Can ChatGPT Decipher Fedspeak?

Anne Lundgaard Hansen and Sophia Kazinnik*

October 25, 2023

Abstract

Abstract This paper investigates the ability of Generative Pre-trained Transformer (GPT) models to decipher Fedspeak, 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 announcements relative to human assessment. We show that GPT models deliver a considerable improvement in classification performance over other commonly used methods in natural language processing (NLP). We then demonstrate that the GPT models can provide explanations for its classifications that are on par with human reasoning. Finally, we show that the GPT-4 model can successfully identify macroeconomic shocks using the narrative approach of Romer and Romer (1989, 2023).

Keywords: Natural Language Processing (NLP), Generative Pre-trained Transformer (GPT), Artificial Intelligence (AI), Central Bank Communication, Monetary Policy Shocks.

JEL Code: E52, E58, C88.

*The authors are with the Quantitative Supervision and Research (QSR) 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, Mehdi Beyhaghi, Brian Ferrell, Yuriy Gorodnichenko, Robin L. Lumsdaine, Marcus Mølbak Ingholt, Alessandro Martinello, David H. Romer, and Christoffer Jessen Weissert for valuable feedback. We also thank seminar participants at the European Central Bank and conference participants at the Fourth New York Fed Conference on Fintech. We thank Bryson Alexander, Nadia Audzeichuk, and Ethan Butler for excellent research assistance. The views expressed in this paper do not necessarily reflect the views of the Federal Reserve Bank of Richmond or the Federal Reserve System.

"It is hard to imagine that we could train a computer to read Federal Reserve transcripts the way we do. [...] We thoroughly expect to be made largely redundant by computers eventually, but perhaps not for a few years to come."

Christina D. Romer and David H. Romer (2023)

1 Introduction

Generative Pre-trained Transformer (GPT) models, proposed by Radford et al. (2018), have received a tremendous amount of public attention in recent months for their ability to analyze and generate text. Since its release in November 2022, ChatGPT, an artificial intelligence (AI) chatbot, has become a prominent topic of discussion across digital platforms and academic fields alike.

In this paper, we empirically evaluate the ability of GPT models to decipher Fedspeak, the language used by the Federal Reserve (Fed) to communicate monetary policy decisions. While much progress has been made toward improving central bank communication, Fedspeak is still known to be notoriously difficult to understand and messages are often delivered in a convoluted manner.[1] We find that GPT models are able to decipher this complex and nuanced language.

Specifically, we show that GPT models accurately classify the policy stance of Federal Open Market Committee (FOMC) announcements against a manually labeled benchmark. We find that GPT models deliver a considerable improvement in classification performance over other commonly employed natural language processing (NLP) methods, including models that were previously considered cutting-edge tools in this domain, such as Bidirectional Encoder Representations Transformers (BERT) from Devlin et al. (2018). Among the considered NLP methods, the GPT model achieves the lowest numerical errors, the highest accuracy, and the highest measure of agreement relative to human classification.

Besides improved performance in classifying FOMC texts, GPT models set apart from

[1]Farber, A. (2013, April 19). Historical Echoes: Fedspeak as a Second Language. Liberty Street Economics.

existing methods by their ability to provide explanations and reasoning. We study this ability by soliciting and comparing the explanations behind the classification of a few selected sentences provided by a human research assistant with those generated by ChatGPT using both GPT-3 and GPT-4. The results show that both models generally present a logic that successfully justifies their classifications and is very similar to that of human reasoning. GPT-4 offers an improvement over GPT-3 with more correct classifications and improved reasoning.

Having established that GPT models can successfully classify Fedspeak and even provide reasoning for the classifications, we employ it on a more complex task. Specifically, we use it to identify monetary policy shocks using the narrative approach of Romer and Romer (1989; 2023). The narrative approach involves careful reading of FOMC texts to establish the policy makers' motivation in changing policy, which determines whether a policy change can be considered a shock. In their discussion of the future of the narrative approach, Romer and Romer (2023) express concerns delegating this task to non-experts and computers. We show that GPT models are getting close to successfully interpreting policy language to identify monetary policy shocks. Consequently, the era where computers can interpret Fed communication using the narrative approach may have arrived earlier than anticipated.

This result has considerable implications not only for the field of monetary economics, but for qualitative methods in general. Broadly, this technology has the potential to bring forth the renaissance of the narrative approach, as large language models (LLMs) are effectively redefining the constraints once established by manual analysis. What used to be a labor-intensive and error-prone process could now become automated and scalable.

Looking beyond the narrative approach to macroeconomic identification, accurate interpretation of policy messages is crucial for financial market participants, policy makers, and the general public. Our results suggest that GPT models are valuable tools in this pursuit. Establishing the ability of GPT models to decode Fedspeak also suggests that these tools can help improve the clarity, transparency, and effectiveness of policy communication. This is highly relevant in an era where central banks are becoming more focused on making their communication more accessible to the public. Blinder et al. (2022) argue that non-experts often do not receive the policy messages correctly, and they propose a solution where com-

munication is tailored to specific audiences. Our results, showcasing the performance of GPT models to interpret central bank communication, suggest that this technology can provide a feasible way to implementing such solutions.

The literature on applications of GPT models in the realm of economics and finance is growing rapidly. For example, Lopez-Lira and Tang (2023) evaluate the performance of ChatGPT in forecasting returns; Jha et al. (2023) use ChatGPT to extract managerial expectations of corporate policies from earnings call transcripts; Leippold (2023) uses GPT models to demonstrate the vulnerabilities of the use of dictionaries in NLP tasks; and 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.

This paper provides a first attempt at evaluating GPT models for the purpose of quantifying Fed communication. Hence, we contribute to a large literature that uses NLP to study the content and sentiment of central bank communication and its impact on the financial markets (Ehrmann and Fratzscher, 2007; Hansen et al., 2019; Hayo and Neuenkirch, 2015; Curti and Kazinnik, 2023) and the general public (Ehrmann and Wabitsch, 2022). Whereas earlier contributions quantify central bank texts based on topic modeling and sentiment analysis using pre-defined dictionaries such as the Loughran and McDonald (2011) dictionary (Chen, 2016; Hansen and McMahon, 2016; Jegadeesh and Wu, 2017; Benchimol et al., 2020), more recent papers use pre-trained LLMs, such as the BERT models (Doh et al., 2022; Bertsch et al., 2022; Gorodnichenko et al., 2023). In this paper, we compare the performance of these popular methods with that of GPT models, and we establish that GPT models outperform these previously used methods.

2 GPT Models and ChatGPT

The GPT model series belongs to a family of pre-trained LLMs, i.e., AI algorithms that use deep learning and are trained on a massively large set of data. Specifically, GPT models are trained using a transformer architecture with self-attention, allowing them to capture nuanced language understanding (Zhang et al., 2022).

With each iteration of the GPT models, they have increased in size and complexity.

GPT-3 was initially released in June 2020 with important version updates released in March 2022 and March 2023.[2] GPT-4, released on March 14, 2023, is currently the most advanced version of the OpenAI language models. This model sets a new milestone in deep learning development, ensuring greater reliability and trustworthiness by generating more factually accurate statements than, e.g., the GPT-3 models.

In the analysis we use both the GPT-3 and GPT-4 model suites, as well as ChatGPT, an interactive chatbot that utilizes both GPT versions. Specifically, the GPT-3 model is used for the classification exercise,[3] we use both GPT-3 and GPT-4 to assess the ability of the models to explain their classifications, while we only rely on the GPT-4 model for the narrative approach because of its superior performance compared to GPT-3. We study the models in a zero-shot setup, i.e., without providing additional training. For comparison, the online appendix also reports classification results for the GPT-3 model when fine-tuned using a subset of our labeled data.[4]

3 Classifying Policy Stance

This section presents the analysis of the ability of GPT models to classify the policy stance of FOMC public communication.

3.1 Data

The Federal Open Market Committee (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 policy decisions. Unlike minutes and transcripts, FOMC statements are released immediately after the meetings, making them key elements in the Fed's communication to the public. We therefore choose to study these texts in our analysis.

[2]In November 2022, OpenAI started to refer to this model as belonging to the "GPT-3.5" series.

[3]The GPT-4 model was not released at the time we conducted this exercise

[4]The online appendix is available at Hansen's website (direct link).

Our analysis focuses on FOMC statements published between 2010 and 2020, which we divide into a set of individual sentences. We manually annotate each sentence with respect to its policy stance, i.e., whether it expresses a belief that the economy may be growing too slowly and need monetary policy stimuli or that the economy may be growing too quickly and need to be slowed down through monetary policy. Naturally, this definition of policy stance is correlated with sentiment. We measure policy stance on a discrete scale using five categories: "dovish", "mostly dovish", "neutral", "mostly hawkish", and "hawkish". Table 1 provides their definitions along with numerical values that we assign to the categories on a scale of -1 to 1, where 0 represents a neutral stance.[5] We employ five categories instead of simply "dovish", "neutral", and "hawkish" to examine the GPT models' ability to discern subtle differences between closely related labels, a common challenge in machine learning.

[Insert Table 1 about here]

The manual annotation is completed by research assistants from the Federal Reserve Bank of Richmond, all with educational backgrounds in economics, finance, or political science. To mitigate the risks of human bias and error, each sentence is processed independently by three reviewers, and the final label is computed as the average given the assigned numerical value for each category.[6] When classifying a sentence, only the context within the confines of each sentence is considered. The reviewers are thus not informed what statement the sentence comes from, including the date of the meeting that the sentence relates to. Since manual classification is time consuming and costly, we use a subset of 500 sentences which are drawn randomly from the full set of sentences using uniform selection. The summary statistics of the manual classifications and the disagreement among reviewers are provided in Table 2.

[Insert Table 2 about here]

[5]The online appendix provides examples of sentences and their classifications, which were given to the research assistants prior to them completing the classification task.

[6]Overall, the work was distributed between four reviewers.

We note that our sample is imbalanced in the sense that there are more sentences with dovish sentiment than with hawkish sentiment, which is a result of the sample period in question. The human reviewers agree most on the classification of "mostly dovish", "neutral", and "mostly hawkish" sentences.

3.2 Method

We use the TEXT-DAVINCI-003 version of the GPT-3 model, which we prompt through the application programming interface (API). The API allows us to control the temperature, which we set to zero to maximize determinism of model output.

To benchmark the results of the GPT models, we also classify the FOMC sentences using the BERT model[7] and dictionary-based methods. Similar to GPT models, BERT is a pre-trained LLM based on the transformer architecture. It has achieved state-of-the-art results on many benchmark data sets (Bertsch et al., 2022; Huang and Hui Wang, 2022). BERT models differ from GPT models in terms of the specifications of their transformer architectures and the way they are pre-trained.

Dictionary-based methods are based on word counts using pre-defined lexicons to label the sentiment of each word. These methods are popular for their simplicity and transparency, but their performance is limited by the coverage of the lexicons and they struggle with nuances and context. Therefore, dictionary-based analyses often use multiple dictionaries to cross-validate the findings. For our analysis, we use the Loughran and McDonald (LM, 2011) and Henry (2008) dictionaries, which are both finance-specific, along with the NRC Word-Emotion Association Lexicon of Mohammad and Turney (2015), which is based on everyday language.[8]

---

[7]Specifically, we use the SENTENCE-TRANSFORMERS/PARAPHRASE-MPNET-BASE-V2 model.

[8]To provide dictionary-based policy stance classifications, we use the dictionaries to compute a sentiment score for each sentence as the ratio of the difference in number of positive and negative words to the total number of words. Thus, we rely on the strong correlation between our definition of policy stance and sentiment. Then, we normalize the score so that it belongs to the $(-1, 1)$ interval and has a median equal to zero. Finally, the categories are defined based on the sentiment score s_i for sentiment i as "dovish" $s_i \in (-1, -0.5($, "mostly dovish" if $s_i \in (-0.5, 0($, "neutral" if $s_i = 0$, "mostly hawkish" if $s_i \in)0, 0.5)$, and

3.3 Results

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. BERT overestimates the number of "dovish" sentences and dictionary-based methods rarely label sentences as "dovish" or "mostly dovish". Both the GPT-3 and BERT models produce fewer "neutral" classifications than the human benchmark, possibly because humans, unlike algorithms, have a tendency to use this label when uncertain.

[Insert Figure 1 about here]

Table 3 shows error metrics and performance metrics based on the confusion matrix. Overall, the GPT-3 model obtains the lowest numerical errors, the highest accuracy, and the highest measure of agreement. The MAE and RMSE are 0.41 and 0.58, which are comparable with the errors associated with disagreement among reviewers resulting in an MAE of 0.27 and RMSE of 0.62, see Table 2. Since accuracy can be misleading for imbalanced data sets such as ours, we pay particular attention to the F1 score and balanced accuracy, which are highest for GPT-3 for most labels and similar to the other methods for other labels. Consistent with the previous literature (Frankel et al., 2022; Zhu et al., 2022), our results confirm that transformer-based LLMs (GPT-3 and BERT) outperform dictionary-based methods.

[Insert Table 3 about here]

Mere classification aside, GPT models can explain why a certain sentence was labeled in a certain way, a capability beyond any existing NLP model. To explore this capability, we ask both ChatGPT, using both GPT-3 and GPT-4 models, and a human research assistant, Bryson, to classify and provide explanations for their classifications for a few selected sentences.[9]

"hawkish" if $s_i \in)0.5, 1)$.

[9]The prompt and detailed results from this exercise are provided in the online appendix.

The results show that the GPT models generally present a logic that successfully justifies their classifications, even when their classifications disagree with those of Bryson. For example, the sentence,

> *"In light of the current shortfall of inflation from 2 percent, the committee will carefully monitor actual and expected progress toward its inflation goal"*

is categories as "neutral" by GPT-3 with the explanation,

> *"This sentence states that the committee will monitor progress towards its inflation goal, without leaning towards any particular policy stance"*

Although this sentence was classified as "dovish" by Bryson, the GPT-3 model does argue in a way that is consistent with its assigned label.

The GPT-4 model offers an improvement over GPT-3 in the sense that it has more cases of agreement with Bryson both in terms of classifications and explanations. For instance, for the above sentence, Bryson and GPT-4 provide the following explanations:

GPT-4: *"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.*"

Bryson: *"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"*

4 Automating the Narrative Approach

Given the ability of GPT models to interpret Fed communication, we hypothesize that they may also have the capacity to accurately implement the narrative approach to macroeconomic identification.[10] The narrative approach, pioneered by Friedman and Schwartz (1963), and

[10]We thank Yuriy Gorodnichenko for this idea.

extended and formalized by Romer and Romer (R&R, 1989), sets out to uncover monetary policy shocks from rigorous manual analysis of qualitative data, e.g., transcripts and minutes of FOMC meetings. The manual treatment of these texts is resource and time consuming.[11] GPT models can reduce this burden tremendously and thus have the potential to make the narrative approach easier and cheaper to implement. In addition, we note that GPT models satisfy the requirements for rigorous narrative analysis as outlined in R&R (2023, Table 1): (i) well-constructed prompts can give the GPT models a clear idea of what to look for in the source; (ii) GPT models approach the source without passion and in a consistent manner, especially when controlling the temperature; and (iii) documenting the work is straightforward and GPT models can provide explanations for their reasoning.

4.1 Method

Similar to R&R (1989) and R&R (2023), we use FOMC transcripts from 1946 to 2017, and minutes for years 2017 through 2023, where transcripts are not available.[12]

We construct a prompt that reflects the principles outlined in R&R (2023).[13] First, the prompt clearly outlines the criteria for identifying a monetary policy shock: policy makers believing the economy was at potential output, changing money growth and interest rates due to high inflation, and understanding and accepting the potential adverse consequences for output and unemployment. Second, it asks for an analysis based on the provided criteria, thus trying to avoid any bias or preconceived notions. Finally, the prompt asks for a detailed explanation of why the provided text does or does not meet the criteria for a monetary policy shock.

Since most shocks identified by R&R (1989; 2023) are contractionary in nature, we focus only on these and leave the identification of expansionary shocks for future work.

[11] As R&R (2023) note, *"there are roughly 50 to 100 pages of material per [FOMC] meeting —so, with eight to twelve meetings per year (or in some periods even more), we are talking about a lot of information (and reading!)."*

[12] Transcripts are not available for the full sample because the Fed releases transcripts of FOMC meetings with a five-year lag.

[13] The prompt is available in the online appendix.

We assess each document in our sample one by one, in no particular order, following the method of R&R as closely as possible. We employ the ChatGPT user interface using the GPT-4 model, augmented with the PDF file plugin to overcome token limits imposed by OpenAI. To improve reliability of responses, we query each document ten times as ChatGPT is not deterministic (Reiss, 2023). For illustrative purposes, we include several examples of the model responses in the online appendix.

4.2 Results

We first test whether the GPT-4 model can identify the same monetary policy shocks as R&R (1989) and R&R (2023). Then, we discuss additional shocks identified by the GPT-4 model, and which were not present in R&R (1989) or R&R (2023). All responses that capture a shock are listed in the online appendix.

4.2.1 Comparison with R&R Shocks

Table 4 lists the shocks identified by the R&R studies alongside with those identified by the GPT-4 model.

[Insert Table 4 about here]

Both R&R studies concur on six contractionary shocks during the period, namely in October 1947, August 1955, December 1968, April 1974, August 1978, and October 1979. They disagree on other shocks. The 1989 study finds a shock in September 1955 not recognized in the 2023 study. Conversely, R&R (2023) identify additional shocks in September 1958, May 1981, and December 1988, not found in R&R (1989). Furthermore, having access to a more recent sample, R&R (2023) discover a shock in June 2022.

The GPT-4 model overall generates comparable results to R&R. It identifies several contractionary policy shocks in the sample, but does not find evidence for shocks in the majority of the documents. Specifically, as in R&R, the model identifies shocks in August 1955, December 1968, April 1974, October 1979, May 1981, December 1988, and June 2022 with 1955, 1974, and 1981 being classified as *maybe* being monetary policy shocks.

The shocks on which R&R (2023) and GPT-4 agree all share the following elements: discussions of policy makers perceiving the economy as operating at or near potential output, often accompanied by remarks on the strength of the current economy or its post-recession expansion phase; discussions of changes in monetary policy in response to high inflation rates, typically manifested in shifts in money growth, interest rates, and strategic deceleration of bank reserves and money supply expansion; and evidence for policy makers' awareness and acceptance of possible adverse effects of their policy decisions on output and unemployment, expressed through an acknowledgment of the risk of market reactions or a significant shift in economic expectations.

For the meetings that *maybe* contain a policy shock (August 1955, April 1974, and May 1981), the associated transcript texts are categorized by GPT-4 as having some, but not all criteria for monetary policy shock present. They feature some elements of a policy shock, but not consistently or not strongly enough to qualify as definite instances. These *"maybes"* present a challenge in defining the exact boundaries of what constitutes a policy shock.

The model disagrees with both R&R studies on the October 1947 and August 1978 shocks, and it doesn't agree with R&R (2023) that September 1958 should be added to the list of shocks. Furthermore, R&R (1989) concludes that there is a shock in September 1955, but the model doesn't identify a shock during this month. The model does, however, find a shock in November 1955. These are likely to be related; as noted in R&R (2023), shocks do not appear suddenly but are results of gradual changes in views. The R&R approach uses the earliest date at which they can argue that the shock criteria are satisfied. It is likely that R&R and GPT-4 simply disagree about the timing of this shock.

To understand the disagreement around these shocks, we revisit the model. Specifically, we query the transcripts of the 1947, 1958, 1978, and September 1955 meetings and ask the model to provide evidence as to why these texts could not be classified as containing a monetary policy shock. Based on the model output, we argue that these instances are not characterized as policy shocks due to the absence of one or more critical components in the text. For instance, in most of these case, the evidence for the belief that the economy was operating at full potential is missing. Specifically, the analysis of the 1947 and 1958 transcripts shows no explicit statements supporting this belief. Likewise, the analysis of the

September 1955 transcript notes a discussion on the deceleration in economic expansion. Additionally, these texts lack the discussion of changes in monetary policy and recognition of potential adverse consequences on output and unemployment.

4.2.2 New Shocks Identified by GPT-4

Our analysis also reveals a series of shocks that were not previously identified in R&R (1989, 2023). Specifically, we discover six distinct shocks that are exclusive to the GPT-4 model, as outlined in Table 5.

[Insert Table 5 about here]

As detailed in the table, the documents identified as containing shocks are transcripts of the FOMC meetings taking place in March 1957, March 1968, April 1968, May 1969, June 1969, August 1979, September 1979, November 1979, December 1980, and March 1997. According to the R&R (2023) approach, the way to date a shock is to place the date of the shock at the earliest point it is identified, and have the criteria satisfied for at least a few meetings afterwards. So, the shocks occurring within the same quarter should not be considered as separate or new shocks, but rather grouped as one. Therefore, we identify six distinct shocks in total: March 1957, March 1968, May 1969, August 1979, December 1980, and March 1997.

We stress that these results are a function of the prompt and the input text. Further guidance in the prompt may change the model's conclusions on these dates. Specifically, the model is not instructed to distinguish between attempts to prevent inflation from rising and attempts to lower inflation. As such, the model explains its reasoning using quotes from the input text that refers to curbing inflation or resisting inflationary pressures, e.g., in 1957: *"Shepardson did not think the Committee should accept inflation as inevitable, and it should take every step that it could to curb such a development"* and in 1997: *"I see a significant risk of an increase in underlying inflation in the years ahead"*. In contrast, R&R only consider policy tightening as shocks if the motivation stems from the need to reduce inflation.[14]

[14]We thank David H. Romer for valuable discussion on this point.

4.2.3 Discussion

Do these impressive results mean that experts and researchers are now obsolete? On the contrary. Our results show that there is a tremendous potential for boosting the capabilities of researchers in the realm of qualitative analysis, as the usual barriers to this method have been lowered.

However, it is important to reiterate that these tools are not infallible, and their output depend crucially on the prompt, whose specification is non-trivial. If a prompt is designed to prioritize certain criteria or guide the model towards specific features, it can produce more nuanced responses. However, a prompt that is too narrow or overly focused can potentially cause the model overlook broader information.

Can the narrative approach be delegated? R&R (2023) argue: *"at some point, in some cases, if it is done very carefull"*. Given our findings, we argue that the answer can be updated to *"probably now, in a lot of cases, but still very carefully"*. Qualified researchers are still needed to determine which sources to use, what type of information can be extracted from the source, and in what shape. While we were able to construct a prompt with which the GPT-4 model identifies the majority of shocks found by R&R (1989, 2023), we relied heavily on their work for that.

5 Concluding Remarks

The analyses presented in this paper show that GPT models demonstrate a strong performance in interpreting Fed communication. First, the GPT-3 model classifies the policy stance of FOMC announcements more accurately than other NLP methods, and GPT models can explain their classifications using arguments similar to a human benchmark. Second, the GPT-4 model is mostly successful in identifying monetary policy shocks from FOMC transcripts and minutes as in R&R (1989, 2023).

Overall, the performance of GPT models fundamentally transforms our understanding of what qualitative analysis can achieve. It opens up possibilities for a new era of research that combines the depth of human insight with the breadth and speed of AI. As we continue to harness and refine these tools, we can expect to see an even deeper impact on how we

analyze and understand complex phenomena, ranging from economic policy to other areas demanding thorough textual analysis. While GPT models may not be able to fully replace human evaluators, they can serve as a highly valuable tool for assisting researchers in this domain.

References

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

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

Blinder, A., M. Ehrmann, H. de Haan, and D.-J. Jansen (2022): "Central Bank Communication with the General Public: Promise or False Hope?" .

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

Curti, F. and S. Kazinnik (2023): "Let's Face It: Quantifying the Impact of Nonverbal Communication in FOMC Press Conferences," *Journal of Monetary Economics*.

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

Doh, T., D. Song, and S.-K. Yang (2022): "Deciphering Federal Reserve Communication via Text Analysis of Alternative FOMC Statements," *Federal Reserve Bank of Kansas City Working Paper (forthcoming)*.

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

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

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

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

Friedman, M. and A. J. Schwartz (1963): *A Monetary History of the US 1867-1960*, Princeton University Press.

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

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

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

Hayo, B. and M. Neuenkirch (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, 363–407.

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

Jegadeesh, N. and D. Wu (2017): "Deciphering Fedspeak: The Information Content of FOMC Meetings," *Available at SSRN 2939937*.

Jha, M., J. Qian, M. Weber, and B. Yang (2023): "ChatGPT and Corporate Policies," *Chicago Booth Research Paper*, 23-15.

Korinek, A. (2023): "Language Models and Cognitive Automation for Economic Research," Tech. rep., National Bureau of Economic Research.

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

Lopez-Lira and Tang (2023): "Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models," *SSRN Working Paper*.

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

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

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

Reiss, M. (2023): "Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark," *arXiv preprint arXiv:2304.11085*.

Romer, C. D. and D. H. Romer (1989): "Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz," *NBER Macroeconomics Annual*, 4, 121–170.

——— (2023): "Presidential Address: Does Monetary Policy Matter? The Narrative Approach after 35 Years," *American Economic Review*, 113, 1395–1423.

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

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

Figures

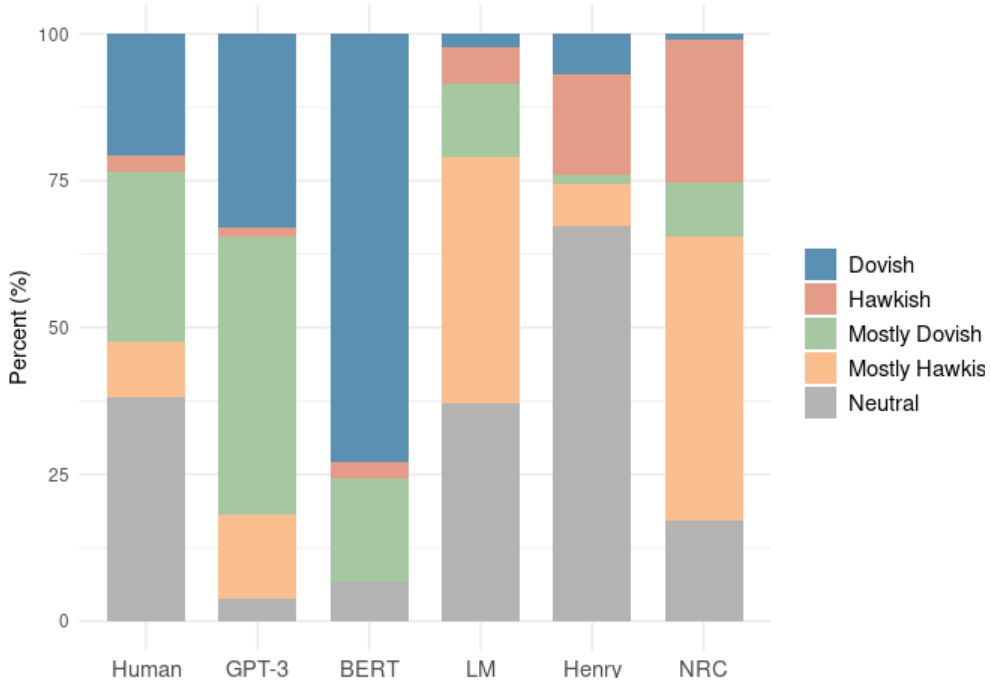

Figure 1: Distribution of categories by method

Tables

Table 1: Definitions of categories and assigned numerical values

Category	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.

Table 2: Summary statistics of manually classified data

	Total	Dovish	Mostly Dovish	Neutral	Mostly Hawkish	Hawkish
Count	500	104	144	191	47	14
MAE	0.27	0.41	0.28	0.18	0.28	0.42
RMSE	0..62	0.61	0.40	0.33	0.40	0.61

Notes: Summary statistics are computed using the numerical values assigned to the classifications in Table 1. MAE and RMSE are computed based on the differences between each reviewer's classification and the final assigned classification.

Table 3: Classification performance evaluation

	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.

Table 4: Contractionary monetary policy shocks from 1946–2023 as identified by Romer and Romer (1989, 2023) and the GPT-4 model

R&R (2023)	R&R (1989)	GPT-4
Oct. 1947	Oct. 1947	
Aug. 1955		Aug. 1955*
	Sept. 1955	
		Nov. 1955
Sept. 1958		
Dec. 1968	Dec. 1968	Dec. 1968
Apr. 1974	Apr. 1974	Apr. 1974*
Aug. 1978	Aug. 1978	
Oct. 1979	Oct. 1979	Oct. 1979
May 1981		May 1981*
Dec. 1988	Dec. 1988	Dec. 1988
June 2022		June 2022

Notes: For shocks denoted with $*$, the GPT-4 model concludes that there *maybe* was a policy shock.

Table 5: Contractionary shocks from 1946–2023 identified by the GPT-4 model only

Shock 1	Shock 2	Shock 3	Shock 4	Shock 5	Shock 6
Mar. 1957	Mar. 1968	May 1969	Aug. 1979	Dec. 1980	Mar. 1997
	Apr. 1968	June 1969	Sep. 1979		
			Nov. 1979		

Notes: Each column represent a shock; the rows list the associated FOMC meeting dates for which the GPT-4 model identified policy shocks.