Deciphering non-responses in earnings calls with Chat-GPT

Abstract: We developed a prompt that can decipher managers' non-responses during earnings calls with the most state-of-the-art large language model——ChatGPT4-turbo. We validate the measurement of managers' non-responses by comparing it with human-annotated labels, conducting granular call-level analyses, and assessing its correlations with analyst forecasts and stock market reactions. We find …

Consistent with the prediction that managers deliberately withhold bad news/proprietary information/…, we find that managers' non-responses predict negative future earnings growth/. We demonstrate ChatGPT's great potential in capturing subtle managerial behaviors in an effective way.

1. Introduction

Among various channels of information disclosures, earnings conference calls are distinctive by providing "two-way" interactions where investors, analysts, and other participants can ask the company's management for further information regarding the presented performances. However, the company's managers have the right to decline or defer their answers to certain questions (hereafter "managers' non-responses"). As a form of proactive behavior, managers' non-responses during an earnings call can also convey valuable signals. Thus, researchers are concerned with the motivations and economic consequences of managers' non-responses during a conference call. To empirically answer this question, a foremost problem unsolved is how to measure managers' non-responses on a large scale. It can be challenging to identify managers' non-responses given the intricacies and nuances of human conversations in a real situation. The recent booms of large language models (LLMs) possessing basic reasoning and semantic comprehension abilities, significantly improved on existing textual analysis workflows (Ties, 2024; Dong et al., 2024). Among them, the most representative model is ChatGPT, developed by OpenAI. With around 1.7 trillion parameters and pre-trained on the most updated dataset, the latest

ChatGPT4-turbo model can make better estimates of sentences and grasp human-level understanding given the context provided. It can handle up to 128K tokens and accept both texts and images as inputs, underlining its great potential to capture critical information from narrative disclosures at a large scale. In this paper, we develop a novel measure of managers' non-responses during an earnings conference by leveraging the ChatGPT4-turbo model, aiming to answer the following questions: 1. Can ChatGPT understand human languages and precisely capture managers' non-responses during an earnings call? 2. With our new measures to examine, what are managers' motivations for their non-response during earnings conference calls? 3. With our new measures to examine, what are the economic consequences of managers' non-responses in the short and long window?

ChatGPT, like other large language models, performs downstream tasks via prompting, which means all relevant task specifications and data process is formatted as textual input context, and the model returns a generated text completion. ChatGPT with its powerful ability of comprehending and generating texts, enjoys some priorities in processing large-scale textual data. Firstly, their standards seldom vary once the prompt and settings are determined, while human assistants restrained by limited attention can change from time to time. Theoretically speaking, ChatGPT can work as an indefatigable research assistant who will never get tired or bored. The classifications generated by ChatGPT could be more consistent relative to human coding. Secondly, Chat-GPT can "understand" the context, therefore it can recognize more implicit non-responses characterized by blathering or driveling. Thirdly, with adequate prompting, ChatGPT is more flexible in dealing with textual data relative to traditional machine learning approaches, which usually require sophisticated coding skills. Although quite appealing, one should take cautious since its responses can be deceptive and fallacious.

Our sample includes XXX copies of earnings call transcripts of S&P500 companies ranging from 2013 through 2023. A conference call usually starts with management presentations through which managers provide prearranged interpretations of the firm's performance during the quarter, followed by Q&A sections in which the analysts can request further details by asking questions. The Q&A sections of earnings conference calls provide "two-way" interactions which allow us to observe managers' disclosure strategies (Bushee and Huang, 2024) in real-time communications. Our analysis is based on the Q&A sections and utilizes a consecutive conversation between an analyst and the management as the research unit.

Using a manually coded dataset of 200 Q&A sections as the benchmark, we developed a few-shot prompt that involves three tasks. Task 1 instructs ChatGPT to identify managers' non-answers. To gain deeper insights into how managers signal their reluctance to disclose information during an earnings call, Task2 further asks ChatGPT to classify the non-responses into five categories including "Refusal", "Lack of Info", "Legal Affairs", "Recall", and "Irrelevant". Task 3 further instructs ChatGPT4-turbo to evaluate managers' non-answers on a scale of 0-10 per three effective communication guidelines from Gricean Norms. With the responses from ChatGPT, we construct measures of managers' non-responses both on the Q&A-section level and the call level.

We perform a battery of analyses to validate the appropriateness of the measures for managers' non-responses. Initially, we require ChatGPT to provide excerpts from the research unit to substantiate its identifications. We manually reviewed 200 copies of these responses to verify their accuracy. As can be seen from Appendix A Table???: To our best understanding, the excerpts extracted by ChatGPT can support its selections of managers' non-responses. This requirement also demonstrates ChatGPT's advantages over traditional natural language process tools about the interpretability of its results. Second, we find that Q&A sections with managers' non-responses tend to have lower scores in terms of the three aspects of Gricean Norms, indicating ChatGPT is consistent with its evaluations. Third, we compare ChatGPT's responses with another dataset of XXX human-annotated labels. The results show that ChatGPT's responses achieve a precision of up to 80% and an accuracy of up to 75%, further confirming the validity of our prompt. Finally, we look into the narrative features of conversations between analysts and the management, firms' earnings performances as well as the personalities of managers and analysts to conduct granular analysis on the Q&A-section level. We find that questions with more negative tone or financial numbers, and inquiring about firms' R&D activities or M&A affairs are more likely to be left unanswered by managers. Questions from analysts who are more pessimistic and less experienced are more likely to be refused. Managers who are over-confident and engage in higher levels of earnings management are more likely to provide non-responses.

Subsequently, utilizing the new measure for managers' non-responses, we try to explore the underlying motivations that may compel managers to withhold information despite explicit requests from analysts. Building on prior findings in the voluntary disclosure literature, we focus on the most frequently investigated managerial incentives including proprietary costs, uncertainty about information endowment, and unfavorable bad news, investigating whether those factors

identified can provide managers with incentives to withhold information by refusing to answer analysts' questions during a conference call. In line with proprietary costs preventing managers from making disclosures, we find managers from firms in highly competitive industries are more likely to withhold information. Aligned with managers' conservative disclosure strategies in the presence of heightened uncertainty, our findings indicate that managers who provide more uncertainty-related disclosures in their presentations are more inclined to avoid analysts' questions. Consistent with existing evidence on managers' tendency to withhold bad news that, if revealed, can negatively impact the firm stock price and exacerbate managers' career concerns, our findings indicate that poor earnings performance is positively associated with the likelihood of managers providing non-responses.

In the next, we investigate the short-term consequences of managers' non-responses by examining their impact on analysts' forecasts and stock market reactions. Despite the multiple motivations of managers' non-responses, such non-responses are likely to exacerbate uncertainties regarding a firm's value by reducing the availability of critical information. Therefore, we predict that these non-responses can impact information processors, including analysts and investors. First, we investigate whether managers' non-responses can affect analysts' revisions and forecasts after the call (Alternative version: we examine whether managers' non-responses can affect the precision and accuracy of analysts' forecasts released 15 days after the earnings call). Given that prior literature posits that analysts revise their EPS forecasts when new information arrives, we predict managers' non-responses during the conference calls would convey a negative signal, thus leading to more analysts' revisions. Consistent with this assumption, we find that managers' non-responses are positively related to analysts' revisions in 10 (or 15?) days after the call. Additionally, we find managers' non-responses are positively associated with analysts' downward revisions while negatively associated with analysts' upward revisions. This result reveals that analysts perceive managers' non-responses as bad signals and incorporate the message into their forecasts. Second, we examine the effect of managers' non-responses on two more specific measures of analysts' forecasts: forecast errors and dispersions. Similarly, we find that managers' non-responses during a conference call will lead to an increase in analyst forecast errors and forecast dispersions, consistent with prior research finding that managers' answers during earnings conference calls serve as a crucial information channel for analysts (Matsumoto et al., 2011, Huang et al., 2018). Third, examine the market reactions to the earnings call. If managers' non-responses

are interpreted as a signal of strategic withholding of negative information, we would expect to observe more negative market reactions in response to these non-responses. In line with our prediction, we document a negative association between managers' non-responses and cumulative abnormal return over the window [-1,1] surrounding the earnings call date.

To explore the long-term effect of managers' non-responses, we examine the relationship between managers' non-responses and future performance. On the one hand, managers' refusal to respond, particularly in face of intense public scrutiny during an earnings call, may signal a strong motivation to withhold negative information, indicating potential poor future earnings performance. On the other hand, managers may evade analysts' questions due to the protection of proprietary information, such as R&D progress or new product promotion. Find a negative/positive relation between managers' non-responses and future earnings performances, indicating that... Plays a dominant effect.

This paper contributes to the literature that investigates the motivations and consequences of managers' non-responses during a conference call. Despite the importance and ubiquity of managers' withholding information, measurement challenges make it difficult to evaluate its effect on the capital market. Specifically constrained by the inherent intricacies of human languages, quantitatively capturing managers' non-responses remains a hard endeavor. Three precedent papers undertook preliminary explorations regarding this problem. Hollander et al. (2010) investigate whether managers withhold information from the public by manually reviewing and labeling the call scripts of publicly listed U.S. firms between Jan 1 and Dec 31 of 2004. Gow et al. (2021) develop the measures of non-responses by crafting a set of regular functions. Barth et al. (2022) identified 1364 trigrams which indicates nonanswers by adopting a supervised machine learning framework on a large training set of questions and answers from firms' earnings calls. These papers typically adopt rule-based methods to quantify textual data, which is primarily effective at identifying non-responses that are explicit and straightforward but are limited in detecting more subtle or euphemistic occasions where managers evade analysts' questions. By taking advantage of recent advancements in natural language process techniques, our research provides a more comprehensive measure for managers' non-responses, which can be used in future research to explore managers' motivations for hiding information or the economic consequences of managers' lack of specific information.

On top of that, this paper adds to the burgeoning literature regarding the applications of AI, NLP techniques, and large language models in research in accounting and finance (Kim and Nikolaev,2023; Bernard et al.,2023…….). By introducing Chat-GPT's evaluations of managers' answers from three aspects of Gricean Norms—clarity, relevance, and quantity of new information—our method offers enhanced flexibility in addressing the complexities and ambiguities of textual data, offering guidance for similar research that utilizes large language models to conduct textual analysis.

Beyond simply identifying instances of non-response, we additionally classify them into five distinct categories, providing deeper insights into both the nature of these non-responses and their underlying causes. ChatGPT allows for a granular analysis of managers' non-responses, enabling us to link different types of non-responses to their perceived outcomes.

2. Literature reviews

2.1 Managers non-responses during earnings calls

The adverse selection theory in voluntary disclosures predicts a full disclosure equilibrium of firm information in the capital market (Milgrom, 1981; Grossman and Hart 1980; Grossman,1981). However, it is common to see managers withholding information in the real world. The considerable discrepancy between theory and practice has promoted numerous studies to determine the underlying motivations of firms' non-disclosures. Existing research has identified several factors that influence managers' disclosure strategies, including proprietary costs (Verrecchia, 1983; Berger and Hann, 2007), poor earnings performance (Dye, 1985; Harper, 2003; Graham et al.,2005; Bloomfield, 2002; Kothari et al., 2009), information uncertainty (Dye, 1985; Jung and Kwon, 1988), career concerns (Nagar et al., 2003; Armstrong et al., 2010), and litigation risks (Skinner, 1997; Field et al., 2005; Houston et al., 2019). However, these studies are limited in fully capturing managers' deliberate withholding of information (Hollander et al., 2010). Chen et al. (2011) use firms' giving-up of earnings guidance as a proxy for the management's deliberate holding of information. They find that firms stop guidance because of poor prior performance, increased uncertainty and decreased informed investors. Hollander et al. (2010) manually reviewed and labelled 1194 copies of earnings call transcripts with managers' non-responses. Their results show that firm size, CEO's stock_based incentives, company age and performance, litigation risks and whether analysts are actively involved during the call's Q&A section are best predictors of managers' non-responses. Moreover, they find investors interpret silence negatively. Gow et al.

(2021) constructs a novel measure of managers' non-responses with linguistic analysis. They mainly investigate the motivations of managers' non-responses from earnings performance and proprietary costs separately, but remains quite limited on the consequence of managers' non-responses. Barth et al. (2023) identify 1364 trigrams signaling nonanswers in earnings calls with a supervised machine learning framework on a large training set of questions and answers. Their measures are significantly associated with lower cumulative abnormal stock return and higher implied volatility. The most recent paper pertinent to our research is a literature review by de Kok, 2024, in which he performs a case study using ChatGPT recognize nonanswers. His case provides a comprehensive description of nonanswers identified by ChatGPT. However, his prompts are different from ours in many ways. For instance, he takes a one-shot strategy without any examples, while we affiliate two distinct examples for Chat-GPT to better understand tasks. Moreover, he uses individual question-and-answer pairs as the unit of analysis, whereas we make the best use of ChatGPT's comprehension capabilities by providing it with the complete back-and-forth dialogues between analysts and managers, ensuring no implicit information around the context is overlooked.

2.2 Large language models and textual analysis

The advent of large language models, with their advanced abilities in understanding, summarizing, and generating texts, provides new methods to perform traditional textual analysis tasks. Several studies have explored the potential of using large language models as sophisticated research assistants for processing textual data in the fields of accounting and finance. Kim et al. (2023 a) apply 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. Kim et al. (2023b) leverage ChatGPT's ability to summarize MD&A and earnings conference calls to develop a novel measure of information "bloat," offering a direct assessment of information processing costs. Similarly, Zang et al. (2024) use GPT-2 and BERT after pre-training and fine-tuning, to develop a new measure of language predictability score (LPS) of MD&A by imitating the language ability of investors. They find that LPS outperforms the fog index, the bog index, and file size in explaining analysts' processing costs. Armstrong (2023) uses the fine-tuned ChatGPT model to create new measures of tax enforcement based on 10-Ks. He finds the model can actively identify ongoing IRS audits at a 96% accuracy rate compared to a tax researcher manually labeling the same disclosure. Bai et al. (2023) introduce a novel measure of information content 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. Yang (2023) leverages ChatGPT to design a predictive model for patent value. He finds LLM embeddings can capture many qualitative aspects of textual information previously unavailable to researchers. Hansen and Kazinnik (2023) evaluate both the zero-shot and fine-tuned ChatGPT-3 model to classify the policy stance of Federal Open Market Committee announcements. Their results show that ChatGPT models obtain the lowest numerical errors, the highest accuracy, and the highest measure of agreement relative to human classification compared with alternative NLP methods. Lopez-Lira and Tang (2024) document that ChatGPT can predict price movements using news headlines without direct financial training, suggesting the great potential of AI systems in benefiting information diffusion and decision-making. Li et al. (2024) employ ChatGPT to automatically analyze analysts' reports, detecting their views of corporate culture. Kirtac and Germano (2024) analyze and compare the performances of the GPT-3 model, BERT, FINBERT, and traditional McDonald dictionary on sentiment analysis tasks. Their results reveal that the GPT-3 model significantly outperforms the others. These pioneering papers conduct preliminary explorations into the application of large language models (LLMs), particularly generative large language models (GLLMs), to assist research in the accounting and finance domains. They demonstrate ChatGPT's significant potential in understanding large-scale textual data and summarizing abstract concepts for further quantitative and empirical analysis.

3. Research Design

3.1 Data and Sample

To start, we downloaded the earnings call transcripts of S&P500 firms from 2013 to 2023 from the S&P Global Capital-IQ database(sample number). Then we exclude firms in the financial and utility sectors (sample number). If the firm held more than one conference call, only the most recent one is kept. To calculate measures for analysts' forecasts, we require there to be at least 2 forecasters in each firm quarter. Meanwhile, for each analyst, we retain the earliest forecasts made within 15 days following the earnings calls and the latest forecasts made within 90 days before the earnings calls. We further eliminate firms for which fundamental financial features, disclosure characteristics, and the stock market are unavailable. During an earnings call, the management would make presentations regarding the firms' performances of the current quarter, followed by the Q&A exchanges between the analysts and managers. In real-situation communication, one

should interpret sentences in conjunction with the surrounding contexts. So, distinct from previous studies that code on single question-answer pairs, we use the entire consecutive ongoing back-and-forth conversations between an analyst and the firm's managers as our research unit to avoid missing information in surrounding contexts. The dataset from Capital-IQ provides us with the convenience of identifying the point in time where one analyst-manager interaction ends and the next one begins. Our final sample contains xxx transcripts of quarterly conference calls from xxx firms, consisting of Q&A exchanges between analysts and managers. Table 1 shows the sample selection process.

[Insert Table1 Sample Selection]

3.2 **Variable constructions**

3.2.1 Managers' non-responses

Model set-up and prompt development. We utilize the Chat-GPT4-turbo model which supports up to 128K tokens per request to avoid any truncation of our transcripts. We get access to the model through the API key provided by the OpenAI ChatCompletion endpoint. A challenge associated with working with private LLMs is result replicability, as the outputs generated by Chat-GPT can vary significantly with each instance. To mitigate this concern as much as possible, we set the 'temperature', 'frequency_penalty', and 'presence_penalty' parameters of the model to 0. Meanwhile, we set the max_tokens parameter to 5000, large enough to prevent the truncation of outputs in the Q&A pairs.

After setting up and configuring the model, we manually code 300 random Q&A sections where the section is labeled as a non-response and meanwhile, the corresponding questions and answers are retrieved if any questions from the analyst were left unanswered through the interactions between an analyst and the management. We adopt two strategies to code the Q&A exchanges. The first approach is straightforward, where managers directly refuse to answer questions, either with or without a justification. The second approach relies on researchers' subjective interpretation to identify more subtle instances of managers' non-answers, such as when responses are incomplete or not directly relevant. Enlightened by de Kok (2023), we employ a few-shot approach to formulate the prompt, using the coded 300 Q&A sections as a benchmark,

and iteratively refining it until the prompt's accuracy surpasses 70%. As expected, we found that ChatGPT achieved a higher recognition rate for non-responses identified using the first approach.

Generation. In the finally settled prompt, we explicitly instruct the model to perform three consecutive tasks, affiliating two examples for guidance. Task 1 requires ChatGPT to identify managers' non-responses and extract relevant texts. Based on the non-responses identified in Task1, Task2 further instructs ChatGPT to classify the non-responses into five categories including "Refusal", if managers directly refuse to answer a question without giving any reasons; "Lack of Info", if managers justify their non-responses by the lack of information; "Legal Affairs", if managers can't respond due to legal barriers; "Recall", if managers explicitly indicate they will revisit this question sometime in the future[1]; and "Irrelevant", if the management rambled on with irrelevant nonsense to dodge the questions. We ask Chat-GPT to return "Null" if no such non-responses are identified in Task 1. To validate Chat-GPT's responses for Task1, we implement an evaluation system in which ChatGPT is instructed to assess managers' responses based on three aspects of Gricean norms: the quantity of new information (New_info), relevance (Relevance), and clarity (Clarity). Ratings are assigned on a scale of 0 to 10, with higher scores indicating responses that are more informative, relevant, and clear. To make our results traceable and verifiable, we instruct ChatGPT to output its responses in JSON format with 6 keys including a mark whether there are any no responses, an excerpt of conversations in which non-responses were identified, evaluations of managers' answers from three aspects mentioned above[2] and the categories for non-responses. A description of the responses shows that the ChatGPT4-turbo model identified ???% of all Q&A exchanges between an analyst and managers during earnings conferences. At the call-level sample, ???% out of the XXX quarterly conference calls include managers' non-responses. There are on average xxx non-responses for every Q&A section and xxx questions with non-responses for every call.

Measure calculation. We construct two measures for managers' non-response at the Q&A-section level-----an indicator (*NOA*) that equals 1 if at least one non-response is recognized during a Q&A exchange and 0 otherwise. If more than one question from a Q&A section is retrieved as unanswered, we manually look into such sections for a further check. So, we also get a specific

[1] Answers like "*The earnings call is not the right place for — to make major product announcement*.", "*So, yeah, we'll get there. We will provide it later*".
[2] Quantity of incremental information, Relevance and Clarity.

number of non-responses for every Q&A exchange (*NOA_number*). Additionally, we compute two non-response measures at the call level. One is an indicator for managers' non-responses at the calls' level (*NOA_Call*), which equals 1 if any non-responses are recognized during an earnings call and 0 otherwise; the other is defined as the rate of conversations with non-responses relative to the total number of conversations during a call (*NOA-Rate*).

3.2.2 Analyst revisions

We assume it takes analysts no more than 15 days from the conference day (day 0) to make such revisions. So, we only retain analysts' forecasts within 15 days following a conference call. On the firm level, we construct three measures for analyst revisions separately including the total number of revisions within 15 days after the call (*Revision*), the ratio of upward revisions to total revisions (*Up_Rev*), and the ratio of downward revisions to total revisions (*Down_Rev*).

3.2.3 Analyst forcastes

We measure forecast error for each firm quarter as the absolute value of the difference between the mean of analysts' estimates and actual earnings per share for a given quarter. We measure dispersion for each firm quarter as the standard deviation of analysts' forecast estimates. The final sample consists of xxx observations for xxx firms. Following Cheng et al. (2013), We use the negative forecast width (the magnitude of the range for range forecasts and zero for point forecasts) to measure individual forecast precision.

3.2.4 Stock market reaction

We measure the post-conference call stock return (*CAR*) as the cumulative abnormal return over the period starting on day +1 after the quarter's earnings announcement and ending on day +3 after the earnings announcement data for the quarter's earnings announcement.

3.2.5 Firms performance

In order to examine the economic consequences of managers' non-responses in the long run, we use the earnings-per-share (*EPS*) of next year to proxy for firms' earnings performances.

3.3 Model designs

We validate our measures for managers' non-responses by using Eq. (1), with i and t denoting the firm and calendar quarter, respectively:

$$DVs = \beta_0 + \beta_1 NOA_{i,t} + \beta_2 Controls_{i,t} + Firm\ Dummy + Year\ Dummy + \varepsilon_{i,t} \quad (1)$$

Following prior research (Brown et al., 2001; Bowen et al.,2002; Barth et al., 2023; Bushee and Huang, 2024), we add the following control variables: firm size ($Size_{i,t}$),earnings surprise ($EarSup_{i,t}$),stock price ($StoPer_{i,t}$),earnings growth ($Growth_{i,t}$),managerial presentation tone ($Tone_{i,t}$), the complexity of managerial presentations ($Complex_{i,t}$) ,forward looking information disclosures in the managerial presentations ($Forward_{i,t}$).