Submitted to *Information Systems Research*
manuscript (Please, provide the manuscript number!)

Authors are encouraged to submit new papers to INFORMS journals by means of a style file template, which includes the journal title. However, use of a template does not certify that the paper has been accepted for publication in the named journal. INFORMS journal templates are for the exclusive purpose of submitting to an INFORMS journal and should not be used to distribute the papers in print or online or to submit the papers to another publication.

Conversation Analytics: Can Machines Read between the Lines in Real-Time Strategic Conversations?

Yanzhen Chen
HKUST, imyanzhen@ust.hk

Huaxia Rui
University of Rochester, huaxia.rui@simon.rochester.edu

Andrew Whinston
University of Texas at Austin, abw@utexas.edu

Strategic conversations involve one party with an informational advantage and another with an interest in the information. This paper proposes machine-learning-based methods to quantify the evasiveness and incoherence of the more-informed party during real-time strategic conversations. To demonstrate the effectiveness of these methods in a real-world setting, we consider the question-and-answer sessions of earnings conference calls during which managers face scrutinizing questions from analysts. Being reluctant to disclose adverse information, managers may resort to evasive answers and sometimes respond less coherently than they otherwise would. Using data from the earnings calls of S&P 500 companies from 2006 to 2018, we show that the proposed measures predict worse next-quarter earnings. The stock market also perceives incoherence as a negative signal. This paper contributes methodologically to business analytics by developing machine-learning methods to extract behavioral cues from real-time strategic conversations. We believe the wide adoption of these tools can increase the efficiency of various markets and institutions where real-time strategic conversations routinely occur, which ultimately benefits business and society.

Key words: business analytics; artificial intelligence; machine learning; conference calls

History:

1. Introduction

We humans are good at, and often proud of our ability to read between the lines — that we can infer information implicitly conveyed and sometimes inadvertently revealed by a speaker. Being able to discern subtleties in conversations is such a unique human skill that we generally consider it an important aspect of intelligence that machines can hardly emulate. This, however, is changing, as we witness a new technology revolution.

Thanks to recent innovations in learning algorithms, the growing computing power, as well as the increasing data availability, the performance of artificial intelligence (AI) has rapidly advanced over the past two decades. Not surprisingly, companies in almost all industries are racing to augment their human intelligence with machine intelligence. Indeed, if we think of human intuitions and experiences as algorithms embodied in biological rather than artificial neurons, it is only logical that AI algorithms, equipped with faster computation and more data, may eventually surpass where our intuitions and experiences can take us. For example, a Wall Street Journal article[1] reported how MetLife improved its customer experience scores by teaching humans to be *more human* using AI. By assessing human performance on social skills such as empathy and patience, the AI software monitors customer service conversations so as to guide call-center agents in real time as they engage with customers. Even for professional investors, as Two Sigma co-founder David Siegel argued, "*the time will come that no human investment manager will be able to beat the computer.*"

Inspired by this new technology revolution, we use algorithms to evaluate real-time strategic conversations in which one party has an informational advantage (henceforth the more-informed party) and the other party (henceforth the less-informed party) has an interest in such information. The two primary contexts where such real-time strategic conversations routinely occur are the principal-agent setting and the labor-market setting, as we illustrate in Table 1. For example, in a principal-agent setting, the agent is the more-informed party and the principal is the less-informed party. Senior managers of a firm need to regularly report firm performance to investors who are the principal, and such reports are often strategic and real-time (e.g., conference calls).

[1] `https://www.wsj.com/articles/call-center-agents-get-a-human-touch-1528984801`

Table 1 Examples of Real-Time Strategic Conversations

	More-informed party	**Less-informed party**
Principal-agent setting	Corporate managers	Investors (through analysts)
Labor-market setting	Election candidates, job seekers	Citizens (through journalists), employers (through recruitment committee)

The objective of this research is to measure the evasiveness and the incoherence of the more-informed party, which we believe are particularly meaningful in such contexts because of the following three elements: the incentives of the two parties are not aligned; the more-informed party usually has certain flexibilities regarding whether, to what extent, and how to reveal information; and the more-informed party often must improvise in answering unexpected questions due to the real-time nature of such conversations. As social psychologists (see Goffman 1959) have long recognized, there are two types of expressions in social interactions: expressions *given* and expressions *given off*. Expressions given are the verbal or non-verbal signals that we intend for others to receive, while expressions given off are those we do not intend for others to receive. Traditionally, the less-informed party relies on their shrewdness and acumen to detect expression given off. This paper is based on the premise that the performance of modern machine-learning algorithms has reached the tipping point whereby AI is able to detect expression given off and may even do so better than average humans in certain contexts.

The specific context for our empirical evaluation is earnings conference calls in which the managers of public companies discuss the financial results of a reporting period. Unlike the management presentation part of an earnings conference call, which is typically scripted and well prepared, the question-and-answer (Q&A) part of the call is conversational. Managers are required to improvise in answering the questions posed to them, which are often hard to predict. They must also do so under time constraints, without real-time support from their staff members. Such a "conversational dance" between managers and analysts is an example of real-time strategic conversation. An appealing feature of this empirical context is that we can study the implications of evasiveness and incoherence using objective quantities such as stock market reactions and next-quarter earnings. Analyses based on earnings conference calls of the Standard and Poor's 500 (S&P 500) companies from February 2006 to December 2018 suggest that both evasiveness and incoherence forecast worse next-quarter earnings and that incoherence also predicts worse next-day abnormal stock returns.

2. Literature and Theoretical Foundation

The analysis of conversations focuses on recorded or naturally occurring talk-in-interaction (Hutchby and Wooffitt 2008). Traditionally, it is mostly a qualitative research method for talk, as in Silverman (2020). In the age of business analytics, using quantitative methods for the analysis of conversations, which we refer to as conversation analytics, is a promising line of research (Meadows and O'Brien 2020). Analogous to the distinction between cooperative games and non-cooperative games, we can roughly categorize conversations into cooperative conversations and non-cooperative conversations. An example of the former is customer service conversation for which the goal of analytics is to improve customer satisfaction. For example, studies suggest that the use of concrete language and personal pronouns can enhance customer satisfaction and boost sales in customer service interactions (Li et al. 2020, Packard et al. 2018). In particular, Li et al. (2020) show that these strategies inject warmth and competence into various stages of a conversation and highlight the important role of temporal flow in conversational outcomes. For non-cooperative conversations such as negotiations and persuasions, the goal of analytics is different. For instance, Zhou et al. (2019) focus on identifying optimal bargaining strategies in price negotiations. By examining past negotiation patterns, their algorithm predicts the most advantageous tactic for a seller in real-time, maximizing final prices and providing general rhetoric suggestions. An interesting but distantly related literature stream is at the intersection of reasoning and text generation. One approach is to map language to domain-specific logical forms, followed by selecting the next dialogue action (Cuayáhuitl and Lemon 2015, Keizer et al. 2017). Another is to employ an end-to-end design for a domain-independent, data-driven comprehension, reasoning, and text generation. For example, Lewis et al. (2017) propose end-to-end learning for natural language negotiations using recurrent neural networks and reinforcement learning, addressing both the linguistic and the strategic reasoning aspects.

Our research differs from the prior literature by studying a unique but also commonly observed type of non-cooperative conversations and by focusing on quantifying subtle behavioral cues of conversation participants. These cues, evasiveness and incoherence, are particularly relevant for real-time strategic conversations and draw upon both academic literature and established conventional wisdom. Conceptually, we categorize an answer as evasive if it is **incomplete** or **irrelevant**. For example, the more-informed party may use shifting and refocusing tactics by barely commenting on a thorny aspect, dwelling instead on favorable or irrelevant issues. Although intuitive, such a strategy has been formally analyzed in the theoretical work of Crawford and Sobel (1982) and a class of game-theory models sometimes referred to as "persuasion games." In these models, the sender, who has an information advantage, can withhold information but cannot lie because the receiver can verify any information that the sender reports. Milgrom and Roberts (1986) showed that under mild conditions, the receiver adopts a unique equilibrium strategy—the "*assume the worst*" strategy—in which the receiver makes the inference that leads to the least favorable decision for the sender, conditional on the information available. This result confirms the intuition that an evasive strategy is negatively interpreted by a rational receiver.

In our specific empirical context, such a strategy is often embodied in the management obfuscation hypothesis. This hypothesis argues that managers obfuscate information when firm performance is unsatisfactory so that the processing cost of adverse information increases, which may delay or even prevent an adverse stock market reaction to the information. Although outright lying and silence during the Q&A part of an earnings call are out of the question, answering in an evasive way that clouds, disguises, or even distorts private information remains a feasible option for some managers (Khalmetski et al. 2017). Indeed, there is empirical evidence suggesting that managers sometimes present less relevant information to avoid giving a direct answer (Larcker and Zakolyukina 2012). Naturally, analysts and investors are interested in evaluating behavioral and linguistic cues that suggest evasiveness. In fact, the value proposition of some companies is precisely their expertise in uncovering deceptive behavior, including detecting executives' evasive responses. For example, Business Intelligence Advisors (BIA), a hedge-fund consulting firm, hires former Central Intelligence Agency employees to analyze language clues (Javers 2010). Specialists at BIA try to gauge whether managers directly answer questions or, instead, dance around difficult matters. An important type of behavioral cue that they look for is "*management replies larded with irrelevant specifics.*" For example, by analyzing how managers of UTStarcom diverted questions during one earnings call, BIA successfully predicted the company's profitability.[2] In accounting practice, the Public Company Accounting Oversight Board (PCAOB) has also emphasized that auditors should consider directly observing earnings calls or reading their transcripts as a part of the procedure of assessing material misstatement (PCAOB 2008). In addition to revealing information about firm fundamentals, how definitively and directly managers answer questions during conference calls also seems to influence stock price movement. In one anecdotal account, Jim Cramer of CNBC credited Google's share surge after its conference call in July 2015 to its new CFO, Ruth Porat, and her being more down-to-earth when answering questions than her predecessors.[3]

[2] After analyzing UTStarcom's August 2, 2005 conference call, BIA rated UTStarcom's second quarter conference call as having a "medium high level of concern" and highlighted that the communication from its managers "avoids providing information [and] indicates underlying concern." The BIA team therefore warned its client that skirting the question would point to poor third-quarter results. This BIA report flagged the concerns with revenue recognition before the stock price shrank to two-thirds of its former value on the day after the next earnings announcement. The main reason that the BIA analysts considered UTStarcom Executive Vice President and Chief Operating Officer Michael Sophie to have avoided commenting on revenue recognition is that they found the following answer to be a "detour statement." During the Q&A part of the earnings call, Mike Ounjian, an analyst with Credit Suisse First Boston, asked, "Are there any issues related to recognizing revenues on these?" Michael and the interim CFO tap-danced around the subject by responding, "Yes, with the backlog, the vast majority of the wireless backlog is clearly PAS [an acronym for one of the company's products, Personal Access System]. I think you saw the announcement at the end of June where we announced on the PAS infrastructure orders in China. And again, it's just the timing of deployment and achieving final acceptance; we've also got some CDMA [an acronym for a type of mobile phone standard] to a lesser extent in the backlog. ... But Q3 is clearly a little more handset-oriented than we would typically run." For more examples of BIA analysts evaluating how managers handle questions, please refer to the Wall Street Journal article at https://www.wsj.com/articles/SB115110330795289453.

[3] `http://www.cnbc.com/2015/07/20/cramer-ruth-porat-key-to-googles-success.html.`

Our incoherence measure is motivated by theories and evidence from the psychology literature on deception. According to this literature, deception induces cognitive load due to people's need to avoid contradicting statements that they made previously or facts that the observer may know about. As a result, deceptive accounts appear less coherent (Hauch et al. 2015). In our empirical context, managers who hide or distort some adverse information experience a cognitive load that may lead to less coherent answers. To construct the incoherence measure, we quantify how smoothly a manager's thoughts flow within the answer. This is an inherently challenging task, if possible at all, because thought is a very abstract construct. In particular, we rely on a state-of-the-art deep-learning model that is native to coherence in text tasks (Lan et al. 2019) to measure the smoothness of the thought flow.

3. Measure Constructions

3.1. Evasiveness

Our measure of evasiveness is designed for conversational data and is based on statistical language modeling. It can be computed in real time without human involvement and thus is scalable. Moreover, the measure is domain-agnostic because the underlying topic space is not predefined. Before describing our approach, we first review how the current business literature measures concepts that are related to our notion of evasiveness. A popular one is informativeness, which is often measured by the number of words in, or the length of, a statement (Miller 2010, Ertugrul et al. 2017, Loughran and McDonald 2014). We believe that such a crude measure is too noisy to accurately capture the notion of evasiveness because evasive answers can be long and direct answers can be short. Two other closely related concepts are readability and vagueness. Readability is often measured by word complexity (Miller 2010, Lehavy et al. 2011, Lawrence 2013) and the presence of grammatical errors (Hwang and Kim 2017), and vagueness is typically measured by the proportion of words from a predefined vagueness lexicon, a catalog of words that indicate uncertainty (Ertugrul et al. 2017, Dzieliński et al. 2017). The reasoning for this approach is that function words such as "may" and "could" reflect the attitude or mood of the speaker. A significant drawback of this approach, however, is that a person's use of such function words may also result from the desire to be polite. According to one of the most influential politeness theories (Brown et al. 1987), speaking less directly is a strategy to mitigate face threats directed at the listener during social interaction. Thus, the use of words from the vagueness lexicon may reflect the speaker's desire to reduce face threats rather than indicating an evasive strategy.

Our approach is fundamentally different from the above-mentioned approaches. Rather than measuring evasiveness at the lexical level, we measure it at the semantic level, and rather than examining only the text of an answer, we evaluate its evasiveness conditional on the text of the question addressed by the answer. Of course, the challenge is to algorithmically evaluate how the content of an answer aligns with the content of the question addressed by the answer.

Conceptually, some literature has defined evasiveness in terms of relevance (Gabrielsen et al. 2020) and incompleteness (Bull 2003). For example, in the context of political interviews, Plüss and Piwek (2016) considered the relevance and completeness of an answer to be its fundamental features. Similarly, Clayman (2001) identified incomplete answers as a negative dimension of resistance, pointing to a speaker's shifting topics and refusing to answer as covert tactics for evading questions during interviews. In the context of political communication, Bull (2003) identified completeness as a major category in its well-known typologies of equivocation in political interviews. Following these studies, we measure evasiveness from the perspective of incompleteness and irrelevance. Rather than doing so manually, we take advantage of machine learning by automatically computing the mismatch between the topics covered in a question and the topics covered in the corresponding answer. The rationale is that irrelevance can be considered as using topics absent in the question, while incompleteness can be considered as answering without touching or emphasizing the topics in the question.

Operationally, we adopt the Latent Dirichlet Allocation (LDA) algorithm (Blei et al. 2003), an unsupervised algorithm that relies on a set of parametric assumptions and the co-occurring patterns of words in different documents to uncover the latent topics in each document. By analyzing the questions and answers using LDA, we summarize each question and each answer using a dense topic vector where an element of the numerical vector indicates the weight of the corresponding topic covered by the question or answer. Once we represent each question or answer as a topic vector from the same topic space, we assess the dissimilarity between the topic vector of a question and that of the corresponding answer to evaluate how evasive the answer is in terms of properly covering each topic in the question. The LDA model has been used widely to assess content similarity between documents. For example, Shi et al. (2016) represented a company's textual description on CrunchBase as a topic vector using LDA, in order to compute a business proximity measure for any pair of companies. They then validated the measure using an application of mergers and acquisitions in the U.S. high technology industry. Chen et al. (2021) compared an executive's job description and the executive's tweets using LDA to gauge the job relevance of those tweets, which they then used to construct a measure of social media personal branding. They then studied whether social media personal branding improved a job candidate's labor market performance in the context of executive employment and compensation.

Technically, the LDA model assumes a two-step document-generating process using a Dirichlet distribution for topic proportion per document and another Dirichlet distribution for topic generation per corpus, where a topic is defined as a distribution over a fixed vocabulary and each document is assumed to cover a mixture of topics. For each document, LDA draws a topic proportion from the first Dirichlet distribution. Then, for each word of that document, it first draws a topic based on the realized topic proportion and draws the word based on the topic definition, which, shared by all documents in the corpus, is drawn from the second Dirichlet distribution. Once we have applied the algorithm to a collection of textual documents, the algorithm estimates all of the model parameters, especially the topic vector for each document. In our context, if the topic vector of the j-th question for conference call i is $T^Q_{i,j}$ and the topic vector of the corresponding answer is $T^A_{i,j}$, we calculate $e_{i,j}$, the evasiveness of the j-th answer, as the cosine distance of the two corresponding topic vectors of the question and answer, i.e.,

$$e_{i,j} \equiv 1 - \frac{T^Q_{i,j} \cdot T^A_{i,j}}{||T^Q_{i,j}|| \cdot ||T^A_{i,j}||}.$$

The evasiveness of an entire conversation is then calculated by averaging $e_{i,j}$ for all of the questions in conference call i. To reduce noise, we ignore trivial questions, which are defined as routine greetings or checks (e.g., a check for connections during a conference call). To overcome the limitation of LDA for short documents, we also apply the following heuristic rule: If the response from the more-informed party is short and definite, i.e., fewer than five words and containing words such as *yes*, *sure*, *correct*, *you bet*, *yeah*, *right* or *no*, we set its evasiveness measure to 0.

3.2. Incoherence

Linguistic research on text coherence and cohesion has exploited the connection of neighboring text for measurement (Halliday and Hasan 2014, Grosz et al. 1995, Foltz et al. 1998). Recent advancements in natural language processing (NLP) often involve training deep learning models by predicting words in adjacent sentences (Kiros et al. 2015, Devlin et al. 2019), predicting following sentences (Gan et al. 2016) and discourse markers (Jernite et al. 2017). For example, Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al. 2019) has stirred the machine-learning community by presenting state-of-the-art results in a wide variety of NLP tasks.

Our incoherence measure is based on a variant of BERT, namely the lite BERT for self-supervised learning of language representations (ALBERT), which aims to address the "ineffectiveness of the next sentence prediction (NSP) loss proposed in the original BERT" (Yang et al. 2019, Liu et al. 2019). The ALBERT model introduces a loss for sentence-order prediction (SOP), which focuses on inter-sentence coherence to boost the performance of BERT. Different from BERT, which combines topic prediction and coherence prediction in constructing loss, ALBERT focuses primarily on coherence, which guides the model to "learn finer-grained distinction about discourse-level coherence properties" (Lan et al. 2019), resulting in a drastic improvement in the capturing of text order. Specifically, ALBERT models achieve this goal by using natural sequences and swapping ones as positive and negative examples rather than by treating sentences from different documents as negative examples. Because ALBERT specializes in sentence-order prediction, with a focus on inter-sentence coherence (Lan et al. 2019), we build our incoherence measure based on this algorithm. Our use of the BERT framework is native because the training of BERT is based on a combined loss function of masked language modeling (MLM) and NSP, which capture coherence and cohesion on the word and sentence levels, respectively. In other words, measuring incoherence is a native task for the BERT-family models. Consequently, this fact not only improves the interpretability of our measure of incoherence but also offers face validity at the algorithmic design level, which also differentiates our use of BERT from other BERT applications.

Furthermore, we modify the loss using human perceptions rather than crude and artificially constructed negative examples, thus, extend ALBERT to integrate human intelligence with inter-sentence coherence. To be specific, we take the last layer hidden state of the first token of the sequence (the CLS token) and further process it by a linear layer followed by a softmax activation function. The linear layer weights are learned from the perceived coherence (classification) objective during fine-tuning. We calculate incoherence as 1 minus the output probability of being coherent.

4. Measure Validations

4.1. Evasiveness

Before validating the proposed evasiveness measure, we present examples of conversations that score high and low in terms of evasiveness according to our measure in Table 2. The first example scores as highly evasive and is from Boeing Company's Q2 2012 earnings calls, with an evasiveness score of 0.929. In this example, the analyst asked for a comparison of the pricing pressure between the MAX and NGs models. The chief executive officer (CEO) responded with his outlook on the company's production system and supply chain. The second and third examples show forthright responses (evasiveness score = 0 in both examples). In the second example, the response was short and definite, including the word "correct"; therefore, we set its evasiveness score to 0. The third example shows a direct response, according to its perfect match of topics between the question and answer.

We validate the proposed evasiveness measure using two types of strategic conversations: political interviews and the Q&A part of conference calls. In each setting, we compare our algorithm-generated evasiveness measure with human perception.

First, we use a public dataset of 53 annotated segments from six political interviews [4] available on public channels such as BBC News, CNN and YouTube. These political interviews are between Brodie

[4] $http://mcs.open.ac.uk/nlg/non-cooperation/$

Table 2 Evasive and Direct Examples Identified Using Our Evasiveness Measure

Evasive Example **(evasiveness = 0.929)**	**Direct Answer Case 1** **Supervised by Rules** **(evasiveness = 0)**	**Direct Answer Case 2** **by the naïve LDA** **(evasiveness = 0)**
(Jul. 25, 2012, 3:40 PM ET The Boeing Company (BA)) **Joseph Nadol, JP Morgan Chase & Co.** Is the bigger issue the MAX? Or is it the pricing on the remaining NGs? **Gregory D. Smith** Well, I think it's just ... as I said, it's any time you're ramping down a program and then ramping up a new one, certainly, there's pricing considerations taken in that. Customer by customer, it's different. But again, I think we know where it is. We're working through our production system. We're working through our supply chain. And again, I think we've got good plans in place to address it.	**(Feb. 4, 2014, 12:24 PM ET Yum! Brands, Inc. (YUM))** **Jason West, Deutsche Bank** Sorry for beating a dead horse again on the sort of current trend, but just want to be clear that you guys are seeing an improvement in the business, particularly at KFC, that's independent of the shift in Chinese New Year, because we know the New Year started about 10 days earlier this year but still didn't start till the end of the month. So I just want to be clear that you guys were able to measure that it's actually a real improvement, excluding that shift. **David Novak** That is correct.	**(Feb. 24, 2011, 2:10 PM ET Newmont Corporation (NEM))** **Patrick Chidley, HSBC Holdings** So, would we expect that in 2012 to come down fairly dramatically? **Brian Hill** I would certainly expect it to come down in 2012.

and Blair, Green and Miliband, O'Reilly and Hartman, Paxman and Osborne, Pym and Osborne and Shaw and Thatcher. Prior research (Plüss and Piwek 2016) used these interviews to study conversations of various levels of cooperation between the interviewer and the interviewee. These interviews are selected because none of the exchanges broke down or turned into a debate. They are rated by seven annotators in terms of irrelevance and incompleteness. Because our proposed measure captures both aspects, we test its consistency from both perspectives and as a whole. To consolidate the opinions of multiple raters, we use both the average and the majority opinions. To test both the linear and general monotonic relationships between the two evasive measures, we examine the Pearson and Spearman correlations in each test. Table 3 reports the test results. Overall, our evasiveness measure is positively correlated with human perceptions of irrelevance and incompleteness, with statistical significance at the 5% level and correlation coefficients ranging from 0.29 to 0.48.

Table 3 Validations of the Machine-Learning-Based Evasiveness Measure and Human Perceptions

(N=53, p-values in brackets)

	Irrelevance		**Incompleteness**		**Both Aspects**	
	Average	**Majority**	**Average**	**Majority**	**Average**	**Majority**
Pearson	0.3846	0.3065	0.3657	0.3032	0.4568	0.4187
	(0.0045)	(0.0256)	(0.0071)	(0.0273)	(0.0006)	(0.0018)
Spearman	0.3661	0.2866	0.3932	0.3318	0.4754	0.4357
	(0.0070)	(0.0375)	(0.0036)	(0.0152)	(0.0003)	(0.0011)

Second, to better connect the measure validation with our specific empirical context, we recruit 30 undergraduate business majors from a large U.S. university and ask them to rate the degree of evasiveness of each answer on a scale from 0 to 9 for 335 pairs of questions and answers (161 conversations) from our data. Again, we examine both linear and rank-order correlations between our measure and human perceptions. Although this task is highly challenging, we nevertheless find a statistically significant correlation in both cases (Pearson correlation = 0.166, p = 0.036; Spearman correlation = 0.178, p = 0.024).

4.2. Incoherence

To validate our incoherence measure, we compare incoherence as measured by our method and as perceived by humans in various domains, including community-based question-answering sites

(CQAs), product reviews, business correspondence that may associated with a scandal, political correspondence that may associated with crises and persuasive writing in high-stakes exams, as well as our main context of conference calls.

For the first four domains, we use the high-quality, human-annotated coherence score data of Lai and Tetreault (2018). The paragraphs cover four themes: Q&A from *Yahoo Answer*, product reviews from *Yelp*, emails released by the State Department from Hillary Clinton's office and the Enron Corpus. Each domain has 1,200 paragraphs. After dropping 60 very short paragraphs, we are left with 4,740 paragraphs, each rated by 8 raters on coherence using a 3-point scale, from 1 (less coherent) to 3 (very coherent). Among the eight raters Lai and Tetreault (2018) recruited for each text, three are experts with linguistic annotation experience and the rest are recruited via Amazon Mechanical Turk.

Table 4 compares the ALBERT-based incoherence measure with human annotations for the corpus in various domains. We find that the ALBERT-based incoherence measure is significantly correlated with human perceptions, with a significance level of 1% in all four domains. In particular, our measure significantly outperforms the LSA-based incoherence measure (Foltz et al. 1998), which has correlations around -0.04. In summary, evidence from these four domains suggests that our incoherence measure accurately reflects the coherence of a text.

The fifth set of validation data is from the seminal works of Crossley and McNamara (2016), which aimed to understand human perception of coherence by analyzing linguistic experts' ratings of coherence and other dimensions of SAT argumentative essays. Similar to our context, these texts are high-stakes and persuasive in nature. The correlation between the average human ratings and our measure is -0.47, with p-values less than 0.01.

Finally, we directly validate our measure using a sample of earnings call data. We recruit 16 undergraduate business majors from a public university in the U.S. to rate 347 earnings call responses. Table 5 reports a significant and negative correlation between the student ratings of coherence and our measures.

In summary, we find that the deep-learning-based incoherence measure captures the human perception of incoherence well in various domains.

Table 4 Comparing Human Perceived Coherence and the Incoherence Measures Using Datasets from Four Domains

Domain	**Total data num.**	**Train pos. num.**	**Train neg num.**	**Batch size**	**Epoch num.**	**Test num**	**Corr Coef**
Yahoo	1,200	200	200	64	5	800	-0.472***
Enron	1,200	200	200	64	5	800	-0.469***
Clinton	1,200	200	200	64	5	800	-0.408***
Yelp	1,200	200	200	64	5	800	-0.413***

Note: *** $p< 0.01$, ** $p< 0.05$, * $p< 0.1$.

Table 5 Comparing Human-Perceived Coherence and the Incoherence Measures Using a Sample of Earnings Calls

Domain	**Total data num.**	**Train pos. num.**	**Train neg. num.**	**Batch size**	**Epoch num.**	**Test num.**	**Corr. Coef.**
Earnings Call	347	50	50	64	8	247	-0.411***

5. Business Application: Conference Calls

Inspired by the earnings call literature that suggests the informative role of real-time business conversation during earnings calls (Matsumoto et al. 2011), we analyze conference calls using our proposed measures of evasiveness and incoherence. Conference calls are held in conjunction with earnings announcements as a form of voluntary disclosure. These calls typically include a management presentation, during which managers interpret company performance, followed by a Q&A part, during which analysts may question those interpretations and request additional information. We use our proposed measures to evaluate the evasiveness and incoherence of the managers' responses to these questions and requests, and we investigate whether such information predicts next-day abnormal stock return and next-quarter earnings surprises.

Relevant research has primarily focused on word choices in earnings call language analyses, such as tone (Chen et al. 2018) and vagueness (Dzieliński et al. 2017). In addition, vagueness in 10-K statements has been linked to stricter loan contracts and a higher risk of price plunges (Ertugrul et al. 2017). Similarly, increased vagueness in Initial Public Offering (IPO) prospectus disclosures, as measured using a lexical approach, can lead to higher first-day returns and increased volatility (Loughran and McDonald 2013).

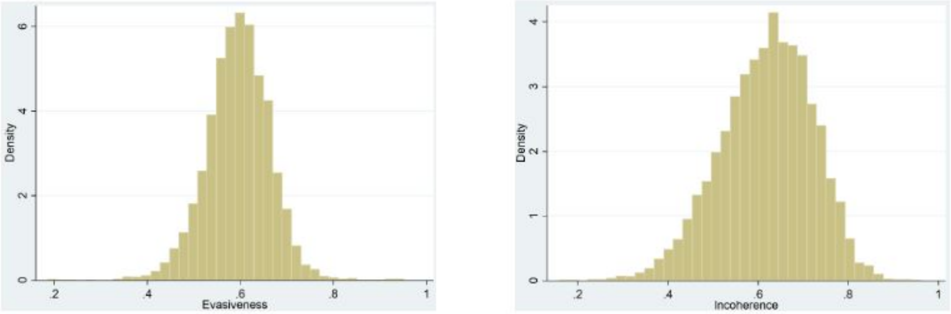
Unlike studies that have focused either on financial statements that were carefully prepared and released (Ertugrul et al. 2017, Loughran and McDonald 2014, 2013) or on managers' responses without relation to the corresponding questions (Dzieliński et al. 2017), our measures are designed to evaluate real-time conversations. By examining the dynamic interplay between managers and analysts, our research builds upon the existing body of work regarding the information intermediary roles of analysts, who uncover market-relevant information and enhance corporate disclosures (Huang et al. 2017).

5.1. Data and Variables

We focus on transcripts of all earnings conference calls held by S&P 500 companies between 2006 to 2018. To construct our evasiveness measure, we first organize the transcript of each call into document pairs, each pair consisting of one document containing a question raised by an analyst and another document containing its answer given by a manager. Our evasiveness measure is based on how matched the two documents are on the topic level. To measure incoherence, we resort to an ALBERT-based model to analyze the managers' responses.

Although our results are robust to various numbers of topics in measuring evasiveness (e.g., 50), we set it to 30 in this application for two reasons. First, this model better describes the text according to the model perplexity. Second, the topics learned seem meaningful based on human reading. When the number of topics is too large, keywords often overlap significantly across topics, leading to less interpretable results.

Our two dependent variables are the next-day abnormal stock return and next-quarter earnings surprise. We obtain stock returns data from the CRSP database and analysts' forecast data from the I/B/E/S database. We also obtain firm balance sheet information from the Compustat database. After removing earnings calls with missing data from any of the above-mentioned databases, our final sample consists of 6,802 earnings calls. In order to mitigate the impact of outliers, we applied a winsorization process to the most extreme 0.15‰ values on both ends, specifically focusing on the next quarter's earnings surprises and the subsequent day's abnormal returns. We plot the histograms of evasiveness and incoherence in Figure 1. We also preview our variables in Table 6 and report the descriptive statistics and the Pearson correlation coefficients between our proposed measures and firm characteristics in Table 7.



(a) Evasiveness

(b) Incoherence

Figure 1 **Visualizations of Distributions of Evasiveness and Incoherence**

Table 6 Variable Definitions

Variable	Definition
Next Earnings Surprise	The standardized earnings surprise based on analysts' forecast errors (AFE) for the next quarter after the earnings call. AFE is calculated as actual earnings per share minus the median of the forecast by all equity analysts using their most recent forecasts and adjusted by the volatility of seasonal changes in earnings in the past up to 18 quarters.
Forecast Revision	The sum of the scaled moving changes in the median one-quarter-ahead earnings forecasts within the past 3 months: $\sum_{j=0}^{2}(f_{i,t-j}-f_{i,t-j-1})/p_{i,t-j}$ where $f_{i,t}$ is the median analyst's quarterly forecast of firm i in month t as in Chan et al. (1996)
Forecast Dispersion	The standard deviation of the most recent earnings forecasts before the next earnings announcement scaled by the volatility of seasonal changes in earnings
log(Market Equity)	The logarithm of market equity at the end of the preceding year
log(Book/Market)	The logarithm of its book-to-market ratio at the end of the preceding year
log(Share Turnover)	The logarithm of annual shares traded adjusted by outstanding shares at the end of the preceding year
$FF\alpha$	The estimated intercept from the event study regression that spans the [–252,–31] time window
$FFCAR_{-2,-2}$	The abnormal return on day -2, i.e., 2 days prior to the next quarterly earnings announcement
$FFCAR_{-30,-3}$	The cumulative abnormal return during the trading window [-30, -3] prior to the next quarterly earnings announcement
Evasiveness	Machine-learning-based alignment between analysts' questions and executives' answers
Incoherence	AI-based answer coherence measure
Negativity	The percentage of negative sentiment measured by Loughran and McDonald (2011) in executives' answers
Log(Total Words)	The log of total words delivered by executives
$FFCAR_{t+1,t+1}$	The next-day abnormal return of the earnings call
Question Complexity	The logarithm of the number of topics with positive probabilities in questions
$FFCAR_{t,t}$	The same-day abnormal return of the earnings call
$FFCAR_{t-1,t-1}$	The previous-day abnormal return of the earnings call
$FFCAR_{t-2,t-2}$	The 2-day-ahead abnormal return of the earnings call
$FFCAR_{t-30,t-3}$	The previous month's abnormal return of the earnings call from 30 to 3 days ahead of the earnings call

Table 7 Descriptive Statistics

Panel A: Means and Standard Deviations of the Variables

	Mean	Std. Dev.		Mean	Std. Dev.
Next Earnings Surprise	0.12	0.412	Evasiveness	0.60	0.067
Forecast Dispersion	0.30	0.445	Incoherence	0.62	0.102
Forecast Revisions	0.0001	0.017	Negativity	10.55	7.409
log(Market Equity)	9.80	1.111	Log(Total Words)	8.23	0.455
log(Book/Market)	-0.23	0.805	$FFCAR_{t+1,t+1}$	0.0002	0.04
log(Share Turnover)	14.64	0.538	Question Complexity	2.050	0.218
$FF\alpha$	-0.0001	0.001	$FFCAR_{t,t}$	0.001	0.041
$FFCAR_{-2,-2}$	0.0005	0.015	$FFCAR_{t-1,t-1}$	0.0004	0.015
$FFCAR_{-30,-3}$	0.0021	0.081	$FFCAR_{t-2,t-2}$	0.0001	0.014
			$FFCAR_{t-30,t-3}$	-0.005	0.081

Panel B: Pearson Correlations with Variables of Interest

	Evasiveness	Incoherence		Evasiveness	Incoherence
$FFCAR_{t,t}$	0.029*	-0.026*	Negativity	-0.070*	0.026*
$FFCAR_{t-1,t-1}$	-0.006	0.010	Question Complexity	-0.018	-0.096*
$FFCAR_{t-2,t-2}$	0.010	-0.010	Forecast Dispersion	0.034*	-0.076*
$FFCAR_{t-30,t-3}$	0.017	-0.039*	Forecast Revisions	0.017	-0.020
$FF\alpha$	0.026*	-0.069*	$FFCAR_{-30,-3}$	0.014	-0.021
log(Market Equity)	-0.068*	-0.068*	$FFCAR_{-2,-2}$	0.022	-0.001
log(Book/Market)	0.016	0.146*	Recent Earnings Surprise	0.006	-0.051*
log(Share Turnover)	0.078*	-0.054*	Log(total words)	-0.047*	-0.055*

Note: * $p < 0.05$

5.2. Abnormal Return

We first investigate how the stock market responds to managers' evasiveness and incoherence during an earnings conference call. To do so, we predict the next-day abnormal stock return, using these measures along with many control variables. We measure abnormal returns based on the three-factor model (Fama and French 1993).

To control for the earnings shock, we include the earnings surprise of the immediate previous quarter. We also control for firm characteristics and risk factors according to the literature. Specifically, we include the abnormal return of the earnings call day ($FFCAR_{t,t}$), where the earnings call date is set as day t. The prior day ($FFCAR_{t-1,t-1}$) and the 2 days ahead ($FFCAR_{t-2,t-2}$) to control for short-term stock returns; we include the abnormal return of the previous month ($FFCAR_{t-30,t-3}$) to control for medium-term stock returns; and we include the momentum effect ($FF\alpha$) from the past year and the trading volume. In addition, we include the firm size and book-to-market ratio to control for the priced factors (Fama and French 1992). Because of the conversational nature of our data, we also account for question complexity, using the logarithm of the number of topics with positive probabilities in a question ($QuestionComplexity$). Moreover, we consider content information and sentiment negativity using the log of total words delivered by executives and its percentage of negative sentiment as measured by Loughran and McDonald (2011).

Table 8 reports the results of the four specifications. Incoherence consistently predicts lower stock returns at a 5% level of statistical significance. Comparing the results from columns 1 and 2, in terms of economic magnitude, the two proposed measures boost the adjusted R^2 by 21.35% when we predict the next-day abnormal return ($FFCAR_{t+1,t+1}$). These findings remain robust if we include year and/or firm fixed effects, as we report in columns 3 and 4. The abnormal stock return is about 0.196 (i.e., 0.049 × 4) standard deviations lower if the incoherence measure is two standard deviations above instead of below its mean. This corresponds to roughly 0.74% of the 1-day risk-adjusted return. Clearly, managers' failure to answer coherently during earnings calls is negatively perceived by investors. The estimated coefficients of evasiveness are negative, although they are not statistically significant. Regarding the control variables, we find that the just-announced earnings surprise and the total information proxied by $log(Total\ Words)$ strongly explain the abnormal return. In addition, managers' negative sentiment leads to lower risk-adjusted stock returns. Not surprisingly, many control variables are insignificant because in a highly efficient stock market, few variables should predict stock returns.

Table 8 Predicting Stock Risk-Adjusted Returns

DV: Risk-Adj. Returns ($FFCAR_{t+1,t+1}$)	(1)	(2)	(3)	(4)
$FFCAR_{t,t}$	-0.022	-0.022	-0.023*	-0.027*
	(0.013)	(0.013)	(0.013)	(0.014)
$FFCAR_{t-1,t-1}$	-0.002	-0.002	-0.003	-0.005
	(0.021)	(0.021)	(0.021)	(0.021)
$FFCAR_{t-2,t-2}$	0.001	0.001	0.001	0.003
	(0.008)	(0.009)	(0.009)	(0.010)
$FFCAR_{t-30,t-3}$	-0.018	-0.021	-0.024	-0.032*
	(0.013)	(0.012)	(0.013)	(0.015)
$FF\alpha$	-0.037	-0.040*	-0.041	-0.064**
	(0.022)	(0.022)	(0.024)	(0.023)
Earnings Surprise	0.084***	0.083***	0.084***	0.103***
	(0.026)	(0.026)	(0.027)	(0.030)
$log(Market\ Equity)$	0.006	0.001	0.004	-0.193*
	(0.015)	(0.015)	(0.014)	(0.094)
$log(Book\ Market)$	-0.009	-0.003	-0.002	-0.022
	(0.013)	(0.012)	(0.012)	(0.074)
$log(Share\ Turnover)$	0.011	0.005	0.013	0.026
	(0.013)	(0.013)	(0.014)	(0.026)
$log(Total\ Words)$	0.044***	0.041***	0.035**	0.054***
	(0.013)	(0.012)	(0.012)	(0.014)
Negativity	-0.054***	-0.052**	-0.052**	-0.083***
	(0.018)	(0.017)	(0.017)	(0.024)
Question Complexity	-0.009	-0.013	-0.014	-0.013
	(0.009)	(0.009)	(0.008)	(0.008)
Incoherence		**-0.048****	**-0.047****	**-0.049****
		(0.016)	(0.016)	(0.018)
Evasiveness		**-0.009**	**-0.011**	**-0.007**
		(0.012)	(0.013)	(0.015)
Year FE			YES	YES
Firm FE				YES
Observations	6,796	6,796	6,796	6,795
Adjusted R-squared	0.89%	1.08%	1.10%	1.93%

Note: t-statistics based on standard errors clustered by firm and time are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

5.3. Earnings Surprises

To understand whether trading based on evasiveness and incoherence detected during earnings conference calls is well-grounded in firm fundamentals or merely results from market overreaction, we investigate whether evasiveness and incoherence actually predict worse next-quarter earnings surprises. To do so, we follow the accounting literature to assess firm fundamentals using the standardized earnings surprise based on analysts' forecast errors (AFE), which is calculated as the median of the forecast errors of a firm's quarterly earnings by all equity analysts using their most recent forecasts. Following the literature, we adjust AFE by the volatility of seasonal changes in earnings, which are calculated using seasonal changes in earnings in the past up to 18 quarters.

We include as control variables the lagged dependent variable, firm size as measured by the logarithm of its market equity, and the logarithm of its book-to-market ratio, and the trading volume using the logarithm of annual shares traded adjusted by outstanding shares, all evaluated at the end of the previous year. To remove the predictive power from past returns, we include three control variables for a firm's recent returns, which are calculated from an earnings announcement event study using the benchmark returns based on the three-factor model (Fama and French 1993). Specifically, we include the cumulative abnormal return during the trading window [-30, -3] ($FFCAR_{-30,-3}$) and the abnormal return on day -2 ($FFCAR_{-2,-2}$), where the next quarterly earnings announcement date is set as day 0. These two variables capture the return information of the 29 trading days prior to the next earnings announcement, which should incorporate the most recent information about firm fundamentals. To control for the firm's return momentum (Jegadeesh and Titman 1993) over the year before the earnings call, we include the control variable $FF\alpha$, which is the estimated intercept from the event study regression measuring the in-sample cumulative abnormal return of the previous year. To remove the predictive power from analysts' forecast dispersion and forecast revision before the next earnings announcement, we control both variables. To construct forecast revision, we sum the scaled moving changes in the median one-quarter-ahead earnings forecasts within the past 3 months: $\sum_{j=0}^{2}(f_{i,t-j}-f_{i,t-j-1})/p_{i,t-j}$, where $f_{i,t}$ is the median analyst's quarterly forecast of firm i in month t. The monthly revision is scaled by the stock price, $p_{i,t-j}$. We calculate forecast dispersion as the standard deviation of the most recent earnings forecasts before the next earnings announcement, scaled by the volatility of seasonal changes in earnings.

We estimate a pooled OLS model with and without year-fixed effects. In accordance with the earnings surprise literature (Hobson et al. 2012, Kelley and Tetlock 2013, Chen et al. 2014), we do not include firm fixed effects, as the firm trend in earnings is already adjusted. Indeed, because the earnings surprise measures analysts' forecast errors, any persistent under- or over-estimation of the earnings for a firm over years cannot exist in equilibrium. Table 9 reports the estimation results. Our findings point toward a significant correlation between both incoherence and evasiveness and unexpectedly lower earnings for the upcoming quarter. Comparing the results from columns 1 and 2, in terms of economic magnitude, the two proposed measures boost the adjusted R^2 by 12.5% when we predict the next quarter's earnings surprise. In terms of scale, if the incoherence metric is two standard deviations above rather than below its mean, the conditional expectation of SAFE is reduced by four times (0.034×4=13.6%) the standard deviation. This translates to an unforeseen drop of $0.056 per unit of seasonal earnings change volatility. In a similar vein, the predicted surprises in earnings for the next quarter carry the same economic implications, as shown in the coefficients' alignment in column 3 for incoherence and evasiveness. Among the control variables, we find that lagged earnings surprise, return momentum ($FF\alpha$) and recent returns ($FFCAR_{-2,-2}$ and $FFCAR_{-30,-3}$) serve as robust indicators of future earnings surprises. We also find analyst forecast dispersion to be informative in predicting earnings surprises in both specifications.

Table 9 Predicting Earnings Surprises

	(1)	(2)	(3)
Earnings Surprise (lagged)	0.180***	0.179***	0.175***
	(0.040)	(0.040)	(0.039)
$FFCAR_{-2,-2}$	0.034***	0.035***	0.033**
	(0.011)	(0.011)	(0.011)
$FFCAR_{-30,-3}$	0.048***	0.048***	0.053***
	(0.013)	(0.013)	(0.012)
$FF\alpha$	0.069***	0.069***	0.081***
	(0.016)	(0.016)	(0.015)
Forecast Dispersion	0.170**	0.169**	0.170**
	(0.075)	(0.074)	(0.075)
Forecast Revision	0.011	0.012	0.016
	(0.009)	(0.009)	(0.010)
log(Market Equity)	0.043	0.039	0.037
	(0.026)	(0.024)	(0.024)
log(Book/Market)	-0.033*	-0.029	-0.028
	(0.017)	(0.017)	(0.018)
log(Share Turnover)	0.022	0.021	0.009
	(0.026)	(0.025)	(0.026)
log(Total Words)	-0.003	-0.005	0.005
	(0.015)	(0.015)	(0.015)
Negativity	-0.009	-0.009	-0.015
	(0.014)	(0.014)	(0.015)
Question Complexity	0.013	0.010	0.006
	(0.013)	(0.013)	(0.012)
Incoherence		**-0.031****	**-0.034****
		(0.014)	(0.014)
Evasiveness		**-0.035****	**-0.034****
		(0.012)	(0.013)
Year FE			YES
Observations	6,800	6,800	6,800
Adjusted R-squared	0.08	0.09	0.09

Note: t-statistics based on standard errors clustered by firm and time are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

To summarize, after controlling for a comprehensive set of factors that the literature suggests predict earnings, we find that evasiveness and incoherence still contain additional information on a firm's future earnings surprise, although only incoherence is currently well captured by investors. These findings raise an interesting question: why might analysts not fully adjust their forecasts in light of evasiveness and incoherence observed in an earnings call? There are at least two plausible explanations.[5] The first is a conflict-of-interest story: the pressure to maintain positive relationships with management may lead to distorted earnings forecasts (Malmendier and Shanthikumar 2014). Second, the overconfidence of an analyst may lead him/her to ignore implicit but valuable language cues such as evasiveness and incoherence. An exploratory analysis using an analyst's frequency of attending a firm's earnings call as a proxy for both overconfidence and the lack of independence

[5] We thank the AE and the reviewers for inspiring and suggesting these intriguing ideas.

yield supporting[6] evidence. While these potential mechanisms are consistent with our preliminary evidence, more empirical works using more fine-grained data are needed to further our understanding of these mechanisms.

5.4. Robustness Check

In this section of our exploration, we proceed to strengthen our analysis in the earnings call context. We delve into various concepts that could potentially correlate with evasiveness. Furthermore, our study extends to potential heterogeneous expectations associated with evasiveness and incoherence given the ranks of different speakers. Lastly, we investigate the reasons why evasiveness fails to predict next-day abnormal returns.

5.4.1. Evasiveness and Related Concepts We first check the correlations between our evasiveness measure and two related measures in the accounting literature: vagueness and linguistic complexity. We construct these two measures following Dzieliński et al. (2017) and Bushee et al. (2018), respectively. The results show that the linguistic complexity measure is significantly correlated with our evasiveness measure, as shown in Table 10. However, our evasiveness measure provides valuable incremental information beyond previous measures of vagueness and complexity, as the correlations between our evasiveness measure and complexity and vagueness measures are small.

Because the linguistic complexity measure is significantly correlated with our evasiveness measure, we further test the robustness of our findings by controlling for this variable. Consistent with our main findings, the results reported in Table 11 and Table 12 show that both evasiveness and incoherence predict worse next-quarter earnings and that the stock market perceives incoherence as a negative signal.

[6] Specifically, we consider an analyst as a frequent analyst of a firm if he/she attends the firm's earnings call at least once every three years. Both incoherence and evasiveness predict negative next-quarter earnings surprise only for those earnings call with more frequent analysts.

Table 10 Correlations between Our Evasiveness Measure and Related Concepts

	Correlation	p-value
Question Complexity	-0.018	0.14
Vagueness	0.01	0.41
Linguistic Complexity	-0.067**	0.00

Note: $Vagueness$ is the lexicon-based measure of vagueness as described in Dzieliński et al. (2017). *Linguistic Complexity* is the measure of language complexity widely adopted in accounting and finance research (Bushee et al. 2018). It measures the complexity of a written passage by analyzing the average sentence length and the percentage of words that are not commonly understood by the intended audience.

Table 11 Robustness Test: Predicting Stock Risk-Adjusted Returns Considering Linguistic Complexity

DV: Risk-Adj. Returns ($FFCAR_{t+1,t+1}$)	(1)	(2)	(3)
$FFCAR_{t,t}$	-0.022	-0.023*	-0.026*
	(0.013)	(0.013)	(0.014)
$FFCAR_{t-1,t-1}$	-0.002	-0.003	-0.005
	(0.021)	(0.021)	(0.021)
$FFCAR_{t-2,t-2}$	0.001	0.001	0.003
	(0.009)	(0.009)	(0.010)
$FFCAR_{t-30,t-3}$	-0.021	-0.024	-0.031*
	(0.012)	(0.013)	(0.015)
$FF\alpha$	-0.040*	-0.041	-0.064**
	(0.022)	(0.024)	(0.023)
Earnings Surprise	0.083***	0.084***	0.103***
	(0.026)	(0.027)	(0.030)
$log(Market\ Equity)$	0.002	0.004	-0.194*
	(0.015)	(0.015)	(0.094)
$log(Book\ Market)$	-0.003	-0.002	-0.022
	(0.012)	(0.012)	(0.074)
$log(Share\ Turnover)$	0.006	0.013	0.025
	(0.015)	(0.015)	(0.026)
$log(Total\ Words)$	0.041***	0.035**	0.053***
	(0.013)	(0.012)	(0.015)
Negativity	-0.052**	-0.052**	-0.083***
	(0.017)	(0.017)	(0.025)
Question Complexity	-0.014	-0.014*	-0.012
	(0.009)	(0.007)	(0.009)
Linguistic Complexity	-0.001	-0.000	0.003
	(0.004)	(0.005)	(0.007)
Incoherence	**-0.048****	**-0.047****	**-0.050****
	(0.016)	(0.016)	(0.019)
Evasiveness	**-0.010**	**-0.011**	**-0.007**
	(0.012)	(0.013)	(0.015)
Year FE		YES	YES
Firm FE			YES
Observations	6,796	6,796	6,795
Adjusted R-squared	1.07%	1.08%	1.92%

Note: t-statistics based on standard errors clustered by firm and time are in parentheses. *** $p< 0.01$, ** $p< 0.05$, * $p< 0.1$.

Table 12 Robustness Test: Predicting Earnings Surprises Considering Linguistic Complexity

	(1)	(2)
Earnings Surprise (Lagged)	0.179***	0.175***
	(0.040)	(0.039)
$FFCAR_{-2,-2}$	0.035***	0.033***
	(0.011)	(0.011)
$FFCAR_{-30,-3}$	0.048***	0.053***
	(0.013)	(0.012)
$FF\alpha$	0.069***	0.081***
	(0.016)	(0.016)
Forecast Dispersion	0.169**	0.170**
	(0.074)	(0.075)
Forecast Revision	0.012	0.016
	(0.009)	(0.010)
log(Market Equity)	0.037	0.036
	(0.024)	(0.023)
log(Book/Market)	-0.028	-0.027
	(0.018)	(0.018)
log(Share Turnover)	0.019	0.008
	(0.023)	(0.025)
log(Total Words)	-0.006	0.004
	(0.015)	(0.016)
Negativity	-0.009	-0.015
	(0.014)	(0.015)
Question Complexity	0.011	0.007
	(0.013)	(0.013)
Linguistic Complexity	0.003	0.002
	(0.005)	(0.005)
Incoherence	**-0.030***	**-0.034****
	(0.014)	(0.014)
Evasiveness	**-0.034****	**-0.033****
	(0.013)	(0.013)
Year FE		YES
Observations	6,800	6,800
Adjusted R-squared	0.09	0.09

Note: t-statistics based on standard errors clustered by firm and time are in parentheses. *** $p< 0.01$, ** $p< 0.05$, * $p< 0.1$.

5.4.2. Heterogeneous Expectations on Evasiveness and Incoherence for Executives of Various Ranks In practice, audiences may anticipate varying levels of evasiveness and incoherence, depending on the rank of the speaker in question. In the earnings call context, an CEO may use ambiguity in their discourse. This strategic use of evasiveness can be traced back to their broad, holistic viewpoint of every aspect of the company's functioning. On the other hand, senior executives such as chief financial officers (CFOs) may be able to deliver more clarity and precision in their communication on specific issues. Indeed, their proficiency in managing the company's financial intricacies positions them uniquely to provide insights that are both lucid and accurate.

In our sample, we observe that the median number of executives engaging with analysts during conference calls is three and that 27% of earnings calls involve no more than two executives fielding questions. Consequently, the possible challenge of differing expectations for details based on the speaker's rank is unlikely to cause major ripples in the earnings call landscape, due to the limited variation in executive ranks participating in these conference calls. However, to further strengthen the application value of our method, we control for whether the Q&A part of an earnings call is handled entirely by a CEO. The results, reported in Tables 13 and 14, remain consistent in predicting both the risk-adjusted stock returns and earning surprises.

5.4.3. Why Evasiveness Fails to Predict the Next-day Abnormal Return Unlike incoherence which predicts both the next-quarter earnings surprise and the next-day stock abnormal return, evasiveness seems to predict only the next-quarter earnings. To shed light on this puzzle, we offer a potential explanation based on investors' perceived accuracy of the evasiveness signal. Although the general construction of evasiveness through topic alignment is validated using both political interviews and responses from college students, its real-world effect on decision makers also depends on the their perceived accuracy of the signal. In our application context, the number of the question-answer pairs and the inherent complexity in identifying arbitrage opportunity might have led to the insignificant result of evasiveness in predicting the next-day abnormal return. Recall that in our application, we measure the evasiveness of a conference call by averaging the evasiveness of each

Table 13 Robustness Test: Predicting Stock Risk-Adjusted Returns Considering Executive Ranks

DV: Risk-Adj. Returns ($FFCAR_{t+1,t+1}$)	(1)	(2)	(3)
$FFCAR_{t,t}$	-0.023*	-0.023*	-0.027*
	(0.013)	(0.013)	(0.014)
$FFCAR_{t-1,t-1}$	-0.002	-0.003	-0.005
	(0.021)	(0.021)	(0.021)
$FFCAR_{t-2,t-2}$	0.001	0.002	0.003
	(0.009)	(0.009)	(0.011)
$FFCAR_{t-30,t-3}$	-0.021	-0.024	-0.032*
	(0.013)	(0.014)	(0.015)
$FF\alpha$	-0.040*	-0.041	-0.064**
	(0.022)	(0.024)	(0.023)
Earnings Surprise	0.083***	0.085***	0.103***
	(0.027)	(0.027)	(0.030)
$log(Market\ Equity)$	0.002	0.004	-0.191*
	(0.015)	(0.015)	(0.093)
$log(Book\ Market)$	-0.001	-0.001	-0.020
	(0.012)	(0.011)	(0.073)
$log(Share\ Turnover)$	0.006	0.014	0.026
	(0.013)	(0.014)	(0.026)
$log(Total\ Words)$	0.049***	0.042***	0.060***
	(0.013)	(0.012)	(0.013)
Negativity	-0.061***	-0.060***	-0.089***
	(0.019)	(0.019)	(0.027)
Question Complexity	-0.014	-0.015	-0.014
	(0.009)	(0.008)	(0.008)
Incoherence	**-0.047****	**-0.046****	**-0.048****
	(0.016)	(0.016)	(0.018)
Evasiveness	**-0.009**	**-0.010**	**-0.007**
	(0.012)	(0.013)	(0.015)
CEO Only	0.237**	0.216**	0.193
	(0.090)	(0.091)	(0.131)
Year FE		YES	YES
Firm FE			YES
Observations	6,796	6,796	6,795
Adjusted R-squared	1.13%	1.13%	1.94%

Note: t-statistics based on standard errors clustered by firm and time are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Q&A pair. If the number of Q&A pairs is relatively small, investors may not be able to form a strong enough perception of executives' evasiveness. In other words, they may not be confident about their perceived evasiveness due to the large variance caused by the relatively small sample size of Q&A pairs. As a result, they are less likely to punish the firm on the stock market. Alternatively, with a relatively small number of Q&A pairs, our evasiveness measure itself might be too noisy, resulting in a large variance estimation, hence a large p-value. To test, we compare the predictive power of the evasiveness measures across earnings calls with small or large number of Q&A pairs. The results reported in Table 15 are consistent with our explanation. More specifically, in this analysis, we created two binary variables, MoreQA is 1 if the number of Q&A pairs is more than 25 and 0 otherwise. LessQA is defined as 1-MoreQA. As we can see from Table 15, for conference calls with more pairs of Q&A, evasiveness does negatively predict stock return with statistical significance. That both

Table 14 Robustness Test: Predicting Earnings Surprises Considering Executive Ranks

	(1)	(2)
Earnings Surprise (Lagged)	0.179***	0.175***
	(0.040)	(0.039)
$FFCAR_{-2,-2}$	0.034***	0.032**
	(0.011)	(0.011)
$FFCAR_{-30,-3}$	0.048***	0.053***
	(0.013)	(0.012)
$FF\alpha$	0.069***	0.081***
	(0.016)	(0.015)
Forecast Dispersion	0.169**	0.170**
	(0.075)	(0.075)
Forecast Revision	0.012	0.016
	(0.009)	(0.010)
log(Market Equity)	0.039	0.037
	(0.025)	(0.024)
log(Book/Market)	-0.029	-0.028
	(0.017)	(0.018)
log(Share Turnover)	0.020	0.008
	(0.025)	(0.026)
log(Total Words)	-0.008	0.002
	(0.016)	(0.016)
Negativity	-0.004	-0.011
	(0.015)	(0.016)
Question Complexity	0.010	0.007
	(0.012)	(0.012)
Incoherence	**-0.031****	**-0.035****
	(0.014)	(0.013)
Evasiveness	**-0.035****	**-0.034****
	(0.012)	(0.013)
CEO Only	-0.109	-0.091
	(0.091)	(0.099)
Year FE		YES
Observations	6,800	6,800
Adjusted R-squared	0.09	0.09

Note: t-statistics based on standard errors clustered by firm and time are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

analysts and investors have difficulty detecting and/or incorporating evasiveness also implies the limitation of human investors and analysts, suggesting the advantages of the AI approach advocated in the current paper.

Table 15 Predictive Power of Evasiveness for Risk-Adjusted Returns

DV: Risk-Adj. Returns ($FFCAR_{t+1,t+1}$)	(1)	(2)
$FFCAR_{t,t}$	-0.001**	-0.001*
	(0.000)	(0.001)
$FFCAR_{t-1,t-1}$	-0.000	-0.000
	(0.001)	(0.001)
$FFCAR_{t-2,t-2}$	0.000	0.000
	(0.000)	(0.000)
$FFCAR_{t-30,t-3}$	-0.001	-0.001*
	(0.001)	(0.001)
$FF\alpha$	-0.002	-0.002**
	(0.001)	(0.001)
Earnings Surprises	0.003***	0.004***
	(0.001)	(0.001)
$log(Market\ Equity)$	0.000	-0.007*
	(0.000)	(0.004)
$log(Book\ Market)$	-0.000	-0.001
	(0.000)	(0.003)
$log(Share\ Turnover)$	0.001	0.001
	(0.001)	(0.001)
$log(Total\ Words)$	0.001***	0.002***
	(0.000)	(0.001)
Negativity	-0.002***	-0.003***
	(0.001)	(0.001)
Question Complexity	-0.000	-0.000
	(0.000)	(0.000)
LessQA*Incoherence	-0.002	-0.002
	(0.001)	(0.001)
LessQA*Evasiveness	0.001	0.001
	(0.001)	(0.001)
MoreQA*Incoherence	**-0.002*** **	**-0.002*** **
	(0.001)	**(0.001)**
MoreQA*Evasiveness	**-0.001*** **	**-0.001***
	(0.000)	**(0.001)**
Year FE		YES
Firm FE	YES	YES
Observations	6,796	6,795
Adjusted R-squared	1.14%	1.98%

Note: t-statistics based on standard errors clustered by firm and time are in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

6. Conclusion

In this paper, we propose two machine-learning–based measures to quantify evasiveness and incoherence in real-time strategic conversations. We validate these measures in different contexts and demonstrate their business value through a concrete application in which we analyze managers' responses to questions during earnings conference calls. We show that both measures provide additional information about a firm's earnings in the following quarter, and the incoherence measure also predicts a lower next-day abnormal stock return after the firm's earnings call.

This paper makes three contributions to the academic literature. First and foremost, the paper contributes methodologically by proposing and measuring two machine-learning-powered constructs to analyze real-time strategic conversations which serve as an indispensable component for the effective functioning of many important institutions and markets. Second, the paper contributes to the emerging field of fintech by demonstrating how financially valuable information can be extracted from conversations between managers and analysts during conference calls, a type of data mostly overlooked by fintech algorithms. Finally, this paper pioneers the combination of machine learning and asset pricing, a direction with a potentially high impact given the rapid advances in AI.

The paper also has important practical implications. Consider any institution or market where real-time strategic conversations routinely occur (see Table 1 for examples). The most obvious beneficiary of our toolkit is the less-informed party. To overcome their informational disadvantage, the less-informed party usually relies on professionals (e.g., analysts, journalists, interviewers) to engage in information-revealing, real-time strategic conversations with the more-informed party. Aided with our toolkit, these professionals can more quickly and easily detect evasiveness and incoherence from the responses given by the more-informed party. Accordingly, they can adjust their subsequent questions to better elicit otherwise hidden information, which ultimately allows the less-informed party to make more-informed decisions. For example, investors can learn more about a firm's performance to better adjust their investments; citizens can better understand a political candidate's policy stance to make more-informed voting decisions; and employers can hire job candidates who better fit a job.

For the more-informed party, with increasingly powerful AI technologies scrutinizing their real-time responses, it will be more and more difficult for them to manipulate information disclosure. In this cat-and-mouse game of information-seeking and information-hiding, we bet that the less-informed party aided by AI will eventually win, leaving only one viable option for the more-informed party: to be forthright. But regardless of the more-informed party's strategy, the less-informed party benefits from our toolkit because of the additional information that is either directly disclosed by the more-informed party or indirectly revealed through the detection of evasiveness and incoherence. Whether they take advantage of this by forming profitable trading strategies or by casting more-informed ballots, in the end, the market or the institution will be able to function more efficiently, thereby benefiting the society as a whole. Therefore, the main practical implication of this research is an improvement in market and institution efficiency.

Our paper is not without limitations. First, like any machine-learning algorithm, the quality of the output measures depends on the quality of the input data. For example, the lack of statistical significance of evasiveness in the stock-return analysis might have been caused by the insufficient number of Q&A pairs in conference calls. Second, practitioners should bear in mind that the relative value of evasiveness or incoherence is probably more informative than its absolute value. It would be valuable for future research to further quantify and adjust for the heterogeneity of these measures in different settings. Finally, we tested the value of our proposed measures only in one application, but there are clearly more settings that abound with real-time strategic conversations. Exploring applications in those settings is an important future research direction.

References

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. *Journal of Machine Learning Research* 3(Jan):993–1022.

Brown P, Levinson SC, Levinson SC (1987) *Politeness: Some universals in language usage*, volume 4 (Cambridge university press).

Bull P (2003) *The microanalysis of political communication: Claptrap and ambiguity*, volume 7 (Routledge).

Bushee BJ, Gow ID, Taylor DJ (2018) Linguistic complexity in firm disclosures: Obfuscation or information? *Journal of Accounting Research* 56(1):85–121.

Chan LK, Jegadeesh N, Lakonishok J (1996) Momentum strategies. *Journal of Finance* 51(5):1681–1713.

Chen H, De P, Hu YJ, Hwang BH (2014) Wisdom of crowds: The value of stock opinions transmitted through social media. *Review of Financial Studies* 27(5):1367–1403.

Chen JV, Nagar V, Schoenfeld J (2018) Manager-analyst conversations in earnings conference calls. *Review of Accounting Studies* 23:1315–1354.

Chen Y, Rui H, Whinston A, et al. (2021) Tweet to the top? social media personal branding and career outcomes. *MIS Quarterly* 45(2):499–534.

Clayman SE (2001) Answers and evasions. *Language in society* 30(3):403–442.

Crawford VP, Sobel J (1982) Strategic information transmission. *Econometrica* 50(6):1431–1451.

Crossley SA, McNamara DS (2016) Say more and be more coherent: How text elaboration and cohesion can increase writing quality. *Journal of Writing Research* 7(3):351–370.

Cuayáhuitl KS H, Lemon O (2015) Strategic dialogue management via deep reinforcement learning. *arXiv preprint arXiv:1511.08099* .

Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186 (Minneapolis, Minnesota: Association for Computational Linguistics).

Dzieliński M, Wagner AF, Zeckhauser RJ (2017) Straight talkers and vague talkers: The effects of managerial style in earnings conference calls. Technical report, National Bureau of Economic Research.

Ertugrul M, Lei J, Qiu J, Wan C (2017) Annual report readability, tone ambiguity, and the cost of borrowing. *Journal of Financial and Quantitative Analysis* 52(2):811–836.

Fama EF, French KR (1992) The cross-section of expected stock returns. *Journal of Finance* 47(2):427–465.

Fama EF, French KR (1993) Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics* 33(1):3–56.

Foltz PW, Kintsch W, Landauer TK (1998) The measurement of textual coherence with latent semantic analysis. *Discourse processes* 25(2-3):285–307.

Gabrielsen J, Jønch-Clausen H, Pontoppidan C (2020) Answering without answering: Shifting as an evasive rhetorical strategy. *Journalism* 21(9):1355–1370.

Gan Z, Pu Y, Henao R, Li C, He X, Carin L (2016) Unsupervised learning of sentence representations using convolutional neural networks. *arXiv preprint arXiv:1611.07897* .

Goffman E (1959) *The presentation of self in everyday life* (Anchor).

Grosz BJ, Joshi AK, Weinstein S (1995) Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics* 21(2):203–225.

Halliday MAK, Hasan R (2014) *Cohesion in english* (Routledge).

Hauch V, Blandón-Gitlin I, Masip J, Sporer SL (2015) Are computers effective lie detectors? a meta-analysis of linguistic cues to deception. *Personality and Social Psychology Review* 19(4):307–342.

Hobson JL, Mayew WJ, Venkatachalam M (2012) Analyzing speech to detect financial misreporting. *Journal of Accounting Research* 50(2):349–392.

Huang AH, Lehavy R, Zang AY, Zheng R (2017) Analyst information discovery and interpretation roles: A topic modeling approach. *Management Science* 64(6):2833–2855.

Hutchby I, Wooffitt R (2008) *Conversation analysis* (Polity).

Hwang BH, Kim HH (2017) It pays to write well. *Journal of Financial Economics* 124(2):373–394.

Javers E (2010) Cia moonlights in corporate world. *Politico.com* .

Jegadeesh N, Titman S (1993) Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance* 48(1):65–91.

Jernite Y, Bowman SR, Sontag D (2017) Discourse-based objectives for fast unsupervised sentence representation learning. *arXiv preprint arXiv:1705.00557* .

Keizer S, Guhe M, Cuayáhuitl H, Efstathiou I, Engelbrecht KP, Dobre M, Lascarides A, Lemon O (2017) Evaluating persuasion strategies and deep reinforcement learning methods for negotiation dialogue agents. *15th EACL 2017 Software Demonstrations*, 480–484 (Association for Computational Linguistics).

Kelley EK, Tetlock PC (2013) How wise are crowds? insights from retail orders and stock returns. *Journal of Finance* 68(3):1229–1265.

Khalmetski K, Rockenbach B, Werner P (2017) Evasive lying in strategic communication. *Journal of Public Economics* 156:59–72.

Kiros R, Zhu Y, Salakhutdinov RR, Zemel R, Urtasun R, Torralba A, Fidler S (2015) Skip-thought vectors. *Advances in neural information processing systems*, 3294–3302.

Lai A, Tetreault J (2018) Discourse coherence in the wild: A dataset, evaluation and methods. *arXiv preprint arXiv:1805.04993* .

Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R (2019) ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR* abs/1909.11942.

Larcker DF, Zakolyukina AA (2012) Detecting deceptive discussions in conference calls. *Journal of Accounting Research* 50(2):495–540.

Lawrence A (2013) Individual investors and financial disclosure. *Journal of Accounting and Economics* 56(1):130–147.

Lehavy R, Li F, Merkley K (2011) The effect of annual report readability on analyst following and the properties of their earnings forecasts. *The Accounting Review* 86(3):1087–1115.

Lewis M, Yarats D, Dauphin YN, Parikh D, Batra D (2017) Deal or no deal? end-to-end learning for negotiation dialogues. *arXiv preprint arXiv:1706.05125* .

Li Y, Packard G, Berger J (2020) Conversational dynamics: When does employee language matter? *working paper* .

Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* .

Loughran T, McDonald B (2011) When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *Journal of Finance* 66(1):35–65.

Loughran T, McDonald B (2013) Ipo first-day returns, offer price revisions, volatility, and form s-1 language. *Journal of Financial Economics* 109(2):307–326.

Loughran T, McDonald B (2014) Measuring readability in financial disclosures. *Journal of Finance* 69(4):1643–1671.

Malmendier U, Shanthikumar D (2014) Do security analysts speak in two tongues? *Review of Financial Studies* 27(5):1287–1322.

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

Meadows M, O'Brien FA (2020) The use of scenarios in developing strategy: An analysis of conversation and video data. *Technological Forecasting and Social Change* 158:120147.

Milgrom P, Roberts J (1986) Relying on the information of interested parties. *The RAND Journal of Economics* 17(1):18–32.

Miller BP (2010) The effects of reporting complexity on small and large investor trading. *The Accounting Review* 85(6):2107–2143.

Packard G, Moore SG, McFerran B (2018) (i'm) happy to help (you): The impact of personal pronoun use in customer–firm interactions. *Journal of Marketing Research* 55(4):541–555.

PCAOB (2008) Proposed auditing standards related to the auditor's assessment of and response to risk. *PCAOB release No. 2008-005.PCAOB Washington, DC* .

Plüss B, Piwek P (2016) Measuring non-cooperation in dialogue. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 1925–1936.

Shi Z, Lee GM, Whinston AB (2016) Toward a better measure of business proximity. *MIS Quarterly* 40(4):1035–1056.

Silverman D (2020) *Qualitative research* (sage).

Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV (2019) Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32.

Zhou Y, He H, Black AW, Tsvetkov Y (2019) A dynamic strategy coach for effective negotiation. *CoRR* abs/1909.13426.