preserve the context of the conference call when posing a question raised by an analyst to a LLM as the same question might lead to very different answers in different occasions. A typical conference call includes a presentation made by senior executives (CEOs and/or CFOs) and an open-floor Q&A session. While the former is usually an uninterrupted speech, the latter features a two-way interaction in which the financial analysts ask a question that is to be answered by the executives.

To ensure that the LLM has the most updated information pertaining to the firm in question, we follow the following procedures to achieve context preservation:

1. For question n raised by an analyst, we specifically provide two pieces of information:
   a) The executive presentation (summarized)
   b) The full set of preceding (n-1) sets of questions and their corresponding human answer. The analyst questions are the original text and human answers are summarized.
2. Note that context of a question becomes more specific for latter questions In a conference call as it *dynamically* encompasses the conversation between executives and other analysts prior to the specific question being posed.
3. For ChatGPT, the model we used is gpt-3.5-turbo, which comes with a limit of 4096 tokens (roughly 4096*0.75=3072 words). We therefore need to be mindful to the token limit while preserving conversation context. We achieve this goal by pre-process the presentation and executive answers using a state-of-the-art model for abstractive text summarization—BART, a denoising autoencoder for pretraining sequence-to-

sequence model developed by Facebook AI in 2019. BART is built on a transformer-based neural machine translation architecture which can be viewed as generalizing BERT (Bidirectional encoder), GPT (left-to-right decoder).[11]

An Illustrative Example:

We provide an example to illustrate the process of iteratively using both the system role and the assistant role in combination to achieve context preservation.

The table below illustrates the input to ChatGPT [context of the executive presentation and the preceding (n-1) questions and answers]. Then, the '**user**' role returns ChatGPT answer. Below is the input flow (the context for question n) to ChatGPT that leads to its answer to question n raised by an analyst.

"role": "`system`", "content": From the perspective of a top executive, please answer the following question raised by a financial analyst during an earnings conference call." and " Knowledge cutoff: {<date>}"
"role": "`assistant`", "content": **executive** presentation (summarized)
"role": "`user`", "content": **analyst** question 1 (full text)
"role": "`assistant`", "content": **executive** answer to question 1 (summarized)
"role": "`user`", "content": **analyst** question 2 (full text)
"role": "`assistant`", "content": **executive** answer to question 2 (summarized)
…
"role": "`user`", "content": **analyst** question (n-1) (full text)
"role": "`assistant`", "content": **executive** answer to question (n-1) (summarized)
"role": "`user`", "content": **analyst** question n (full text)

It is important to note that the context preserves executive presentation and the analyst-executive Q&A before question n. The context does not contain information about ChatGPT's response to previous questions.

The output illustrated below is recorded as ChatGPT's answer to question n.

[11] We confirm the robustness of our finding in a random sample of conference calls, for which the text summarization is conducted using PEGASUS (pre-training with extracted gap-sentences for abstractive summarization) developed by Google AI in 2020.

“role”: “assistant", "content": **ChatGPT** answer to question n

The input is then updated and extended to include executive answer to question n and question (n+1) and then feed to ChatGPT for its answer for question (n+1).

Next, the input flow (the context for question n+1)

"role": "system", "content": From the perspective of a top executive, please answer the following question raised by a financial analyst during an earnings conference call: " and " Knowledge cutoff: {<date>}"
"role": "assistant", "content": **executive** presentation (summarized)
"role": "user", "content": **analyst** question 1 (full text)
"role": "assistant", "content": **executive** answer to question 1 (summarized)
"role": "user", "content": **analyst** question 2 (full text)
"role": "assistant", "content": **executive** answer to question 2 (summarized)
…
"role": "user", "content": **analyst** question (n-1) (full text)
"role": "assistant", "content": **executive** answer to question (n-1) (summarized)
"role": "user", "content": **analyst** question n (full text)
"role": "assistant", "content": **executive** answer to question n (summarized)
"role": "user", "content": **analyst** question (n+1) (full text)

Overall, the system role and the assistant role work together to create a specific and dynamic information environment for every analyst question. And ChatGPT’s answer takes into the consideration of executive presentation and the conversation history between executives and analysts before the specific question is asked.

From an intuitive standpoint, the context does not contain information about ChatGPT’s response to previous questions. This process simulates a chatbot that ‘listens’ through the conference call and then answer question n.

Other LLM Models:

- StableLM has a similar set of three roles, named as <|SYSTEM|>, <|ASSISTANT|>, and <|USER|>.
- Google's Bard API does not require one to specifically label the three roles. The sequence of conversation is recognized by stacking "content" in the order of "system", "assistant", "user",…,"assistant", "user".

Appendix B: Two Cases of Chatbot vs. Executives Answers

Case 1 (Sim_Bert = 0.903; transcriptid = 1561284; gvkey = 1004):
Question: I just had a couple of more things. On the SG&A, I think you got down to 10.5% and you've targeted 10%. So how do we -- how should we think about that trending from here, and are you ahead of plan? I mean, could you get below 10% or are you just tracking as you want to? When do we get there?

Manager A:
We are in line with expectations at this point. Consistent with what we articulated at the Investor Day, we've got a target to get back to that 10%. We get target till we get back to that 10% or slightly below run rate. And as the business continues to grow throughout the year, we expect continued improvement there.

ChatBot A:
Yes, we are tracking well on our SG&A reduction plan, and we expect to realize additional progress during the rest of the year. We're not looking to give guidance beyond what we provide in the earnings release, but we are confident that we can achieve the 10% SG&A target.

Case 2 (Sim_Bert = 0.928; transcriptid = 139369; gvkey = 1050):
Question: And it's nice to see that automotive ticking back in there again. Is the bid activity in automotive improving, or is it just kind of steady here?

Manager A:
Automotive has been improving significantly. And as you know, we've reshaped a couple of our portfolios to pursue the more high-margin activity, the higher gross margin activity in our products. So we have captured some solid automotive projects and we are continuing to bid several. So we're seeing some positive trends in automotive, Dale.

ChatBot A:
Yes, we're definitely seeing a steady improvement in automotive activity. We have a very strong product portfolio in the automotive sector, and we have been seeing steady bid activity as more and more auto manufacturers look to meet global emissions standards. We feel very good about our position in this market and the potential for continued growth going forward.

Appendix C: Variable Definitions

Human-AI Differences (HAID):	
HAID	For each question raised in an earnings conference call, we first calculate textual similarity between executive and chatbot responses using one of the following methods. Then, *HAID* is the average value of (1-textual similarity) across answers given to all questions asked by analysts in the conference call.
HAID(Bert)	HAID based on semantic similarity using the Bidirectional Encoder Representations from Transformers (BERT) developed by Google.
HAID(Cos)	HAID based on cosine similarity of word distributions.
HAID(Word2Vec)	HAID based Word2Vec, a pre-trained embedding provided by Google. The pre-trained embeddings are available at https://code.google.com/archive/p/word2vec/.
Dependent and Control variables:	
Size	The natural logarithm of the market value of equity, calculated as the stock price at the fiscal quarter end (prcc_q) multiplied by the number of common shares outstanding (cshoq). Source: Compustat.
MtoB	The market-to-book ratio, calculated as the market value of the firm (atq + cshoq × prcc_q - ceqq) divided by the book value of the asset (atq). Source: Compustat.
Lev	The sum of long-term debt (dlttq) and the portion of long-term debt due in the coming quarter (dlcq) scaled by total asset (atq). Source: Compustat.
Ret_SD	The standard deviation of monthly stock returns (ret) during fiscal year *t*. Source: CRSP.
ROA	The standard deviation of earnings before extraordinary items and discontinued operations (ib) scaled by end-of-year total assets (at) for fiscal years *t-4* to year *t*. Source: Compustat.
R&D	The change in diluted earnings per share excluding extraordinary items (epsfx) from fiscal year *t-1* to *t*, scaled by the stock price at the end of fiscal year *t* (prcc_f). Source: Compustat.
Numest	The natural logarithm of 1 plus the number of financial analysts (numest) who forecasted the firm's earnings per share for fiscal quarter according to the IBES summary dataset compiled in the last month before the earnings announcement for fiscal quarter *t*. Source: IBES.
Spi	The special items (spiq) of fiscal quarter *t* scaled by the total asset (atq) of fiscal quarter *t*. Source: Compustat.
Num_Q	The natural logarithm of the number of questions in the Q&A session of the conference call. Source: Capital IQ.
Num_Sent	The natural logarithm of the number of sentences in the conference call transcript. Source: Capital IQ.
Abs.CAR	The absolute value of the cumulative abnormal return for a (-1,1) three-days window around the conference call. The abnormal return is calculated as the raw return minus the return of the firms within the same size decile of the index following Kimbrough (2005). Source: CRSP.
AVol	The mean daily trading volume during the event period (-1, 3) minus the daily trading volume during the non-filing period (-49, -5), deflated by the standard deviation of daily trading volume during the non-filing period (-49, 5) following Miller (2010). Source: CRSP.

Forecast Error	The difference between the forecasted EPS and the actual EPS. The forecast is measured as the first forecast on the EPS of fiscal quarter *t+1* right after the conference call of fiscal quarter *t*. Source: I/B/E/S.
Forecast Dispersion	The difference between the highest forecasted EPS and the lowest forecasted EPS. The forecast is measured as the first forecast on the EPS of fiscal quarter *t+1* right after the conference call of fiscal quarter *t*. Source: I/B/E/S.
Bid-Ask	The bid-ask spread in the three-days event window (-1, 1) around the conference call. The bid-ask spread is calculated as the ask price minus the bid price then scaled by the mid-point of bid and ask price. Source: CRSP.
Amihud Ratio	The Amihud illiquidity measure, which is calculated by dividing the absolute value of the daily return of an asset by its trading volume. Source: CRSP.
Guidance	An indicator variable that equals to one if the managers provided quantitative forecast during the conference call and zero otherwise. Source: I/B/E/S.
Cross-sectional variables:	
R&D	An indicator variable that is set equal to one if the firm-quarter R&D expense is greater than zero in the Compustat data and zero otherwise
Segment	Quartile ranking of the total number of business segments
Insider Trade	Net percentage insider trading during the firm-quarter
Participation	Quartile ranking of the total number of managers participated in the conference call

TABLE 1. Summary Statistics

Panel A provides the summary statistics of the key variables; and Panel B shows the correlation matrix. Refer to Appendix C for variable definitions.

Panel A: Summary Statistics

VARIABLES	Obs.	Mean	SD	P10	P25	Median	P75	P90
Bid-Ask Spread	104,718	-6.884	1.253	-8.395	-7.809	-7.023	-6.127	-5.162
Amihud Ratio	104,932	-6.612	2.543	-9.788	-8.476	-6.785	-4.947	-3.138
Abs. CAR	91,248	0.042	0.054	0.004	0.010	0.024	0.052	0.100
AVol	104,878	1.815	2.764	-0.325	0.113	0.974	2.449	4.798
Guidance	104,932	0.144	0.302	0.000	0.000	0.000	0.000	0.667
Forecast Error	92,673	0.021	0.168	-0.100	-0.020	0.020	0.060	0.160
Dispersion	92,673	0.136	0.202	0.010	0.030	0.070	0.150	0.320
HAID(Bert)	104,932	0.311	0.045	0.255	0.280	0.308	0.339	0.370
HAID(Cos)	104,932	0.472	0.060	0.395	0.433	0.473	0.512	0.548
HAID(Word2vec)	104,932	0.118	0.024	0.090	0.101	0.115	0.131	0.149
Num_Q	104,932	2.524	0.507	1.792	2.197	2.639	2.944	3.045
Num_Sent	104,932	2.360	0.426	1.818	2.062	2.342	2.639	2.926
Spi	104,932	-0.003	0.010	-0.007	-0.002	0.000	0.000	0.000
Numest	104,932	1.677	0.987	0.000	1.099	1.792	2.485	2.890
Unexp_Earn	104,932	0.001	0.051	-0.023	-0.006	0.000	0.006	0.024
Ret_SD	104,932	0.115	0.067	0.050	0.068	0.098	0.142	0.200
Size	104,932	7.371	1.821	5.076	6.095	7.310	8.553	9.853
MtoB	104,932	3.205	5.587	0.766	1.232	2.104	3.800	7.114
Lev	104,932	0.584	1.161	0.000	0.046	0.224	0.589	1.342
ROA	104,932	0.002	0.041	-0.032	-0.001	0.008	0.019	0.034
R&D	104,932	0.025	0.057	0.000	0.000	0.000	0.022	0.082

Panel B: Pairwise Correlation

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
(1) HAID(Bert)	1.00												
(2) HAID(Cos)	**0.43**												
(3) HAID(Word2vec)	**0.47**	**0.74**											
(4) Num_Q	**0.09**	**0.12**	**-0.18**										
(5) Num_Sent	**-0.06**	**-0.59**	**-0.42**	**0.04**									
(6) Spi	**0.02**	**0.04**	**0.02**	**0.01**	**-0.01**								
(7) Numest	**-0.02**	**-0.07**	**-0.18**	**0.37**	**0.11**	-0.01							
(8) Unexp_Earn	**-0.01**	0.00	0.00	0.00	0.00	**0.31**	**-0.01**						
(9) Ret_SD	**-0.02**	**-0.02**	**0.02**	**-0.13**	**-0.05**	**-0.11**	**-0.13**	**0.04**					
(10) Size	**-0.05**	**-0.20**	**-0.25**	**0.33**	**0.27**	**0.07**	**0.43**	**-0.02**	**-0.44**				
(11) MtoB	**-0.05**	**-0.07**	**-0.08**	**0.06**	**0.07**	**0.01**	**0.11**	-0.01	**-0.02**	**0.16**			
(12) Lev	**0.03**	0.00	**0.02**	**-0.05**	0.00	**-0.04**	**-0.10**	**0.04**	**0.21**	**-0.12**	**-0.16**		
(13) ROA	**0.02**	**0.07**	0.00	**0.14**	-0.01	**0.35**	**0.08**	**0.28**	**-0.36**	**0.30**	0.00	**-0.07**	
(14) R&D	**-0.06**	**-0.07**	**-0.04**	**-0.08**	0.00	**-0.07**	0.00	0.00	**0.25**	**-0.19**	**0.15**	**-0.18**	**-0.46**

TABLE 2. Human-AI Differences (HAID) and Market Response

This table reports the results of the regression of the market response as the dependent variable and the HAID measures as the key independent variable. Please refer to Appendix C for detailed variable definitions. Panel A reports the results when the market response is defined as the absolute value of the cumulative abnormal return of the (-1, 1) event window and Panel B reports the results when the market reaction is defined by abnormal trading volume following Miller (2010). The model controls for firm and fiscal year-quarter fixed effects, and the standard errors are clustered by industry and fiscal year-quarter. ***, **, * denote statistical significance at the 1%, 5%, or 10% level.

Panel A: Abnormal Volume

VARIABLES	(1) AVol	(2) AVol	(3) AVol
HAID(Bert)	**1.241***** **(6.995)**		
HAID(Cos)		**1.068***** **(4.981)**	
HAID(Word2vec)			**3.281***** **(7.167)**
Num_Q	0.164*** (3.576)	0.159*** (3.435)	0.197*** (4.320)
Num_Sent	0.026 (0.761)	0.100** (2.461)	0.083** (2.218)
Spi	-2.876 (-1.368)	-2.909 (-1.387)	-2.892 (-1.377)
Numest	-0.008 (-0.428)	-0.008 (-0.396)	-0.007 (-0.377)
Unexp_Earn	-0.343 (-1.525)	-0.341 (-1.526)	-0.344 (-1.532)
Ret_SD	1.987*** (3.383)	1.995*** (3.403)	1.998*** (3.416)
Size	0.199*** (4.495)	0.200*** (4.519)	0.201*** (4.538)
MtoB	-0.001 (-0.499)	-0.001 (-0.506)	-0.001 (-0.490)
Lev	0.060* (1.866)	0.060* (1.867)	0.060* (1.870)
ROA	3.879*** (5.084)	3.873*** (5.097)	3.889*** (5.111)
R&D	0.365 (0.487)	0.366 (0.492)	0.366 (0.493)
Firm Year-Qtr FE	Yes	Yes	Yes
Ind. Year-Qtr Cluster	Yes	Yes	Yes
Observations	104,878	104,878	104,878
Adj. R^2	0.301	0.301	0.301

Panel B: **Absolute Value of Cumulative Abnormal Return**

VARIABLES	(1) Abs. CAR	(2) Abs. CAR	(3) Abs. CAR
HAID(Bert)	**0.223*** **		
	(6.186)		
HAID(Cos)		**0.334*** **	
		(10.434)	
HAID(Word2vec)			**0.622*** **
			(10.236)
Num_Q	-0.001	-0.005	0.005
	(-0.376)	(-1.394)	(1.511)
Num_Sent	-0.010	0.014**	0.001
	(-1.602)	(2.045)	(0.195)
Spi	-0.059	-0.074	-0.059
	(-0.241)	(-0.303)	(-0.241)
Numest	-0.004	-0.005	-0.004
	(-0.892)	(-0.895)	(-0.868)
Unexp_Earn	0.054	0.056	0.054
	(1.585)	(1.630)	(1.583)
Ret_SD	-0.104**	-0.103**	-0.103**
	(-2.127)	(-2.088)	(-2.088)
Size	-0.028***	-0.028***	-0.028***
	(-3.128)	(-3.113)	(-3.102)
MtoB	0.000	0.000	0.000
	(0.367)	(0.358)	(0.366)
Lev	-0.005	-0.005	-0.005
	(-1.143)	(-1.159)	(-1.154)
ROA	0.580***	0.582***	0.582***
	(4.669)	(4.658)	(4.645)
R&D	-0.087	-0.084	-0.087
	(-0.853)	(-0.830)	(-0.853)
Firm Year-Qtr FE	Yes	Yes	Yes
Ind. Year-Qtr Cluster	Yes	Yes	Yes
Observations	91,248	91,248	91,248
Adj. R^2	0.087	0.087	0.087

TABLE 3. Human-AI Differences (HAID) and Analyst Forecast

This table reports the results of the regression of the analysts' reaction to the conference call information as the dependent variable and the HAID measures as the key independent variable. Panel A reports the results when the analysts' reaction is defined as the forecast error which is the difference between forecasted EPS and the actual EPS in the forecast right after the conference call, and Panel B reports the results when the analysts' reaction is defined by forecast dispersion, which is the difference between the highest and lowest forecasted EPS. The model controls for firm and fiscal year-quarter fixed effects, and the standard errors are clustered by industry and fiscal year-quarter. ***, **, * denote statistical significance at the 1%, 5%, or 10% level.

Panel A: Forecast Error

VARIABLES	(1) Forecast Error	(2) Forecast Error	(3) Forecast Error
HAID(Bert)	**-0.111*** **		
	(-6.233)		
HAID(Cos)		**-0.083*** **	
		(-5.518)	
HAID(Word2vec)			**-0.224*** **
			(-5.112)
Num_Q	-0.001	-0.001	-0.004***
	(-1.158)	(-0.913)	(-2.954)
Num_Sent	-0.008***	-0.014***	-0.012***
	(-2.906)	(-4.428)	(-4.133)
Spi	-1.358***	-1.356***	-1.359***
	(-7.914)	(-7.888)	(-7.906)
Numest	-0.005*	-0.005*	-0.005*
	(-1.797)	(-1.815)	(-1.858)
Unexp_Earn	0.397***	0.397***	0.397***
	(8.522)	(8.526)	(8.538)
Ret_SD	0.023	0.022	0.022
	(0.660)	(0.639)	(0.645)
Size	-0.006	-0.006	-0.006
	(-1.484)	(-1.506)	(-1.496)
MtoB	0.000**	0.000**	0.000**
	(2.246)	(2.287)	(2.274)
Lev	-0.010***	-0.010***	-0.010***
	(-4.242)	(-4.229)	(-4.227)
ROA	1.294***	1.295***	1.294***
	(12.087)	(12.082)	(12.066)
R&D	0.114***	0.114***	0.114***
	(4.043)	(4.041)	(4.052)
Firm Year-Qtr FE	Yes	Yes	Yes
Ind. Year-Qtr Cluster	Yes	Yes	Yes
Observations	92,673	92,673	92,673
Adj. R^2	0.241	0.241	0.241

Panel B: Forecast Dispersion

VARIABLES	(1) Forecast Dispersion	(2) Forecast Dispersion	(3) Forecast Dispersion
HAID(Bert)	**-0.047*** **		
	(-3.112)		
HAID(Cos)		**-0.081*** **	
		(-6.087)	
HAID(Word2vec)			**-0.173*** **
			(-4.901)
Num_Q	-0.002	-0.001	-0.003**
	(-1.123)	(-0.550)	(-2.197)
Num_Sent	0.004**	-0.001	0.001
	(2.148)	(-0.480)	(0.613)
Spi	-0.686**	-0.681**	-0.685**
	(-2.554)	(-2.535)	(-2.549)
Numest	0.070***	0.070***	0.070***
	(10.145)	(10.127)	(10.142)
Unexp_Earn	0.061**	0.061**	0.061**
	(2.424)	(2.425)	(2.434)
Ret_SD	0.253***	0.253***	0.253***
	(5.241)	(5.233)	(5.238)
Size	0.020***	0.020***	0.020***
	(3.749)	(3.746)	(3.751)
MtoB	-0.000	-0.000	-0.000
	(-1.372)	(-1.372)	(-1.383)
Lev	0.022***	0.022***	0.022***
	(6.023)	(6.034)	(6.035)
ROA	-0.088	-0.089	-0.090
	(-0.532)	(-0.536)	(-0.537)
R&D	0.089*	0.089*	0.089*
	(1.966)	(1.945)	(1.965)
Firm Year-Qtr FE	Yes	Yes	Yes
Ind. Year-Qtr Cluster	Yes	Yes	Yes
Observations	92,673	92,673	92,673
Adj. R^2	0.602	0.602	0.602

TABLE 4. Human-AI Differences (HAID) and Liquidity

This table reports the results of the regression of the capital market liquidity around the conference call as the dependent variable and the HAID measures as the key independent variable. Panel A reports the results when the market liquidity is defined as the bid-ask spread which is the difference between the daily bid and ask price scaled by the average of bid and ask price, and Panel B reports the results when the liquidity is defined as the Amihud illiquidity measure. The model controls for firm and fiscal year-quarter fixed effects, and the standard errors are clustered by industry and fiscal year-quarter. ***, **, * denote statistical significance at the 1%, 5%, or 10% level.

Panel A: Bid-Ask Spread

VARIABLES	(1) Bid-Ask	(2) Bid-Ask	(3) Bid-Ask
HAID(Bert)	**-0.335*** **		
	(-8.443)		
HAID(Cos)		**-0.341*** **	
		(-6.242)	
HAID(Word2vec)			**-0.779*** **
			(-7.744)
Num_Q	-0.034***	-0.032***	-0.042***
	(-4.173)	(-3.772)	(-5.351)
Num_Sent	-0.001	-0.025***	-0.015**
	(-0.144)	(-2.947)	(-2.029)
Spi	0.579*	0.593**	0.581*
	(1.970)	(2.026)	(1.988)
Numest	-0.035***	-0.035***	-0.035***
	(-4.139)	(-4.141)	(-4.158)
Unexp_Earn	0.063	0.062	0.063
	(0.863)	(0.858)	(0.868)
Ret_SD	-0.171	-0.174	-0.174
	(-1.465)	(-1.488)	(-1.487)
Size	-0.567***	-0.567***	-0.567***
	(-27.926)	(-27.890)	(-27.956)
MtoB	0.000	0.001	0.000
	(0.869)	(0.873)	(0.855)
Lev	0.019**	0.019**	0.019**
	(2.433)	(2.432)	(2.426)
ROA	-0.674***	-0.674***	-0.676***
	(-4.040)	(-4.038)	(-4.041)
R&D	-0.093	-0.094	-0.093
	(-0.554)	(-0.561)	(-0.555)
Firm Year-Qtr FE	Yes	Yes	Yes
Ind. Year-Qtr Cluster	Yes	Yes	Yes
Observations	104,718	104,718	104,718
Adj. R^2	0.794	0.794	0.794

Panel B: Amihud illiquidity

VARIABLES	(1) Amihud Ratio	(2) Amihud Ratio	(3) Amihud Ratio
HAID(Bert)	**-0.473*** **		
	(-6.925)		
HAID(Cos)		**-0.313*** **	
		(-3.687)	
HAID(Word2vec)			**-0.691*** **
			(-3.457)
Num_Q	-0.090***	-0.090***	-0.100***
	(-5.980)	(-5.900)	(-6.742)
Num_Sent	-0.016	-0.038**	-0.028**
	(-1.201)	(-2.655)	(-2.123)
Spi	1.397**	1.401**	1.390**
	(2.417)	(2.417)	(2.404)
Numest	-0.047***	-0.047***	-0.047***
	(-4.340)	(-4.361)	(-4.380)
Unexp_Earn	-0.083	-0.083	-0.082
	(-1.077)	(-1.084)	(-1.065)
Ret_SD	-0.825***	-0.828***	-0.828***
	(-4.602)	(-4.624)	(-4.621)
Size	-1.241***	-1.242***	-1.242***
	(-61.057)	(-60.946)	(-61.030)
MtoB	0.001	0.001	0.001
	(1.144)	(1.159)	(1.147)
Lev	-0.023	-0.023	-0.023
	(-1.580)	(-1.584)	(-1.581)
ROA	-1.503***	-1.499***	-1.500***
	(-4.504)	(-4.494)	(-4.487)
R&D	0.167	0.167	0.168
	(0.583)	(0.586)	(0.590)
Firm Year-Qtr FE	Yes	Yes	Yes
Ind. Year-Qtr Cluster	Yes	Yes	Yes
Observations	104,932	104,932	104,932
Adj. R^2	0.895	0.895	0.895

TABLE 5. Human-AI Differences (HAID)and Management Guidance

This table reports the results of the regression of the management guidance specificity during the conference call as the dependent variable and the HAID measures as the key independent variable. The management guidance specificity is defined as an indicator variable that equals to one if the guidance is quantitative and zero otherwise. The model controls for firm and fiscal year-quarter fixed effects, and the standard errors are clustered by industry and fiscal year-quarter. ***, **, * denote statistical significance at the 1%, 5%, or 10% level.

VARIABLES	(1) Guidance	(2) Guidance	(3) Guidance
HAID(Bert)	**0.152*** **		
	(9.913)		
HAID(Cos)		**0.178*** **	
		(7.701)	
HAID(Word2vec)			**0.356*** **
			(8.830)
Num_Q	0.006***	0.005**	0.010***
	(2.744)	(2.165)	(4.287)
Num_Sent	0.001	0.014***	0.008*
	(0.375)	(3.125)	(1.872)
Spi	-0.182*	-0.190*	-0.183*
	(-1.708)	(-1.797)	(-1.732)
Numest	0.005**	0.005**	0.005**
	(2.323)	(2.329)	(2.364)
Unexp_Earn	0.005	0.006	0.005
	(0.320)	(0.345)	(0.311)
Ret_SD	0.044	0.046	0.046
	(1.611)	(1.660)	(1.662)
Size	0.008***	0.008***	0.008***
	(2.730)	(2.784)	(2.794)
MtoB	0.000	0.000	0.000
	(0.483)	(0.474)	(0.490)
Lev	0.002	0.002	0.002
	(1.148)	(1.154)	(1.165)
ROA	0.082**	0.083**	0.083**
	(2.005)	(2.033)	(2.030)
R&D	-0.043	-0.043	-0.043
	(-1.248)	(-1.236)	(-1.259)
Firm Year-Qtr FE	Yes	Yes	Yes
Ind. Year-Qtr Cluster	Yes	Yes	Yes
Observations	104,932	104,932	104,932
Adj. R^2	0.383	0.383	0.383

TABLE 6. Cross-sectional Analyses

This table repeats the regression of the main results on HAID and the interaction between HAID and four cross-sectional variables, defined in Appendix C. ***, **, * denote statistical significance at the 1%, 5%, or 10% level, respectively.

VARIABLES	(1) AVol	(2) AVol	(3) AVol	(4) AVol	(5) Abs. CAR	(6) Abs. CAR	(7) Abs. CAR	(8) Abs. CAR	(9) Guidance	(10) Guidance	(11) Guidance	(12) Guidance
HAID* × *R&D	**0.731***				**0.001**				**0.054***			
	(1.988)				**(0.007)**				**(1.919)**			
HAID* × *Segment		**0.068**				**0.090*****				**0.041****		
		(0.420)				**(2.792)**				**(2.316)**		
HAID* × *Insider Trade			**0.063*****				**0.002**				**0.006***	
			(2.876)				**(0.507)**				**(1.824)**	
HAID* × *Participants				**0.057**				**0.057***				**0.053*****
				(0.335)				**(1.960)**				**(2.819)**
Other Control Variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Firm Year-Qtr FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Ind. Year-Qtr Cluster	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	104,878	93,037	101,971	104,878	91,248	80,024	88,641	91,248	104,932	93,086	102,024	104,932
Adj. R^2	0.301	0.301	0.301	0.301	0.087	0.089	0.087	0.087	0.383	0.398	0.383	0.383

TABLE 7. Using Other Large Language Models to Measure Human-AI Differences (HAID)

This table repeats the regression of the main results on two alternative HAID measures. Panel A reports the results when HAID is calculated based on responses generated by Google Bard, which is powered by Google's PaLM (Pathways Language Model); and Panel B reports the results when HAID is calculated based on the responses produced by an open source StableLM-Alpha model. ***, **, * denote statistical significance at the 1%, 5%, or 10% level, respectively.

Panel A: Google Bard

VARIABLES	(1) AVol	(2) Abs. CAR	(3) Forecast Error	(4) Forecast Dispersion	(5) Guidance
HAID(Bert)	**0.758***** **-5.381**	**0.192***** **-4.225**	**-0.032***** **(-3.448)**	**-0.023**** **(-2.008)**	**0.077***** **-3.311**
Other Ctrl	Yes	Yes	Yes	Yes	Yes
Firm Year-Qtr FE	Yes	Yes	Yes	Yes	Yes
Ind. Year-Qtr Cluster	Yes	Yes	Yes	Yes	Yes
Observations	104,534	90,947	92,428	92,428	104,587
Adj. R^2	0.301	0.087	0.24	0.602	0.383

Panel B: Open sourced StableLM-Alpha

VARIABLES	(1) AVol	(2) Abs. CAR	(3) Forecast Error	(4) Forecast Dispersion	(5) Guidance
HAID(Bert)	**1.444***** **-6.503**	**0.219***** **-5.205**	**-0.144***** **(-7.797)**	**-0.021*** **(-1.987)**	**0.149***** **-7.009**
Other Ctrl	Yes	Yes	Yes	Yes	Yes
Firm Year-Qtr FE	Yes	Yes	Yes	Yes	Yes
Ind. Year-Qtr Cluster	Yes	Yes	Yes	Yes	Yes
Observations	104,864	91,240	92,660	92,660	104,918
Adj. R^2	0.302	0.087	0.242	0.602	0.383

TABLE 8. Determinants of Human-AI Differences (HAID)

This table reports the results of the regression of the determinants of HAID measure. The HAID measures are defined as the dissimilarity between human and AI responses using Google Bert, Cosine similarity and Word2vec. The model controls for the one fiscal quarter lagged HAID, as well as firm and fiscal year-quarter fixed effects, and the standard errors are clustered by industry and fiscal year-quarter. ***, **, * denote statistical significance at the 1%, 5%, or 10% level.

VARIABLES	(1) *HAID(Bert)*	(2) *HAID(Cos)*	(3) *HAID(Word2vec)*
Lag HAID(Bert)	-0.007		
	(-1.365)		
Lag HAID(Cos)		0.000	
		(0.070)	
Lag HAID(Word2vec)			-0.001
			(-0.323)
Num_Q	0.002	0.000	-0.012***
	(0.980)	(0.176)	(-11.953)
Num_Sent	-0.031***	-0.125***	-0.044***
	(-14.012)	(-74.132)	(-41.913)
Sent_Diff	-0.003***	-0.005***	-0.002***
	(-15.148)	(-28.644)	(-23.246)
Tone	0.008	0.005	0.024***
	(1.214)	(0.612)	(8.996)
Tone_Diff	0.022***	0.003	0.023***
	(3.725)	(0.595)	(13.731)
Participants	0.001***	0.001***	0.000***
	(7.331)	(10.722)	(7.436)
Fog_Present	-0.000	-0.000	0.000**
	(-0.342)	(-1.303)	(2.577)
Fog_Q	-0.001***	-0.001***	-0.001***
	(-8.884)	(-9.954)	(-9.120)
Fog_Response	-0.001***	-0.008***	-0.003***
	(-6.120)	(-33.892)	(-23.208)
Spi	0.047***	0.063***	0.017*
	(2.753)	(4.005)	(1.882)
Numest	0.001***	0.001***	0.000**
	(3.235)	(3.237)	(2.049)
Size	0.002***	0.000	0.000
	(3.896)	(0.896)	(1.166)
MtoB	-0.000	-0.000	-0.000
	(-0.722)	(-0.272)	(-0.722)
Lev	0.000	-0.000	-0.000
	(0.860)	(-0.321)	(-0.067)
ROA	-0.029***	-0.026***	-0.014***

	(-3.674)	(-3.996)	(-4.055)
R&D	-0.011*	-0.012**	-0.003**
	(-1.912)	(-2.549)	(-2.137)
Segment	-0.000	-0.000	0.000
	(-0.458)	(-0.702)	(0.357)
Firm Year-Qtr FE	Yes	Yes	Yes
Ind. Year-Qtr Cluster	Yes	Yes	Yes
Observations	85,667	85,667	85,667
Adj. R^2	0.269	0.601	0.466

TABLE 9. Contextless Human-AI Differences (HAID)

This table repeats the regression of the main results when HAID measure is calculated based on a contextless method, in which we do not sequentially summarize and provide the information of the presentation and the *n-1* questions and answers when prompt ChatGPT the *n*th question in the conference call. ***, **, * denote statistical significance at the 1%, 5%, or 10% level, respectively.

VARIABLES	(1) AVol	(2) Abs. CAR	(3) Forecast Error	(4) Forecast Dispersion	(5) Guidance
HAID(Bert)	**-0.216**	**-0.061**	**-0.028**	**0.020**	**-0.003**
	(-1.107)	**(-1.594)**	**(-1.579)**	**(1.301)**	**(-0.161)**
Other Ctrl	Yes	Yes	Yes	Yes	Yes
Firm Year-Qtr FE	Yes	Yes	Yes	Yes	Yes
Ind. Year-Qtr Cluster	Yes	Yes	Yes	Yes	Yes
Observations	104,534	90,947	92,428	92,428	104,587
Adj. R^2	0.316	0.099	0.251	0.603	0.388

TABLE 10. Complexity, Obfuscation and Human-AI Differences (HAID)

This table repeats the regression of the main results and additionally controls for the information and obfuscation components in conference calls as in Bushee et al., (2018). ***, **, * denote statistical significance at the 1%, 5%, or 10% level, respectively.

	(1)	(2)	(3)	(4)	(5)
VARIABLES	AVol	Abs. CAR	Forecast Error	Forecast Dispersion	Guidance
HAID(Bert)	**1.270*** **	**0.211*** **	**-0.116*** **	**-0.047*** **	**0.159*** **
	(6.575)	**(4.893)**	**(-6.333)**	**(-3.065)**	**(8.116)**
Info	0.038***	-0.004	-0.001	0.003**	0.001
	(3.310)	(-1.490)	(-1.031)	(2.096)	(0.634)
Obfus	0.016***	-0.001	-0.001**	0.001***	0.000
	(2.713)	(-1.028)	(-2.575)	(3.263)	(0.267)
Other Ctrl	Yes	Yes	Yes	Yes	Yes
Firm Year-Qtr FE	Yes	Yes	Yes	Yes	Yes
Ind. Year-Qtr Cluster	Yes	Yes	Yes	Yes	Yes
Observations	88,980	77,895	79,011	79,011	89,020
Adj. R^2	0.301	0.093	0.251	0.598	0.388