Can ChatGPT Decipher Fedspeak?

Anne Lundgaard Hansen and Sophia Kazinnik [†]

April 10, 2023

Abstract Yes! This paper investigates the ability of Generative Pre-training Transformer (GPT) models to decipher Fedspeak, a term used to describe the technical language used by the Federal Reserve to communicate on monetary policy decisions. We evaluate the ability of GPT models to classify the policy stance of Federal Open Market Committee (FOMC) announcements relative to a human classified benchmark. The performance of GPT models surpasses that of other popular classification methods.

Keywords: Natural Language Processing (NLP), Generative Pre-training Transformer (GPT), Federal Reserve Communication, Applications, Artificial Intelligence (AI).
JEL Code: E52, E58, C88.

1 Introduction

Generative Pre-trained Transformer (GPT) models, and ChatGPT in particular, have received a tremendous amount of public attention in recent months. Since its release in November 2022, ChatGPT (Radford et al., 2018), an artificial intelligence chatbot, has become a prominent topic across digital platforms and academic fields alike.[1]

Geerling et al. (2023) show that ChatGPT performs well in the Test of Understanding in College Economics (TUCE), a standardized test of economics knowledge, answering 86.7% of the macroeconomics questions correctly. One can therefore think of ChatGPT as a virtual research assistant, potentially qualified for tasks such as classifying central bank communication texts. In principle, GPT models have enough domain knowledge to label economic texts correctly. However, they may not have the same level of nuance and context-awareness as a human research assistant. The technology can, therefore, be either hugely time and resource saving or it can result in misleading or wrong conclusions.

[†]The authors are with the Quantitative Supervision and Research group at the Federal Reserve Bank of Richmond. Address: 530 E Trade St, Charlotte NC 28202. Corresponding author: Sophia.Kazinnik@rich.frb.org.

We thank Steve Baker, Brian Ferrell, and Robin Lumsdaine for valuable feedback. We thank Bryson Alexander, Nadia Audzeichuk, and Ethan Butler for excellent research assistance.

[1]See, e.g., Choi et al. (2023) (law), Frieder et al. (2023) (mathematics), and Biswas (2023) (public health). A review of the literature on ChatGPT and GPT models in the contexts of economics and finance follows below.

We set out to test this hypothesis in the context of Fedspeak, the technical language used by the Federal Reserve (Fed) to communicate monetary policy decisions. Specifically, we consider the exercise of classifying announcements released by the Federal Reserve Open Market Committee (FOMC) based on their policy stance.

We contribute to a large literature that studies the content and sentiment of central bank communication and its impact on the financial markets, (e.g., Ehrmann and Fratzscher (2007); Hansen et al. (2019); Hayo and Neuenkirch (2015); Curti and Kazinnik (2021)) and the general public (Ehrmann and Wabitsch, 2022). Whereas earlier contributions (Chen, 2016; Hansen and McMahon, 2016; Jegadeesh and Wu, 2017; Benchimol et al., 2020) quantify central bank texts based on topic modeling and sentiment analysis using pre-defined dictionaries, (e.g., Loughran and McDonald (2011)), more recent papers (e.g., Doh et al. (2022), Bertsch et al. (2022), Gorodnichenko et al. (2023)) use pre-trained large language models (LLMs), such as the Bidirectional Encoder Representations from Transformers (BERT) developed by Google AI Langauge (Devlin et al. (2018)). In this paper, we compare the performance of GPT to both dictionary-based and BERT methods. We provide a first attempt at evaluating GPT models for the purpose of quantifying Fed communication. Closest to our paper is Leippold (2023a), who uses GPT models to demonstrate the vulnerabilities of the use of dictionaries in NLP tasks, showing that context-aware approaches like BERT serve as a better alternative. Others have explored the impact of GPT models and ChatGPT on the fields of economics and finance. For example, Leippold (2023b) interviews GPT models on the issues of climate change, showcasing the model strengths and deficiencies. Dowling and Lucey (2023) and Korinek (2023) discuss how ChatGPT and LLMs in general can be utilized by financial researchers to increase productivity by automating micro-tasks. Finally, Zaremba and Demir (2023) examine the current state of the GPT technology in finance and argue that it has the potential to improve NLP-based financial applications.[2]

We proceed as follows. Section 2 describes the data and the manual labeling task. Section 3 provides an overview of GPT models and other NLP algorithms used for comparison. The results are presented in Section 4. Section 5 concludes.

2 Data

The FOMC meets eight times a year to discuss the economic outlook and set the direction for monetary policy. These meetings are followed by public statements that summarize the committee's view of the economy and monetary policy decisions.[3] Our analysis focuses on FOMC statements published between 2010 and 2020. Consistent with the literature, we break down each statement into individual sentences.

[2]Interestingly, the paper by Zaremba and Demir (2023) is written entirely with the use of ChatGPT, based on the prompts the two (human) co-authors provide.

[3]FOMC statements are available from the website of the Federal Reserve Board of Governors: `https://www.federalreserve.gov/monetarypolicy/fomccalendars.htm`.

Since manual classification is time-consuming and costly, we randomly draw 500 sentences using uniform selection. This data is then manually annotated with respect to five categories: "dovish", "mostly dovish", "neutral", "mostly hawkish", and "hawkish". The definitions of these are provided in Table 1, while specific examples are provided in the supplementary material. We assign numerical values to the categories on a scale of -1 to 1, where 0 represents neutral, to compute performance metrics, which are also defined in the table. We employ five categories instead of three ("dovish", "neutral", "hawkish) to examine GPT's ability to discern subtle differences between closely related labels, a common challenge in machine learning.

Table 1
Category Definitions

Category label	Value	Definition
Dovish	-1	Strongly expresses a belief that the economy may be growing too slowly and may need stimulus through monetary policy.
Mostly dovish	-0.5	Overall message expresses a belief that the economy may be growing too slowly and may need stimulus through monetary policy.
Neutral	0	Expresses neither a hawkish nor dovish view and is mostly objective.
Mostly hawkish	0.5	Overall message expresses a belief that the economy is growing too quickly and may need to be slowed down through monetary policy.
Hawkish	1	Strongly expresses a belief that the economy is growing too quickly and may need to be slowed down through monetary policy.

To mitigate the risks of human bias and errors, each sentence is processed independently by three human reviewers, and the final label is computed as the average given the assigned numerical value for each category.[4] When classifying a sentences, the context is considered, but only within the confines of each sentence.

We provide summary statistics of the manual classifications and the disagreement among reviewers in Table 2. We note that the data is imbalanced in the sense that there are more sentences with dovish sentiment than with hawkish sentiment, which is consistent with the sample period in question. The human reviewers agree most on the classification of "mostly dovish", "neutral", and "mostly hawkish" sentences. For all sentences classified as either "dovish" or "hawkish", at least one reviewer rated the sentence differently by more than one category, e.g., as "neutral".

[4] Overall, the work was distributed between four reviewers, all with backgrounds within finance and economics.

Table 2
Summary Statistics of Manually Classified Data

	Total	Dovish	Mostly dovish	Neutral	Mostly hawkish	Hawkish
Count	500	104	144	191	47	14
Average disagreement	0.47	0.67	0.52	0.31	0.51	0.67
Count (>1 step disagreement)	264	104	60	67	19	14
Count (>2 steps disagreement)	49	0	21	22	6	0

Notes: Average disagreement is calculated as the average difference between the classifications assigned by the 3 reviewers using the numerical value of each classification as given in Table 1.

3 NLP Methods

3.1 GPT Models and ChatGPT

GPT (Generative Pretrained Transformer) is a family of transformer-based models that are trained on large amounts of text data. GPT models use self-attention architecture to consider sentence and paragraph context, allowing them to excel in various natural language processing tasks and capture nuanced language understanding (Vaswani et al., 2017; Zhang et al., 2022).

Our analysis focuses on GPT-3, which is the third generation of OpenAI's GPT language model.[5] GPT-3 is a family of models (ada, babbage, curie and davinci) that have different capabilities and speed. We use the davinci model in our analysis given its superior capabilities. ChatGPT, on the other hand, is an artifical intelligence chatbot, currently supported by both GPT-3 and GPT-4 models.[6]

Models like GPT-3 generally do not require explicit examples for additional training to perform well. This concept, referred to as zero-shot learning, is a type of machine learning where the model is trained to recognize and classify new objects or concepts that it has not been explicitly trained on before. The reason that zero-shot GPT-3 can deliver satisfactory performance is that the vast amount of information underlying the model enables it to perform unsupervised learning by observing patterns and structures within the text (Akyürek et al., 2022). The model is able to apply this knowledge to a wide range of text-related tasks without the need for explicit instruction or labels. This capability provides a potential for considerable impact in the field of economics and finance with regards to any text-related task. In addition to zero-shot GPT-3, we evaluate

[5] With each iteration of the GPT models, the models have increased in size and complexity. The first iteration, GPT-1, had 117 million parameters, while the second iteration, GPT-2, had 1.5 billion parameters. GPT-3 is an order of magnitude larger than GPT-2, and includes a number of new features such as multi-lingual support, improved reasoning capabilities, and the ability to generate coherent and natural-sounding text.

[6] GPT-4 was released on March 14 2023 to improve upon the previous version of the model with a better ability to accurately follow user intentions.

GPT-3 when fine-tuned on our data.[7]

3.2 BERT

Similar to GPT models, BERT is a pre-trained language model based on the transformer architecture using a masked language modeling approach to pre-train a deep neural network on large amounts of text data. It has achieved state-of-the-art results on many benchmark data sets, making it one of the most widely used models in NLP for applications within finance (e.g., Bertsch et al. (2022) and Huang and Hui Wang (2022)). BERT-based models differ from GPT-based models in terms of their architecture and the way they are pre-trained.[8]

3.3 Dictionary-Based Methods

Dictionary-based methods use pre-defined lexicons to quantify text data. These lexicons contain labeled words or phrases, and are popular for their simplicity and transparency. However, their effectiveness is limited by coverage, and they struggle with nuances and context. We use the following three dictionaries in our analysis:

- Loughran and McDonald (2011) (LM) dictionary: LM was created to classify financial language into positive, negative, and uncertain categories. It is widely used in financial research to classify the tone of financial news and other finance-related communications (e.g., Picault and Renault (2017) and Shapiro and Wilson (2022)).

- Henry (2008) financial dictionary: This dictionary has been widely employed in financial sentiment analysis. However, its shortcomings include a restricted word count and insufficient scope.

- Mohammad and Turney (2015) NRC Word-Emotion Association Lexicon (NRC): NRC is a linguistic resource created by the National Research Council of Canada to capture the sentiment of everyday language.

4 Results

We begin by evaluating how well GPT-3 classifies the FOMC announcements using zero-shot learning. We then consider the performance of GPT-3 when fine-tuned with 400 labeled sentences.

4.1 Zero-Shot Learning

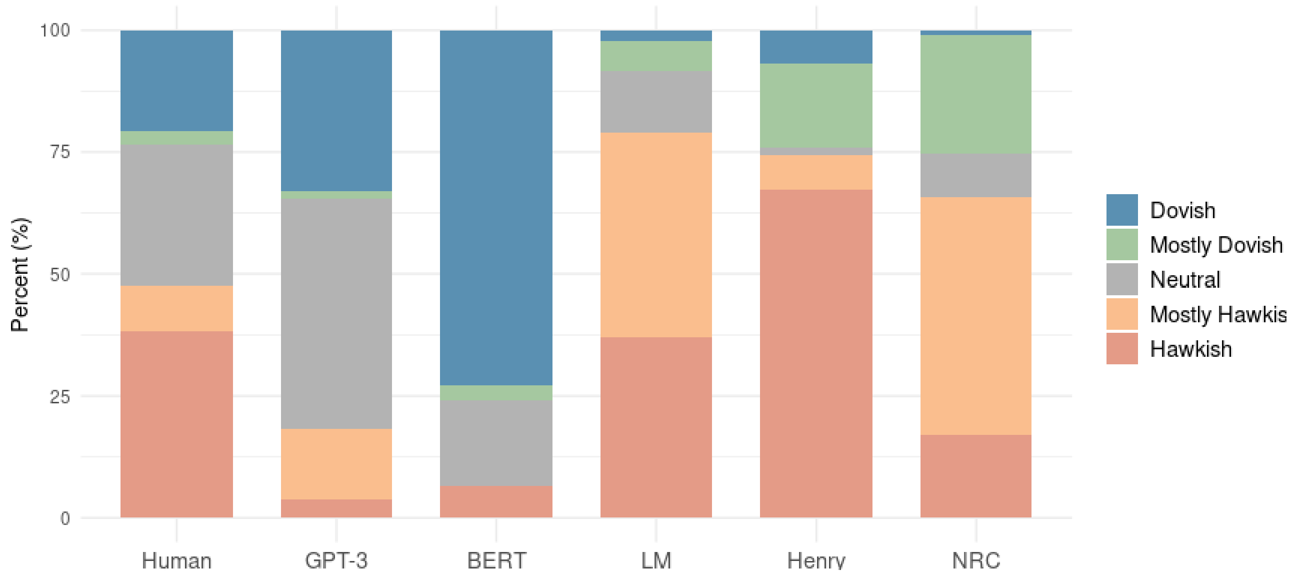
Figure 1 displays the distribution of labels across the classification methods. GPT-3 closely matches the human benchmark for the "dovish", "mostly hawkish", and "hawkish" labels, while BERT overestimates the number of "dovish" sentences and dictionary-based methods rarely label sentences

[7]In the supplementary material, we describe an additional analysis using GPT-3 with embeddings.

[8]We use the SENTENCE-TRANSFORMERS/PARAPHRASE-MPNET-BASE-V2 model in our analysis.

as “dovish” or “mostly dovish”. GPT-3 produces fewer “neutral” classifications than the human benchmark, possibly because humans, unlike algorithms, have a tendency to use this label when uncertain. Overall, GPT-3 outperforms other NLP methods based on this unconditional distribution.

Figure 1
Distribution of Labels by Method



The distribution of labels, however, does not inform us whether GPT-3 classifies each sentence in the same way as humans. To better assess performance, Table 3 shows mean absolute error (MAE), root mean squared errors (RMSE), and the following metrics from confusion matrices: (i) accuracy, the proportion of correct predictions; (ii) kappa, which measures the agreement between predictions and actual values accounting for the chance of agreement;[9] (iii) F1 score, the harmonic mean of true positive prediction proportions; and (iv) balanced accuracy, the average of true positive predictions among all positive cases and true negative predictions among all negative cases.

Overall, we see that GPT-3 obtains the lowest numerical errors, the highest accuracy, and the highest measure of agreement. Since accuracy can be misleading for imbalanced data sets such as ours, we pay particular attention to the F1 score and balanced accuracy, which also are highest for GPT-3 for most labels and similar to the other methods for other labels.

We also note that dictionary-based methods perform worse than transformer-based methods, i.e., GPT-3 and BERT, which is consistent with the previous literature. For example, Frankel et al. (2022) show that machine-learning methods outperform dictionary-based ones in capturing disclosure sentiment for 10-K filings and conference calls. Zhu et al. (2022) provide a comprehensive framework for sentiment analysis in finance, and evaluate 31 different methods using a sample of

[9]Kappa values range from -1 (total disagreement) to 1 (perfect agreement), with 0 indicating agreement due to chance alone.

20,000 Glassdoor company reviews. Results show that BERT models outperform other machine-learning models, which in turn surpass lexicon-based approaches.

Table 3
Performance Evaluation of Zero-Shot Learning

	GPT-3	BERT	LM	Henry	NRC
MAE	**0.41**	0.66	0.62	0.55	0.81
RMSE	**0.58**	0.84	0.80	0.75	0.96
Accuracy	**0.37**	0.25	0.28	0.35	0.11
Kappa	**0.18**	0.03	0.07	0.08	-0.04
F1 score					
Dovish	**0.49**	0.31	0.07	0.17	0.04
Mostly dovish	**0.43**	0.33	0.23	0.04	0.17
Neutral	0.15	0.13	0.48	**0.57**	0.14
Mostly hawkish	**0.36**	NA	0.15	0.07	0.11
Hawkish	**0.10**	0.07	NA	0.08	0.03
Balanced Accuracy					
Dovish	**0.71**	0.48	0.51	0.53	0.51
Mostly dovish	**0.56**	**0.56**	0.53	0.50	0.51
Neutral	0.54	0.51	**0.59**	**0.59**	0.45
Mostly hawkish	**0.67**	0.50	0.49	0.50	0.42
Hawkish	0.53	0.52	0.47	**0.56**	0.45

Notes: For each metric, the best performing model is boldfaced.

4.2 Fine-Tuned Learning

Next, we train the model using a subset of the human classified sentences. Specifically, we randomly draw 400 sentences to fine-tune the GPT-3, leaving 100 sentences for cross validation.
Table 4 reports the performance metrics of all methods on this test sample. First, note that consistent with the full-sample results described above, zero-shot GPT-3 overall outperforms BERT and the dictionary-based methods. Fine-tuning enhances the performance of GPT-3 even futher: the MAE is nearly half of that of zero-shot GPT-3, accuracy increases by almost a factor 1.5, and kappa more than doubles.

Table 4
Performance Evaluation of Fine-Tuned GPT-3

	GPT-3 (fine-tuned)	GPT-3 (zero-shot)	BERT	LM	Henry	NRC
MAE	**0.23**	0.40	0.60	0.58	0.54	0.85
RMSE	**0.40**	0.57	0.77	0.79	0.71	0.98
Accuracy	**0.61**	0.41	0.28	0.33	0.31	0.10
Kappa	**0.46**	0.21	0.01	0.15	0.00	-0.03
F1 score						
Dovish	**0.77**	0.48	0.34	0.07	0.06	0.07
Mostly dovish	**0.53**	0.45	0.31	0.34	0.07	0.26
Neutral	**0.66**	0.24	0.18	0.58	0.52	0.04
Mostly hawkish	0.22	**0.50**	NA	0.12	NA	0.11
Hawkish	**0.80**	NA	NA	NA	NA	NA
Balanced Accuracy						
Dovish	**0.83**	0.65	0.45	0.52	0.47	0.51
Mostly dovish	**0.67**	0.59	0.54	0.57	0.51	0.55
Neutral	**0.73**	0.57	0.53	0.67	0.52	0.40
Mostly hawkish	0.61	**0.80**	0.50	0.54	0.47	0.49
Hawkish	**0.99**	0.49	0.49	0.48	0.45	0.39

Notes: All algorithms are implemented on the test sample of 100 sentences.

5 Discussion and Conclusion

Mere classification aside, GPT models have the ability to explain why a certain sentence was labeled in a certain way, a capability beyond any existing NLP model and a valuable feature for researchers. We test this capability in a short exercise using ChatGPT and the underlying GPT-3 and GPT-4 models. We find that GPT's reasoning successfully justifies its classifications, and furthermore is very similar to the reasoning provided by a human reviewer.[10] GPT-4 offers an improvement over GPT-3 with more cases of agreement with the human classifications and explanations. The newest version of GPT-4 is therefore likely to generate even stronger performance metrics than those of GPT-3 reported in this paper.

The analysis presented in this paper shows that GPT models demonstrate a strong performance in classifying Fedspeak sentences, especially when fine-tuned. However, it is important to note that despite its impressive performance, GPT-3 is not infallible. It may still misclassify sentences or fail to capture nuances that a human evaluator with domain expertise might capture. Thus, while GPT models may not be able to fully replace human evaluators, they can serve as a highly valuable tool for assisting researchers and analysts in this domain.

[10]Detailed results are available in the supplementary material.

References

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. (2022). What Learning Algorithm Is In-Context Learning? Investigations With Linear Models. *arXiv preprint arXiv:2211.15661*.

Benchimol, J., Kazinnik, S., and Saadon, Y. (2020). Communication and Transparency Through Central Bank Texts.

Bertsch, C., Hull, I., Lumsdaine, R. L., and Zhang, X. (2022). Central Bank Mandates and Monetary Policy Stances: Through the Lens of Federal Reserve Speeches. *Sveriges Riksbank Working Paper Series*.

Biswas, S. S. (2023). Role of Chat GPT in Public Health. *Annals of Biomedical Engineering*.

Chen, K. (2016). Interpreting the Fedspeak: Text Analysis on FOMC Statements. *BBVA Research*.

Choi, J. H., Hickman, K. E., Monahan, A., and Schwarcz, D. (2023). ChatGPT Goes to Law School. *Minnesota Legal Studies Research Paper*, (No. 23-03).

Curti, F. and Kazinnik, S. (2021). Let's Face It: Quantifying the Impact of Nonverbal Communication in FOMC Press Conferences. (3782239).

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

Doh, T., Song, D., and Yang, S.-K. (2022). Deciphering Federal Reserve communication via text analysis of alternative FOMC statements. *Federal Reserve Bank of Kansas City Working Paper (forthcoming)*.

Dowling, M. and Lucey, B. (2023). ChatGPT for (Finance) Research: The Bananarama Conjecture. *Finance Research Letters*, 53.

Ehrmann, M. and Fratzscher, M. (2007). Communication by Central Bank Committee Members: Different Strategies, Same Effectiveness? *Journal of Money, Credit and Banking*, 39(2-3):509–541.

Ehrmann, M. and Wabitsch, A. (2022). Central Bank Communication with Non-Experts: A Road to Nowhere? *Journal of Monetary Economics*, 127:69–85.

Frankel, R., Jennings, J., and Lee, J. (2022). Disclosure Sentiment: Machine Learning vs. Dictionary Methods. *Management Science*, 68(7):5514–5532.

Frieder, S., Pinchetti, L., Griffiths, R.-R., Salvatori, T., Lukasiewicz, T., Petersen, P. C., Chevalier, A., and Berner, J. (2023). Mathematical Capabilities of ChatGPT. *arXiv:2301.13867v1*.

Geerling, W., Mateer, G. D., Wooten, J., and Damodaran, N. (2023). Is ChatGPT Smarter than a Student in Principles of Economics? *The American Economist*.

Gorodnichenko, Y., Pham, T., and Talavera, O. (2023). The Voice of Monetary Policy. *American Economic Review*, 113(2):548–84.

Hansen, S. and McMahon, M. (2016). Shocking Language: Understanding the Macroeconomic Effects of Central Bank Communication. *Journal of International Economics*, 99(S1):S114–S133.

Hansen, S., McMahon, M., and Tong, M. (2019). The Long-Run Information Effect of Central Bank Communication. *Journal of Monetary Economics*, 108:185–202.

Hayo, B. and Neuenkirch, M. (2015). Central Bank Communication in the Financial Crisis: Evidence From a Survey of Financial Market Participants. *Journal of International Money and Finance*, 59:166–181.

Henry, E. (2008). Are Investors Influenced by How Earnings Press Releases are Written? *The Journal of Business Communication*, 45(4):363–407.

Huang, A. H. and Hui Wang, Y. Y. (2022). FinBERT: A Large Language Model for Extracting Information from Financial Text. *Contemporary Accounting Research*.

Jegadeesh, N. and Wu, D. (2017). Deciphering Fedspeak: The information content of FOMC meetings. *Available at SSRN 2939937*.

Korinek, A. (2023). Language Models and Cognitive Automation for Economic Research. Technical report, National Bureau of Economic Research.

Leippold, M. (2023a). Sentiment Spin: Attacking Financial Sentiment with GPT-3. *SSRN Working Paper*.

Leippold, M. (2023b). Thus Spoke GPT-3: Interviewing a Large-Language Model on Climate Finance. *Finance Research Letters*, 53:103617.

Loughran, T. and McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.

Mohammad, S. and Turney, P. (2015). NRC Word-Emotion Association Lexicon.

Picault, M. and Renault, T. (2017). Words Are Not All Created Equal: A New Measure of ECB Communication. *Journal of International Money and Finance*, 79:136–156.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving Language Understanding by Generative Pre-training.

Shapiro, A. H. and Wilson, D. J. (2022). Taking the Fed at its Word: A New Approach to Estimating Central Bank Objectives Using Text Analysis. *The Review of Economic Studies*, 89(5):2768–2805.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is All You Need. *Advances in Neural Information Processing Systems*, 30.

Zaremba, A. and Demir, E. (2023). ChatGPT: Unlocking the Future of NLP in Finance. *Available at SSRN 4323643*.

Zhang, B., Ding, D., and Jing, L. (2022). How Would Stance Detection Techniques Evolve After the Launch of ChatGPT? *arXiv preprint arXiv:2212.14548*.

Zhu, Y., Hoepner, A. G., Moore, T. K., and Urquhart, A. (2022). Sentiment Analysis Methods: Survey and Evaluation. *Available at SSRN 4191581*.

6 Examples of Classified Sentences

Classification	Sentence
Dovish	The recent softness in inflation is a concern, and the Committee stands ready to act if necessary to ensure that inflation returns to its target over the medium term.
Dovish	The risks to the economic outlook remain tilted to the downside, and the Committee will closely monitor incoming data for any signs of a deterioration in the outlook.
Mostly dovish	The recent easing of financial conditions is welcome, and should help support the economic expansion over the medium term.
Mostly dovish	Incoming data suggest that the economy is performing well overall, but there are some areas of concern, including ongoing trade tensions and a slowing global economy.
Mostly dovish	Given the current state of the economy, the Committee believes that it will be appropriate to maintain the current target range for the federal funds rate for the foreseeable future.
Neutral	The balance of risks to the economic outlook appears roughly balanced at this time.
Neutral	The unemployment rate remained steady at 4.5%
Mostly hawkish	With the unemployment rate at historic lows and inflation near our target, we believe that some further tightening of monetary policy may be warranted in the coming months.
Mostly hawkish	The Committee is aware that financial imbalances could pose risks to the outlook, and we will be vigilant in monitoring these risks going forward.
Mostly hawkish	While the recent slowdown in economic growth is a concern, we believe that it is largely due to transitory factors and expect growth to pick up in the coming quarters.
Hawkish	Given the current state of the economy, the Committee believes that it will be appropriate to continue raising the target range for the federal funds rate at a gradual pace.
Hawkish	We believe that the risks to the economic outlook are skewed to the upside, and we will act preemptively if necessary to ensure that inflation remains under control.
Hawkish	While we acknowledge that some risks to the outlook remain, we believe that the current stance of monetary policy is too accommodative and that further tightening may be needed to keep inflation in check.

7 Case Study of Explanations: Bryson vs. ChatGPT

In this case study, we analyze selected sentences and compare human explanations provided by Bryson, a Research Associate in our group, with those generated by ChatGPT. Bryson is a 24-year-old male, known for his intelligence and curiosity. ChatGPT, a language model designed as a chatbot, comprises both GPT-3 and GPT-4 models. The comparison begins with the following prompt:

> Imagine you're an RA working for the Fed. You have a Master's degree in Economics. Your task is to classify the following 5 sentences into one of the five categories (dovish, mostly dovish, neutral, mostly hawkish, hawkish). These sentences are taken at random from the texts of FOMC announcements:
>
> 1. In light of the current shortfall of inflation from 2 percent, the committee will carefully monitor actual and expected progress toward its inflation goal.
>
> 2. Labor market conditions have shown some improvement in recent months, on balance, but the unemployment rate remains elevated.
>
> 3. The committee will closely monitor market conditions and is prepared to adjust its plans as appropriate.
>
> 4. The committee judges that the case for an increase in the federal funds rate has continued to strengthen but decided, for the time being, to wait for some further evidence of continued progress toward its objectives.
>
> 5. The committee expects to begin implementing its balance sheet normalization program relatively soon, provided that the economy evolves broadly as anticipated; this program is described in the June 2017 addendum to the committee's policy normalization principles and plans.

Below, we summarize the responses of each participant and offer a comparison for each sentence.

Sentence 1:

Name	Label	Explanation
Bryson	Dovish	This sentence emphasizes the current shortfall of inflation below the Committee's target, suggesting that loose monetary (low FFR or securities purchases) policies will be necessary to bring inflation up towards the Committee's target.
GPT-3	Neutral	This sentence states that the committee will monitor progress towards its inflation goal, without leaning towards any particular policy stance.
GPT-4	Mostly dovish	The sentence emphasizes the shortfall of inflation from the target and the committee's intention to monitor progress, suggesting a cautious approach and potential inclination towards easing monetary policy.

For the first sentence, Bryson and GPT-4 agree that the policy stance is dovish due to a shortfall of inflation from target. However, whereas Bryson interprets the content as an indicator of policy accommodation and classifies the sentence as "dovish", GPT-4 assigns more caution to the statement and classified the sentence as "mostly dovish". GPT-3 fails to recognize the dovish sentiment of the sentence and assigns a neutral policy stance.

Sentence 2:

Name	Label	Explanation
Bryson	Mostly dovish	This sentence reflects the Committee's ongoing concern for the strength of the labor market. While no policy action is specified, the implication is that the Committee will pursue expansionary policy until the unemployment rate decreases to a level the Committee finds acceptable.
GPT-3	Mostly dovish	Although there has been some improvement in labor market conditions, the fact that the unemployment rate remains elevated suggests that the economy is not yet fully recovered. This sentiment leans towards a more cautious or dovish stance.
GPT-4	Dovish	The sentence highlights improvements in the labor market, but also notes that the unemployment rate is still high, which implies that the committee may lean towards further easing to address unemployment concerns.

For this sentence, both algorithms agree with the Bryson about the direction of the policy stance as well as the explanation. Interestingly, whereas the explanation from GPT-4 is strikingly similar to that given by Bryson, their classifications are different ("mostly dovish" vs. "dovish").

Sentence 3:

Name	Label	Explanation
Bryson	Neutral	This sentence offers no insight into the Committee's view of the state of the economy nor the actions the Committee might take. Therefore, it is impossible to say whether it is hawkish or dovish.
GPT-3	Neutral	This sentence is a generic statement that the committee will monitor market conditions and adjust its plans as needed, without revealing any particular policy stance.
GPT-4	Neutral	The sentence indicates that the committee will monitor market conditions and is ready to adjust its plans as needed, but does not provide a clear direction as to whether they lean towards easing or tightening.

All three explanations share the idea that sentence 3 does not reveal any particular policy stance or preference for easing or tightening monetary policy.

Sentence 4:

Name	Label	Explanation
Bryson	Mostly hawkish	This sentence reveals the Committee's view that hawkish policies (i.e., an increase in the policy interest rate) will be necessary in the near term; however, this sentiment is tempered by the Committee's wait and see approach.
GPT-3	Mostly hawkish	The sentence acknowledges that the case for increasing the federal funds rate has strengthened, which implies a hawkish stance. However, the committee has decided to wait for further evidence of progress towards its objectives, which slightly softens its hawkishness.
GPT-4	Mostly hawkish	The sentence suggests that the case for raising the federal funds rate has strengthened, but the committee is waiting for additional evidence before taking action, implying a preference for tightening monetary policy in the near future.

There is striking agreement between Bryson and the models on both the classification and the underlying explanation for this sentence. All explanations refer to the case for raising the policy rate, but note that the committee will wait for further evidence.

Sentence 5:

Name	Label	Explanation
Bryson	Hawkish	The balance sheet normalization program involves reducing the securities on the Fed's balance sheet and reducing the reinvestment of principal payments. This action by the Committee will reduce the money supply and is a clear transition away from loose-money towards a more contractionary policy stance.
GPT-3	Neutral	This sentence simply states that the committee expects to begin implementing its balance sheet normalization program soon, without revealing any particular policy stance.
GPT-4	Hawkish	The sentence indicates the committee's expectation to start implementing its balance sheet normalization program soon, which is a clear signal of tightening monetary policy, as long as the economy evolves as anticipated.

The final sentence is rated "hawkish" by Bryson and GPT-4, but "neutral" by GPT-3. Bryson and GPT-4 have similar explanations for their classification involving the balance sheet normalization program as a clear transition away from monetary policy accommodation towards a more contractionary policy stance.

In summary, GPT models generally present a logic that successfully justifies their classifications, and furthermore is very similar to Bryson's reasoning. GPT-4 offers an improvement over GPT-3 with more cases of agreement with Bryson. For example, GPT-3 fails to capture the classification of the "dovish" and "hawkish" sentence.

8 Additional Analysis: Embeddings

This appendix describes additional analyses of GPT-3 and BERT. First, we run a classification with embeddings using GPT-3 on our data.[11] Second, we run a BERT-based SetFit model, a framework for few-shot fine-tuning of Sentence Transformers.[12]

For GPT-3 with embeddings, we first tokenize our sentences, and then convert them into a sequence of numerical values (i.e., embeddings) using the pre-trained GPT-3 model. These embeddings capture the contextual information of our sample. We then run a classification model, where the model takes the embeddings from the GPT-3 as input features and learns to classify the text based on the provided labels. In the GPT-3 with embeddings model, the results show varying performance across different classifications. Precision is highest for Hawkish (0.67) and Mostly Dovish (0.65) classifications, while Recall is highest for Neutral classification (0.83). The F1-score is highest for Neutral (0.65), followed by Mostly Dovish (0.59) and Hawkish (0.57). The model achieved an overall accuracy of 0.57.

For SetFit, we conduct a similar exercise, where we first tokenize and embed our sentences. We then produce a small number of examples (6 per classification class, or 30 in total), and run SetFit-based classification exercise. In terms of performance, the highest precision was achieved for Dovish classification (0.45), while the highest recall was for Hawkish classification (0.75). However, the F1-scores for all classifications were lower compared to the GPT-3 with embeddings model. The overall accuracy of this model was 0.30.

Comparing the performance of these two models, GPT-3 with embeddings outperforms SetFit model in terms of accuracy, F1-scores, and other performance metrics. The GPT-3 with embeddings model demonstrates higher precision, recall, and F1-scores for most classifications, while the few shot-BERT model struggles to achieve comparable performance even with the additional examples provided.

[11] We use TEXT-EMBEDDING-ADA-002 model to embed our data.

[12] SetFit Model Reference