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

John Bai, Nicole Boyson, Yi Cao, Miao Liu, and Chi Wan

Q Group

May 2024

Motivation

- Public companies hold quarterly earnings conference calls
  - Investors attend and ask questions
  - Important source of information
- Human investors have limited processing capacity and memory
  - Might miss some key information

Research question

- Can computers (i.e., Chat-GPT) help detect new information offered by management during Q&A sessions of quarterly calls?
  - "New information" => Information not discussed in management's prepared remarks or in prior Q&A
    - We train the computer to consider all information already in its corpus **AND up-to-the minute information discussed earlier in the conference call**

Typical earnings conference call

- Presentation and discussion of financial results
- Overview of upcoming goals and milestones
- Discuss how plans may impact the future financial performance
- **Open floor for Q&A**

Typical earnings conference call -- example

Corporate Participants

* Jaideep K. Paul
  Airtel Africa Plc - CFO & Director
* Olusegun Ogunsanya
  Airtel Africa Plc - MD, CEO & Director

Conference Call Participants

* Jonathan D. Kennedy-Good
  JPMorgan Chase & Co, Research Division - Analyst
* Madhvendra Singh
  HSBC, Research Division - Analyst of Telecoms
* Maurice Graham Patrick
  Barclays Bank PLC, Research Division - MD
* Rohit Modi
  Citigroup Inc., Research Division - Senior Associate
* Tajudeen Ibrahim
  Chapel Hill Denham Securities Limited, Research Division - Head of Research

Olusegun Ogunsanya M

Chief executive officer
Date appointed to Board: October 2021
Independent: no
Nationality: Nigerian
Committee: Market Disclosure

Jaideep Paul F

Chief finance officer
Date appointed to Board: June 2021
Independent: no
Nationality: Indian
Committee: Finance (Chair)

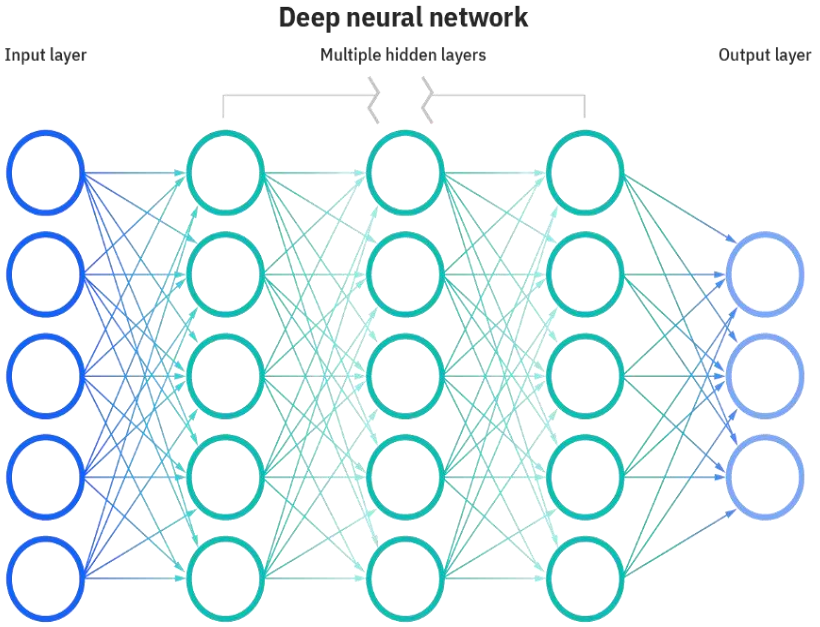
What is AI (in our setting?) (1)*

- Artificial Intelligence: enables computers to simulate human intelligence and problem-solving capabilities
- Generative AI: uses machine learning (ML) to take raw data and "learns" to generate statistically probable outputs when prompted
  - Encodes a simplified representation of training data and draws from it to create a new work similar, but not identical, to original data.
- Uses neural networks— a machine learning program, or model, that uses processes that mimic the way biological neurons work together to identify phenomena, weigh options and arrive at conclusions.

*https://www.ibm.com/topics/artificial-intelligence

What is AI (in our setting?) (2)*

- Neural networks: have an input layer, a number of hidden layers, and an output layout.
  - Classic ML requires human intervention/supervision
    - Uses human-labeled data to help predict outcomes
  - Deep ML can be unsupervised
    - Can ingest unstructured data in its raw form
    - Deep ML has more hidden layers
  - Reinforcement learning:
    - Computer learns by interacting with surroundings and getting feedback (rewards or penalties) for actions.



*https://www.ibm.com/topics/artificial-intelligence

What is ChatGPT (3.5)?*

- Chat Generative Pre-Training Transformer 3.5: a natural language processing (NLP) model trained to produce text
  - Statistically probable output = next word
- Uses Reinforcement Learning with Human Feedback (RLHF)
  – uses human demonstrations and preference comparisons to guide the model toward desired behavior
  – Trained on a massive corpus of text data, around 570GB of datasets, including web pages, books, and other sources
  – ChatGPT uses a transformer neural network
    - Understands relations between words, sentences, and concepts in text
    - Can infer meaning, recognize patterns, and generate relevant and coherent responses based on the context provided in the conversation
  – Updated through January 2022; not connected to Internet!!!

*https://decrypt.co/resources/what-is-chatgpt

Context Preservation

Related literature

- Li, Mai, Shen, Yang, and Zhang (2023) – corporate culture using ChatGPT
- Cao, Jiang, Wang, and Yang (2023) – can AI analyst beat human analysts
- Lopez-Lira and Tang (2023) – predict stock prices using ChatGPT
- Kim, Muhn, and Nikolaev (2023) – reduce the size of documents
- Eisfeldt, Schubert, Taska, and Zhang (2023) – labor implications
- Jha, Qian, Weber, and Yang (2023) – investment disclosure

Key measure

- Q&A section of earnings conference call
- **ChatGPT knows all general knowledge and conference-call specific knowledge up to question n**
- Measure semantic similarity between human's actual answer and ChatGPT answer
- **(1 – Similarity is main independent measure)**
  - **HAID:** Human-AI Discrepancy

Using Chat GPT 3.5-Turbo

- Chat GPT 3.5-Turbo allows us to pre-educate the model with conference call transcript data up to the question of interest
- Sets the context of the information as a "conversation" so that we receive back an "answer."
- Three roles in this model: "system," "assistant," and "user."
- We use a combination of "system" and "assistant" to preserve the context of the "conversation"

System (authors) role:

Two instructions to Chat GPT 3.5 turbo

1. From the perspective of a top executive, please answer the following question raised by a financial analyst during an earnings conference call
2. knowledge cutoff date (important for avoiding forward-looking bias)

Assistant role (executive) and user role (analyst)

INPUT:

1. (Assistant) executive presentation, summarized
2. (User) full set of preceding (n-1) analyst questions and (Assistant) corresponding executive answers
   – Analyst questions original text; executive answers summarized
3. (User) Analyst question n

OUTPUT:

(“Fake” Analyst/User): Answer to question n from Chat GPT

Snapshot

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

System input: our instructions

Assistant input: Executive presentation

User input: Analyst question

Assistant input: Executive answer

Fake user/analyst output: GPT answer

Summarizing presentation and executive answers

- ChatGPT 3.5-turbo has limit of 4,096 tokens
  - roughly 4,096*0.75=*3,072 words*
- Out of necessity, we pre-process presentation and executive answers using BART

  BART: a denoising autoencoder for pretraining sequence-to-sequence model developed by Facebook AI in 2019

Case aimilar answers

Appendix B: Two Cases of Chatbot vs. Executives Answers

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

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

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

Case dissimilar answers

- **Manager answer**: The sequential, July was down 2.7%, and that compares to the 10-year average being a negative 2.1%. In August, it was plus 0.2% versus the long-term average of being 0.6%. And then September, that was what we talked about. It was up 7.3%, and the long-term average is a positive 2.8%. We had that material increase in weight per shipment. And when you look at shipments, and I think I mentioned this before, shipments were pretty much in line with what normal seasonality was for the most part for the quarter. It was just that big step-up in weight per shipment that drove that weight change.
- **Chatbot answer**: We're very focused on that. Our IT group is focused on that. And we're making all every step we can to make that happen. Very, very focused on that right now.

Data & sample

- Sample period 2004 to 2020
- 190,538 conference call transcripts
- Compustat quarterly financial data; I/B/E/S analyst forecast and management guidance
- After dropping observations with missing firm-level controls, final sample: 104,932 earnings conference calls.

Key measure: HAID (Human-AI difference)

HAID(Bert) HAID based on semantic similarity using the Bidirectional Encoder Representations from Transformers (BERT) developed by Google.

HAID(Cos) HAID based on cosine similarity of word distributions.

HAID(Word2Vec) HAID based on Word2Vec, a pre-trained embedding provided

by Google. The pre-trained embeddings are available at https://code.google.com/archive/p/word2vec/

Bert:

- captures contextual meaning.
- Handles complex sentence structures and long-range dependencies

Semantic similarity

Semantic similarity:
 * distance between sentences or phrases based on the likeness of their meaning or semantic content as opposed to lexicographical similarity

Example of likeness of meaning without lexicographical similarity

Sentence 1: The cat ate the mouse
Sentence 2: The mouse was eaten by the kitten

These two sentences are semantically quite similar despite the rearrangement of subject and verb and use of kitten versus cat.

Summary statistics: key variables

Variable	Obs.	Mean	SD	Median
HAID (Bert)	104,932	0.311	0.045	0.308
HAID (Cos)	104,932	0.472	0.060	0.473
HAID (Word2Vec)	104,932	0.118	0.024	0.115
Number of questions in Q&A session (log)	104,932	2.524	0.507	2.639
Number of sentences in transcript (log)	104,932	2.360	0.426	2.342
Number of analysts (log)	104,932	1.677	0.987	1.792

Correlations: key variables

	HAID (Bert)	HAID (Cos)
HAID (Cos)	0.43	
HAID (Word2Vec)	0.47	0.74

Testable hypothesis

- H1a: Higher HAID → higher abnormal trading volume and absolute cumulative return

Rationale: higher HAID presumably contains potentially useful incremental information

- H1b: Higher HAID → higher stock liquidity
  - (bid-ask spread and Amihud)

Results for hypothesis 1a

Panel A: Abnormal Volume

VARIABLES	(1) AVol	(2) AVol	(3) AVol
HAID(Bert)	**1.241*****		
	(6.995)		
HAID(Cos)		**1.068*****	
		(4.981)	
HAID(Word2vec)			**3.281*****
			(7.167)

Panel B: Absolute Value of Cumulative Abnormal Return

VARIABLES	(1) Abs. CAR	(2) Abs. CAR	(3) Abs. CAR
HAID(Bert)	**0.223*****		
	(6.186)		
HAID(Cos)		**0.334*****	
		(10.434)	
HAID(Word2vec)			**0.622*****
			(10.236)

Results for hypothesis 1b

Panel A: Bid-Ask Spread

VARIABLES	(1) Bid-Ask	(2) Bid-Ask	(3) Bid-Ask
HAID(Bert)	**-0.335***** **(-8.443)**		
HAID(Cos)		**-0.341***** **(-6.242)**	
HAID(Word2vec)			**-0.779***** **(-7.744)**

Panel B: Amihud illiquidity

VARIABLES	(1) Amihud Ratio	(2) Amihud Ratio	(3) Amihud Ratio
HAID(Bert)	**-0.473***** **(-6.925)**		
HAID(Cos)		**-0.313***** **(-3.687)**	
HAID(Word2vec)			**-0.691***** **(-3.457)**

Testable hypothesis

- H2: Higher HAID → lower analyst error and analyst dispersion
- H3: Higher HAID → higher probability of managerial guidance

Results for hypothesis 2

Panel A: Forecast Error

VARIABLES	(1) Forecast Error	(2) Forecast Error	(3) Forecast Error
HAID(Bert)	**-0.111***** **(-6.233)**		
HAID(Cos)		**-0.083***** **(-5.518)**	
HAID(Word2vec)			**-0.224***** **(-5.112)**

Panel B: Forecast Dispersion

VARIABLES	(1) Forecast Dispersion	(2) Forecast Dispersion	(3) Forecast Dispersion
HAID(Bert)	**-0.047***** **(-3.112)**		
HAID(Cos)		**-0.081***** **(-6.087)**	
HAID(Word2vec)			**-0.173***** **(-4.901)**

Results for hypothesis 3

VARIABLES	(1) Guidance	(2) Guidance	(3) Guidance
HAID(Bert)	**0.152***** **(9.913)**		
HAID(Cos)		**0.178***** **(7.701)**	
HAID(Word2vec)			**0.356***** **(8.830)**

Cross-sectional prediction

- Impact of HAID should be more pronounced for **complex** firms, or firms with more **information asymmetry** between management and investors

Cross-sectional test results: AVol

VARIABLES	(1) AVol	(2) AVol	(3) AVol	(4) AVol
HAID × R&D	**0.731***			
	(1.988)			
HAID × Segment		**0.068**		
		(0.420)		
HAID × Insider Trade			**0.063*****	
			(2.876)	
HAID × Participants				**0.057**
				(0.335)
Other Control Variables	Yes	Yes	Yes	Yes
Firm Year-Qtr FE	Yes	Yes	Yes	Yes
Ind. Year-Qtr Cluster	Yes	Yes	Yes	Yes
Observations	104,878	93,037	101,971	104,878
Adj. R^2	0.301	0.301	0.301	0.301

Cross-sectional test results: Abs. CAR

VARIABLES	(5) Abs. CAR	(6) Abs. CAR	(7) Abs. CAR	(8) Abs. CAR
HAID × R&D	**0.001**			
	(0.007)			
HAID × Segment		**0.090*****		
		(2.792)		
HAID × Insider Trade			**0.002**	
			(0.507)	
HAID × Participants				**0.057***
				(1.960)
Other Control Variables	Yes	Yes	Yes	Yes
Firm Year-Qtr FE	Yes	Yes	Yes	Yes
Ind. Year-Qtr Cluster	Yes	Yes	Yes	Yes
Observations	91,248	80,024	88,641	91,248
Adj. R^2	0.087	0.089	0.087	0.087

Cross-sectional test results: Guidance

VARIABLES	(9) Guidance	(10) Guidance	(11) Guidance	(12) Guidance
HAID × R&D	**0.054***			
	(1.919)			
HAID × Segment		**0.041****		
		(2.316)		
HAID × Insider Trad			**0.006***	
			(1.824)	
HAID × Participants				**0.053*****
				(2.819)
Other Control Variables	Yes	Yes	Yes	Yes
Firm Year-Qtr FE	Yes	Yes	Yes	Yes
Ind. Year-Qtr Cluster	Yes	Yes	Yes	Yes
Observations	104,932	93,086	102,024	104,932
Adj. R^2	0.383	0.398	0.383	0.383

Results using Alternative LLM

Panel A: Google Bard

VARIABLES	(1) AVol	(2) Abs. CAR	(3) Forecast Error	(4) Forecast Dispersion	(5) Guidance
HAID(Bert)	**0.758***** **-5.381**	**0.192***** **-4.225**	**-0.032***** **(-3.448)**	**-0.023**** **(-2.008)**	**0.077***** **-3.311**

Panel B: Open sourced StableLM-Alpha

VARIABLES	(1) AVol	(2) Abs. CAR	(3) Forecast Error	(4) Forecast Dispersion	(5) Guidance
HAID(Bert)	**1.444***** **-6.503**	**0.219***** **-5.205**	**-0.144***** **(-7.797)**	**-0.021*** **(-1.987)**	**0.149***** **-7.009**

Determinants of HAID (selected results)

VARIABLES	(1) HAID(Bert)	(2) HAID(Cos)	(3) HAID(Word2vec)
Lag HAID(Bert)	-0.007		
	(-1.365)		
Lag HAID(Cos)		0.000	
		(0.070)	
Lag HAID(Word2vec)			-0.001
			(-0.323)
Spi	0.047***	0.063***	0.017*
	(2.753)	(4.005)	(1.882)
Numest	0.001***	0.001***	0.000**
	(3.235)	(3.237)	(2.049)
ROA	-0.029***	-0.026***	-0.014***
	(-3.674)	(-3.996)	(-4.055)
R&D	-0.011*	-0.012**	-0.003**
	(-1.912)	(-2.549)	(-2.137)

Controlling for Bushee's measures of complexity

VARIABLES	(1) AVol	(2) Abs. CAR	(3) Forecast Error	(4) Forecast Dispersion	(5) Guidance
HAID(Bert)	**1.270*****	**0.211*****	**-0.116*****	**-0.047*****	**0.159*****
	(6.575)	**(4.893)**	**(-6.333)**	**(-3.065)**	**(8.116)**
Info	0.038***	-0.004	-0.001	0.003**	0.001
	(3.310)	(-1.490)	(-1.031)	(2.096)	(0.634)
Obfus	0.016***	-0.001	-0.001**	0.001***	0.000
	(2.713)	(-1.028)	(-2.575)	(3.263)	(0.267)

Conclusion

- Provide novel large-scale evidence how LLMs can potentially help investors
- Propose a new measure to help uncover "hidden" information contained in manager responses
- Will be interesting to see how managers' responses evolve over time

To-do List

- Understand better determinants of HAID – a more systematic decomposition

- Signed vs. unsigned

Thank you!

Correlations

Panel B: Pairwise Correlation

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
(1) HAID(Bert)	1.00												
(2) HAID(Cos)	0.43												
(3) HAID (Word2Vec)	0.47	0.74											
(4) Num_Q	0.09	0.12	-0.18										
(5) Num_Sent	-0.06	-0.59	-0.42	0.04									
(6) Spi	0.02	0.04	0.02	0.01	-0.01								
(7) Numest	-0.02	-0.07	-0.18	0.37	0.11	-0.01							
(8) Unexp_Eaern	-0.01	0.00	0.00	0.00	0.00	0.31	-0.01						
(9) Ret_SD	-0.02	-0.02	0.02	-0.13	-0.05	-0.11	-0.13	0.04					
(10) Size	-0.05	-0.20	-0.25	0.33	0.27	0.07	0.43	-0.02	-0.44				
(11) MtoB	-0.05	-0.07	-0.08	0.06	0.07	0.01	0.11	-0.01	-0.02	0.16			
(12) Lev	0.03	0.00	0.02	-0.05	0.00	-0.04	-0.10	0.04	0.21	-0.12	-0.16		
(13) ROA	0.02	0.07	0.00	0.14	-0.01	0.35	0.08	0.28	-0.36	0.30	0.00	-0.07	
(14) R&D	-0.06	-0.07	-0.04	-0.08	0.00	-0.07	0.00	0.00	0.25	-0.19	0.02	-0.18	-0.46

Summary Statistics

Panel A: Summary Statistics

Variables	Obs.	Mean	SD	Median
Bid-Ask Spread	104,718	-6.884	1.253	-7.023
Amihud Ratio	104,932	-6.612	2.543	-6.785
Abs. CAR	91,248	0.042	0.054	0.024
Avol	104,878	1.815	2.764	0.974
Guidance	104,932	0.144	0.302	0.000
Forecast Error	82,673	0.021	0.168	0.020
Dispersion	82,673	0.136	0.202	0.070
HAID(Bert)	104,932	0.311	0.045	0.308
HAID(Cos)	104,932	0.472	0.060	0.473
HAID(Word2Vec)	104,932	0.118	0.024	0.115
Num_Q	104,932	2.524	0.507	2.639
Num_Sent	104,932	2.360	0.426	2.342
Spi	104,932	-0.003	0.010	0.000
Numest	104,932	1.677	0.987	1.792
Unexp_Earn	104,932	0.001	0.051	0.000
Ret_SD	104,932	0.115	0.067	0.098
Size	104,932	7.371	1.821	7.310
MtoB	104,932	3.205	5.587	2.104
Lev	104,932	0.584	1.161	0.224
ROA	104,932	0.002	0.041	0.008
R&D	104,932	0.025	0.057	0.000