

Can ChatGPT Decipher FedSpeak?

Anne Lundgaard Hansen and Sophia Kazinnik*

April 10, 2024

Abstract

Abstract This paper examines the ability of large language models (LLMs) to decipher FedSpeak, the technical language used by the Federal Reserve to communicate on policy decisions. We evaluate the ability of the GPT-3.5 and GPT-4 models to correctly classify the policy stance of FOMC announcements relative to human assessment. We find that these models outperform traditional methods in classification accuracy and provide justifications akin to human rationale. Additionally, we show that the GPT-4 model can successfully perform the elaborate and non-trivial task of identifying macroeconomic shocks using the narrative approach of Romer and Romer (1989, 2023). Finally, we show preliminary evidence that market reactions to FOMC announcements have intensified following the introduction of the GPT-4 model.

Keywords: Large Language Models (LLMs), Artificial Intelligence (AI), Financial Markets, Central Bank Communication.

JEL Codes: 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, Jonathan Benchimol, Mehdi Beyhaghi, Brian Ferrell, Yuriy Gorodnichenko, Robin L. Lumsdaine, Marcus M. Ingholt, Alessandro Martinello, David H. Romer, and Christoffer J. Weissert for valuable feedback. We thank participants at the European Central Bank, Danmarks Nationalbank, Bank of Israel, and the Fourth New York Fed Conference on Fintech for useful comments. We thank Bryson Alexander, Nadia Audzeichuk, Ethan Butler, and Nicole Lindsay 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

Given the importance of the Federal Reserve in determining asset prices, market participants heavily scrutinize Fed’s communications for insights into future policy rates and the context for policy decisions. The methods with which these communications are scrutinized have evolved from simple manual text categorization to leveraging more sophisticated tools such as natural language processing (NLP) methods and machine learning models.¹

Large language models (LLMs) and in particular ChatGPT, an LLM-driven chatbot released by OpenAI in November 2022, have accelerated this trend.² LLMs set a new standard for a wide range of natural language tasks. They are trained on massive amounts of data to predict the likelihood of a given word or phrase following a sequence of words in a text. This design enables them to extract relevant signals from texts in specific domains, even if they were not explicitly trained on data from those domains. Furthermore, a tool like ChatGPT enables anyone with a computer and a network connection to use these cutting-edge models. Since not every market participant is a hedge fund with access to the most advanced machine learning tools for signal extraction, the introduction of ChatGPT levels the playing field. LLMs, and in particular ChatGPT, are therefore natural candidates for automating the interpretation of Fed communications, and they could be transformative in making sophisticated economic analysis more accessible.

In this paper we empirically evaluate the GPT-3.5 and GPT-4 models as tools for interpreting Fed communications. By interpretation, we mean not only classification of policy

¹For example, economists from Aberdeen suggests using NLP to decode Fed communication (Aberdeen, Global Macro Research, “NLP: Using machine learning to decipher Fed speak”, November 4, 2022).

²Large language models (LLMs) are artificial intelligence systems designed to understand, interpret, and generate natural language. Examples include OpenAI’s GPT (Generative Pre-trained Transformer) series.

stance, but also considerably more complex tasks such as reasoning and in-depth policy analysis. Our study is the first to systematically evaluate the effectiveness of LLMs in interpreting and analyzing Fedspeak. We provide evidence for the considerable improvement of these new tools over the techniques previously used in the literature for analyzing texts in finance.

We start by establishing a performance benchmark of GPT models in the realm of Fed communication. We show that GPT models accurately classify the policy stance of Federal Open Market Committee (FOMC) announcements against a manually labeled benchmark, as well as deliver a considerable improvement in classification performance over other commonly used NLP methods, including models that were previously considered cutting-edge tools in this domain.³ Among the considered methods, the GPT models achieve the lowest numerical errors, highest accuracy, and highest measure of agreement relative to human classification.

The importance of this initial experiment is not in creating policy stance labels with GPT models per se – given the brevity of FOMC texts, manual reading is in fact feasible.⁴ Rather, the exercise demonstrates that GPT models can understand the specialized language used by the Fed without additional customization or training, paving the way for their application in interpreting central bank communications and policy-related texts.

Besides improved performance in classifying FOMC texts, GPT models set apart from 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 sub-sample of sentences provided by a human research assistant with those generated by ChatGPT. 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-3.5 with more correct classifications and improved reasoning.

³Such as Bidirectional Encoder Representations Transformers (BERT) from Devlin et al. (2018)

⁴Although their ability to do so suggests analysts could more swiftly interpret FOMC announcements and similar documents. Underscoring the value of this contribution, both Morgan Stanley and J.P. Morgan launched AI algorithms in early 2023 aimed at decoding FOMC communication to extract signals to investors (Morgan Stanley, “Sensing the Fed’s Direction with the Help of AI”, April 25, 2023; Bloomberg, “JPMorgan Creates AI Model to Analyze 25 Years of Fed Speeches”, April 26, 2023).

We then employ the GPT-4 model on a considerably more complex task, namely the identification of monetary policy shocks using the narrative approach of Romer and Romer (1989; 2023). The narrative approach involves careful manual reading of FOMC texts to establish the policy makers' motivation in changing policy, which determines whether a policy change can be considered a shock. This task is highly demanding and time-consuming, and as a result, the narrative approach is a huge undertaking to implement. The narrative approach is therefore an ideal use case for LLMs: if these models can perform such tasks on a level comparable to human experts, we can use them to enhance speed and efficiency in the implementation of the narrative approach.

Romer and Romer (2023) discuss this option but express concerns delegating this task to computers. In contrast, we show that GPT models are getting close to successfully completing this task. 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 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.

Overall, these results suggest that GPT models are valuable tools for accurate interpretation of policy messages from the Fed. While advanced machine learning tools for analyzing central bank communications have been around for decades, their high cost and complexity have mostly restricted their use to institutional investors, putting retail investors at a disadvantage. Thus, the availability of ChatGPT is a potentially transformative development for the financial sector offering professionals an on-demand virtual expert on Fed communications. The question then becomes whether ChatGPT was adopted by a large share of market participants as a tool in this pursuit. To tentatively answer this question, we examine the change in trading volume around the FOMC announcement release before and after the introduction of the GPT-4 model. We find that market reactions to FOMC announcements have intensified in the period after GPT-4's introduction, which might be due to faster dissemination of information, leading to quicker trading decisions. We leave a more thorough

study of this question for future research, while waiting for more data to become available.

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 communication 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 models can be programmed to customize its responses to suit different demographic groups, thereby addressing the disparities in monetary policy knowledge. As such, by harnessing the capabilities of the LLM technology, central banks can enhance the effectiveness and inclusivity of their communications, leading to an improved public understanding of their policies and actions.

We contribute to the rapidly growing literature on applications of GPT models in the realm of economics and finance. 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; Cook et al. (2023) use GPT models to infer the tone and topic of discussion from corporate earnings calls. We also 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). This paper, on the other hand, provides a systematic evaluation of GPT models for the purpose of quantifying Fed communication.

We proceed as follows. Section 2 describes the GPT models and ChatGPT. Section 3 describes the setup and results from the classification exercise, including results on the

ability of GPT models to explain their classifications. Section 4 presents the application of the GPT-4 model for the narrative approach. Section 5 discusses whether market reactions to FOMC announcements changed following the introduction of ChatGPT. Finally, Section 6 concludes.

2 GPT Models and ChatGPT

Generative Pre-trained Transformer (GPT) models (Radford et al., 2018) have received a tremendous amount of public attention in the past year for their ability to analyze and generate human-like text. The GPT model series belongs to a family of pre-trained LLMs, i.e., artificial intelligence 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.⁵ GPT-3 was initially released in June 2020 with important version updates released in March 2022 and March 2023. On March 15, 2022, OpenAI released a new version of GPT-3 called “TEXT-DAVINCI-003”. This model was described as more capable than previous versions. In November 2022, OpenAI started to refer to this model as belonging to the “GPT-3.5” series. The latest version of the GPT-3.5 model, the “GPT-3.5-TURBO”, was released on March 1, 2023 – and it has instantly caused a spike in interest in GPT-3.5. 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, for example, the GPT-3 models.

In the analysis we use both the GPT-3.5 and GPT-4 model suites, as well as ChatGPT, an interactive chatbot that utilizes both GPT versions. We use the GPT-4 model only for the

⁵GPT-1 had 117 million parameters. The second iteration, GPT-2, had 1.5 billion parameters. GPT-3.5 is an order of magnitude larger than GPT-2, and GPT-4 is reportedly about six times larger than GPT-3, with approximately one trillion parameters.

narrative approach because of its superior performance compared to GPT-3.5. We study the performance of these models in a zero-shot setup, i.e., without providing additional training. For comparison, the online appendix also reports classification results for the GPT-3.5 model when fine-tuned using a subset of our labeled data.⁶

2.1 The Training Sample Critique

Applications of LLMs are often criticized for prompting the models with questions to which the training data contain the answers. A similar critique could be raised for the applications presented in this paper: at the time we employed the models, the GPT-3.5 model was trained on data that were available on the internet up until September 2021, and the GPT-4 model was trained on even more recent data. The training data therefore contains information on the policy stance for each FOMC meeting as well as the Romer and Romer (1989) paper.⁷

One way to address whether this critique is a substantial concern is to simply ask the GPT models to repeat back results from the training data. For example, Horton (2023) asks the GPT-3.5 model: “In Charness and Rabin (2002), which fraction of respondents chose “Left” in the Berk29, Berk 26, Berk23 and Barc2 scenarios?” The model answers, “In the Berk29, Berk 26, Berk23, and Barc2 scenarios, approximately 59%, 57%, 43%, and 16% of respondents chose “Left,” respectively,” when the actual set of results was 31%, 78%, 100%, and 52%. These findings mitigate the training sample concern. How so?

GPT models do not work by recollecting information. Rather than looking up answers to the prompted questions in the training data, LLMs find answers based on *patterns* learned from the training. If GPT models were simply reading answers from the training data, they would not be subject to hallucination, a well-known risk of using these models for finding facts. In this way, LLMs can be compared to human brains. For example, a human born in 1987 has been exposed to a lot of information since birth. But, when asked a specific question, e.g., the policy stance of the announcement following the March 2007 FOMC

⁶The online appendix is available at Hansen’s website ([link](#)).

⁷The training data set does not contain the Romer and Romer (2023) paper and has no information on the monetary policy shock identified in June 2022.

meeting, it is unlikely that this person will *know* the answer based on historical facts. It seems more likely that the person will read the provided announcement text and interpret a policy stance. LLMs work in a similar way.

3 Classifying Policy Stance

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.

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.⁸ 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.

⁸The online appendix provides examples of sentences and their classifications, which were given to the research assistants prior to them completing the classification task.

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.⁹ 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]

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 GPT-3.5 and GPT-4 models, which we prompt through the application programming interface (API).¹⁰ 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 and dictionary-based methods.¹¹ 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.

⁹Overall, the work was distributed between four reviewers.

¹⁰We use the TEXT-DAVINCI-003 and GPT-4 versions, respectively.

¹¹Specifically, we use the SENTENCE-TRANSFORMERS/PARAPHRASE-MPNET-BASE-V2 model in combination with a zero-shot classification pipeline.

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.¹²

3.3 Results

Figure 1 displays the distribution of labels across the classification methods. Overall, GPT-4 is the best performing model relative to the human benchmark. GPT-3.5 closely matches the human benchmark for the “dovish” and “mostly hawkish” labels. BERT overestimates the number of “dovish” sentences and dictionary-based methods rarely label sentences as “dovish” or “mostly dovish”. Both the GPT-3.5 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:

[Insert Table 3 about here]

Overall, the GPT-4 model obtains the lowest numerical errors, the highest accuracy, and the highest measure of agreement. The MAE and RMSE are 0.31 and 0.48, which

¹²To 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” if $s_i \in (-1, -0.5)$, “mostly dovish” if $s_i \in (-0.5, 0)$, “neutral” if $s_i = 0$, “mostly hawkish” if $s_i \in (0, 0.5)$, and “hawkish” if $s_i \in (0.5, 1)$.

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-4 and GPT-3.5 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 models and BERT) outperform dictionary-based methods.

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-3.5 and GPT-4 models, and a human research assistant, Bryson, to classify and provide explanations for their classifications for a few selected sentences.¹³

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 categorized as “neutral” by GPT-3.5 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-3.5 model does argue in a way that is consistent with its assigned label.

The GPT-4 model offers an improvement over GPT-3.5 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:

¹³The prompt and detailed results from this exercise are provided in the online appendix.

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”*

In sum, this exercise shows that 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-3.5 with more cases of agreement with Bryson.

4 Automating the Narrative Approach

The narrative approach, pioneered by Friedman and Schwartz (1963), and 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.¹⁴ 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.¹⁵ 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.

¹⁴As 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!).”*

¹⁵We thank Yuriy Gorodnichenko for this idea.

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.¹⁶

We construct a prompt that reflects the principles outlined in R&R (2023).¹⁷ 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.

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.

¹⁶Transcripts are not available for the full sample because the Fed releases transcripts of FOMC meetings with a five-year lag.

¹⁷The prompt is available 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. 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.¹⁸

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

¹⁸We thank David H. Romer for valuable discussion on this point.

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.

Finally, this question leads to a broader issue of human-AI interaction. In a recent work, Agarwal et al. (2023) conduct an experiment varying the information available to radiologists diagnosing patient cases, with or without AI assistance in order to compute the treatment effect of AI on diagnostic quality. They show that while there are considerable potential gains from optimal delegation, it is a non-trivial problem given the systematic biases in human decision-making. In the context of the narrative approach, the delegation problem would involve determining which tasks in the research process can be effectively and reliably delegated to AI and which should be performed by human researchers.

5 Did Investors Use ChatGPT to Decipher Fedspeak?

It took ChatGPT five days to reach 1 million users and just two months to reach 100 million users, shattering all previous records of tool adoption.¹⁹ It is difficult to find another tool, both in the finance space and elsewhere, that has been adopted so quickly and at such scale. The widespread and fast adoption was possible, in part, due to the fact that using the chatbot does not require any specific skills, as it is purely conversational.

The question is, whether market participants have been actively using ChatGPT to decipher Fed communications, and in that case, whether the technology has changed how markets react to Fed messages. We attempt to shed some preliminary light on these questions by comparing trading volumes around FOMC days before and after the release of ChatGPT. We present this analysis under the caveats that numerous confounding factors might be at play, and there has only been a small number of post-GPT FOMC meetings. We are therefore not claiming this exercise to be a detailed statistical analysis or to establish causal linkages.

It's been shown that both the trading volume and market depth increase during the

¹⁹For comparison, it took Instagram more than two years to garner 100 million users.

FOMC announcement days, and, in particular, during minutes surrounding the statement release (Fleming and Piazzesi, 2005) and the press conference (Gómez-Cram and Grotteria, 2022). Additionally, Rosa (2013) finds that the release of FOMC minutes significantly affects the volatility of U.S. asset prices and trading volumes, with the magnitude of these effects being both economically and statistically significant. Kim and Verrecchia (1991) and Shalen (1993) posit that new information can impact trading volume by influencing disagreement among market participants. Increased disagreement may lead to higher trading volume, while reduced disagreement, reflecting converging beliefs, could result in lower volume.

We investigate whether there has been a change in the behaviour of market participants by comparing the trading volume dynamics in the period before and after ChatGPT roll-out, as well as looking at control weeks, following Rosa (2013). Figure 2 plots this comparison:

[Insert Figure 2 about here]

Figure 2 displays the median change in trading volume around the FOMC announcement release (before and after GPT-4 introduction) and on control days (the same weekdays and hours of the previous and following weeks of the FOMC announcement release day).²⁰ Fluctuations in trading volume appear relatively stable and low before the announcement, with some spikes occurring at and after the release, indicating increased trading activity during these times. The graph also suggests that the market's reaction to the FOMC announcements has intensified in the period after GPT-4's introduction. The comparison suggests there is a difference in how the market is reacting to information releases now compared to before.

The marked increase in trading volume changes could mean that the market participants are reacting more strongly to the information provided by the FOMC after the introduction of GPT-4. This might be due to faster dissemination of information, leading to quicker and possibly more significant trading decisions. It may also imply that GPT-4 has enabled better or quicker analysis of the FOMC announcements, allowing traders who use this technology to act more decisively.

²⁰These series appear to statistically differ based on ANOVA analysis.

6 Conclusions

The analysis presented in this paper shows that GPT models demonstrate a strong performance in interpreting Fed communication. First, the GPT-4 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. This exercise establishes a benchmark for the performance of GPT models in the realm of Federal Reserve communication. Second, the GPT-4 model is mostly successful in identifying monetary policy shocks from FOMC transcripts and minutes as in R&R (1989, 2023). Whereas this is a highly cumbersome and time-consuming task for humans, the model generates results almost instantly. GPT models can therefore be an invaluable tool for help implementing the narrative approach for future FOMC meetings to identify new monetary policy shocks.

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

- AGARWAL, N., A. MOEHRING, P. RAJPURKAR, AND T. SALZ (2023): “Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology,” *National Bureau of Economic Research*.
- 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. EHRLMANN, 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*.
- COOK, T. R., S. KAZINNIK, A. L. HANSEN, AND P. MCADAM (2023): “Evaluating Local Language Models: An Application to Financial Earnings Calls,” *Available at SSRN 4627143*.
- 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)*.

- 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.
- FLEMING, M. J. AND M. PIAZZESI (2005): “Monetary policy tick-by-tick,” *Working Paper*, August, 3, 9.
- 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.
- GÓMEZ-CRAM, R. AND M. GROTERIA (2022): “Real-time price discovery via verbal communication: Method and application to Fedspeak,” *Journal of Financial Economics*, 143, 993–1025.
- 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.

- HORTON, J. J. (2023): “Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?” *arXiv preprint arXiv:2301.07543v1*.
- 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.
- KIM, O. AND R. E. VERRECCHIA (1991): “Trading volume and price reactions to public announcements,” *Journal of Accounting Research*, 29, 302–321.
- 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.
- ROSA, C. (2013): “The financial market effect of FOMC minutes,” *Economic Policy Review*, 19.

- SHALEN, C. T. (1993): “Volume, volatility, and the dispersion of beliefs,” *The Review of Financial Studies*, 6, 405–434.
- 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

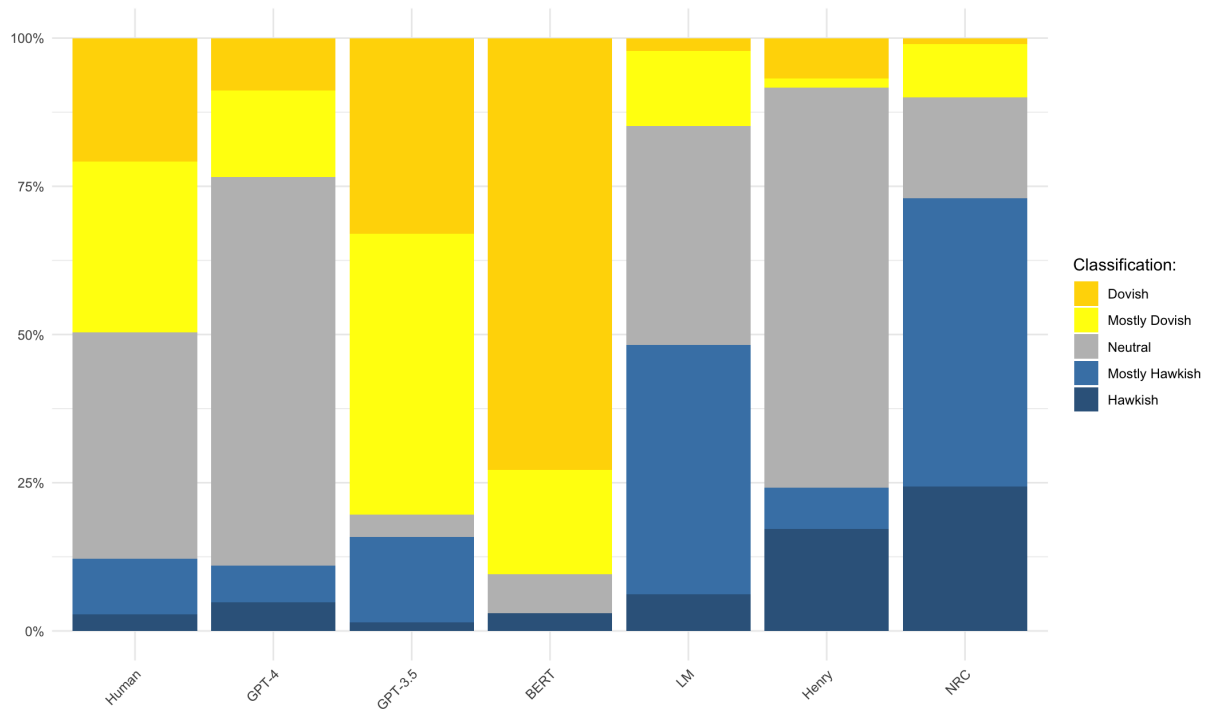


Figure 1: Distribution of labeled categories by method

Note: This figure shows the distribution of classification results across the different methods used in the paper. The stacked bar chart presents the percentage distribution of policy stance classifications.



Figure 2: Volume Dynamics around FOMC Announcement Release

Note: These two figures display the dynamics of the median change in SPY trading volume for FOMC days prior to and after the release of GPT-4 against two control series.

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-4	GPT-3	BERT	LM	Henry	NRC
MAE	0.31	0.41	0.66	0.62	0.55	0.81
RMSE	0.48	0.58	0.84	0.80	0.75	0.96
Accuracy	0.52	0.37	0.25	0.28	0.35	0.11
Kappa	0.30	0.18	0.03	0.07	0.08	-0.04
F1 Score						
Dovish	0.42	0.49	0.31	0.07	0.17	0.04
Mostly dovish	0.38	0.43	0.33	0.23	0.04	0.17
Neutral	0.63	0.15	0.13	0.48	0.57	0.14
Mostly hawkish	0.44	0.36	NA	0.15	0.07	0.11
Hawkish	0.47	0.10	0.07	NA	0.08	0.03
Balanced Accuracy						
Dovish	0.63	0.71	0.48	0.51	0.53	0.51
Mostly dovish	0.60	0.56	0.56	0.53	0.50	0.51
Neutral	0.66	0.54	0.51	0.59	0.59	0.45
Mostly hawkish	0.67	0.67	0.50	0.49	0.50	0.42
Hawkish	0.81	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.