

Can ChatGPT Decipher FedSpeak?

Anne Lundgaard Hansen and Sophia Kazinnik*

August 5, 2023

Abstract

Abstract 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 announcements relative to human assessment. We show that GPT models deliver a considerable improvement in classification performance over other commonly used methods. We then demonstrate how the GPT-4 model can provide explanations for its classifications that are on par with human reasoning. Finally, we show that the GPT-4 model can be used to identify macroeconomic shocks using the narrative approach of Romer and Romer (1989, 2023).

Keywords: Natural Language Processing (NLP), Generative Pre-training 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, and Christoffer Jessen Weissert for valuable feedback. 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, and ChatGPT in particular, have received a tremendous amount of public attention in recent months for their ability to analyze and generate text.¹ Since its release in November 2022, ChatGPT (Radford et al., 2018), an artificial intelligence chatbot, has become a prominent topic of discussion across digital platforms and academic fields alike.²

In this paper, we empirically evaluate the ability of GPT models to decipher FedSpeak, the language used by the Federal Reserve to communicate monetary policy decisions. While much progress has been made in this regard, FedSpeak is still known to be notoriously difficult to understand and messages are often delivered in a convoluted manner.³ 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 note 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 from Transformers (BERT) (Devlin et al., 2018). Among the con-

¹While we mention and use both ChatGPT and GPT models in the paper, it is important to note that these terms are not interchangeable. Built by the same company (OpenAI), GPT-3 and GPT-4 are large language models trained on extensive internet data, while ChatGPT is a conversational agent that leverages GPT-3 and GPT-4 models for human interaction.

²See, e.g., Choi et al. (2023) (law), Frieder et al. (2023) (mathematics), and Biswas (2023) (public health).

³Farber, A. (2013, April 19). Historical Echoes: FedSpeak as a Second Language. Liberty Street Economics.

sidered NLP methods, zero-shot GPT-3 obtains the lowest numerical errors, the highest accuracy, and the highest measure of agreement relative to human classification.⁴ By fine-tuning the GPT-3 model, we achieve even higher accuracy.

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 few selected sentences provided by our research assistants (Bryson) with those generated by ChatGPT.⁵ We note that both GPT-3.5 and GPT-4 models generally present a logic that successfully justifies their classifications and is very similar to Bryson's reasoning. GPT-4 offers an improvement over GPT-3.5 with improved reasoning and more cases of agreement with Bryson.

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 macroeconomic shocks using the narrative approach of Romer and Romer (1989; 2023). The narrative approach involves careful reading of relevant texts to help infer and establish causal relationships. Romer and Romer (1989; 2023) use this approach to read texts of FOMC discussions and establish policymakers motivation in changing policy, in order to determine when macroeconomic shocks occurred.

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 classifying policy language in a variety of settings. Consequently, the era where computers can interpret Federal Reserve 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 are effectively redefining the constraints once established by manual analysis. What used to be a labor-intensive and

⁴Zero-shot learning is a technique that enables a pre-trained model to classify samples from classes that were not present in the training data.

⁵We use both the GPT-3.5 and GPT-4 versions that underlie the ChatGPT user interface.

error-prone process could now become automated, scalable, and efficient.

Looking beyond the 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. For example, GPT models can be used to automate the analysis of Fed communications, potentially leading to a reduction in investor disagreement and more accurate predictions. 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 even when they are listening. They propose a solution where communication is tailored to specific audiences. Our results showcasing the performance of GPT models in the realm of central bank communication suggest that this technology can provide a feasible way to implementing such solutions.

We proceed as follows. After a review of the literature in Section 2, the paper is organized in two parts. The first part, covered in Section 3, presents the exercise of classifying the policy stance of FOMC announcement texts. Specifically, the data and the manual labeling task is discussed in Section 3.1; we provide an overview of GPT models and other NLP algorithms used for comparison in Section 3.2; we present the classification results in Section 3.3; and we compare the explanations of classifications provided by GPT models with human reasoning in Section 3.4. The second part in Section 4 shows how GPT models can be used to identify macroeconomic shocks using the narrative approach. We detail the method in Section 4.1, present the results in Section 4.2, and provide discussion in Section 4.3. Finally, we conclude in Section 5.

2 Literature Review

The literature on application of GPT models in the realm of economics and finance is growing rapidly, with researchers exploring its various capabilities and applications. For example, Lopez-Lira and Tang (2023) evaluate the performance of ChatGPT in forecasting returns,

leveraging the model’s sentiment analysis capabilities and comparing it to existing sentiment analysis methods. Leippold (2023a) 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. 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 large language models (LLMs) in general can be utilized by financial researchers to increase productivity by automating micro-tasks. 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.⁶

We provide a first attempt at evaluating GPT models for the purpose of quantifying Fed communication. We contribute to a large literature that uses natural language processing to study 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 (2023), and many others) and the general public (e.g., Ehrmann and Wabitsch (2022),).

Whereas earlier contributions quantify central bank texts based on topic modeling and sentiment analysis using pre-defined dictionaries such as that developed by Loughran and McDonald (2011) (e.g., Chen (2016), Hansen and McMahon (2016), Jegadeesh and Wu (2017), Benchimol et al. (2020)), more recent papers use pre-trained large language models (LLMs), such as the BERT models (e.g., Doh et al. (2022), Bertsch et al. (2022), Gorodnichenko et al. (2023)). In this paper, we compare the performance of these popular methods with the performance of GPT models, using both zero-shot and few-shot learning, and establish that GPT models outperform these previously used methods. We provide a detailed description of these methods in the Section 3.2.

⁶Interestingly, 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 Classifying Policy Stance

This section presents the analysis of the ability of GPT models to classify the policy stance of FOMC announcement texts.

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 deliberations behind the monetary policy decisions the committee makes.⁷

Our main analysis focuses on these FOMC statements published between 2010 and 2020.⁸ We begin by preprocessing the data the following way. We first divide each FOMC statement into a set of individual sentences. These sentences are then manually annotated with respect to their policy stance. Each sentence is classified outside of the broader context of the statement, i.e., it is treated as a standalone unit of meaning. 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.

We proxy for the policy stance using five pre-defined categories: “dovish”, “mostly dovish”, “neutral”, “mostly hawkish”, and “hawkish”. We assign numerical values to the categories on a scale of -1 to 1, where 0 represents a neutral stance, to compute performance metrics, which are also defined in the table. We employ five categories instead of three (i.e., “dovish”, “neutral”, “hawkish”) to examine GPT’s ability to discern subtle differences between closely related labels, a common challenge in machine learning.

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

⁷FOMC statements are available from the website of the Federal Reserve Board of Governors: <https://www.federalreserve.gov/monetarypolicy/fomccalendars.htm>.

⁸We discuss the data used for the narrative approach exercise in Section 4.

⁹Overall, the work was distributed between four reviewers, all with backgrounds in finance and economics.

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. We provide the definition for each label in Table 1:

[Insert Table 1 about here]

The summary statistics of the manual classifications and the disagreement among reviewers 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 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”.

Next, we describe the natural language processing methods used to quantify the FOMC statements described in this section.

3.2 NLP Methods

3.2.1 GPT Models and ChatGPT

The GPT model series belongs to a family of pre-trained large language models (LLMs).¹⁰ Trained on large data sets of unlabelled text, they are capable of performing a wide range of downstream tasks.¹¹ GPT models use self-attention architecture to consider sentence and paragraph context, allowing them to excel in various natural language processing tasks and

¹⁰Simply put, a large language model is an artificial intelligence algorithm that uses deep learning and was trained on a massively large set of data.

¹¹With each iteration of the GPT models, the models have increased in size and complexity. GPT-1 had 117 million parameters. The second iteration, GPT-2, had 1.5 billion parameters. GPT-3 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.

capture nuanced language understanding (see, for example, Zhang et al. (2022)). In the paper we use both the GPT-3 and GPT-4 model suites, as well as ChatGPT, an interactive chatbot that utilizes both GPT versions.

GPT-3 was released in June 2020, and created a storm of interest with its remarkable language generation capabilities.¹² On March 15, 2022, OpenAI released the new version of GPT-3 called “TEXT-DAVINCI-003”. This model was described as more capable than previous versions.¹³ GPT-4 is currently the most advanced version of the OpenAI language models. Released on March 14, 2023, it is said to be the new milestone in deep learning development. Overall, GPT-4 generates more factually accurate statements than GPT-3 and GPT-3.5, ensuring greater reliability and trustworthiness.

Models like GPT-3 and GPT-4 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 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 is a considerable advantage as fine-tuning takes time and computational resources. We thus test its zero-shot capabilities in evaluating FOMC announcements. For comparison, we also evaluate the performance of fine-tuned GPT on a subset of our data.¹⁴

¹²Both GPT-3 and GPT-4 is a suite of multiple base models with varying numbers of parameters. The most popular ones are ADA, BABBAGE, CURIE, and DAVINCI.

¹³Eight months later, 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.

¹⁴In the supplementary material, we describe an additional analysis using GPT-3 with embeddings.

3.2.2 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 (see, for example, 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.¹⁵

3.2.3 Dictionary-Based Methods

Dictionary-based methods are typically based on order-less document representation, where only the counts of words matters. The counts is usually based on pre-defined lexicons or term lists, containing labeled words or phrases. These methods are popular for their simplicity and transparency. However, their effectiveness is limited by their coverage, and they struggle with nuances and broader context of the language.¹⁶ Therefore, dictionary-based analyses often use multiple dictionaries to cross-validate the findings. For our analysis, we use the following three dictionaries:

- Loughran and McDonald (LM, 2011) 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 (see, for example, Picault and Renault (2017) and Shapiro and Wilson (2022)). There are 2,356 negative words and 354 positive words in this lexicon.
- 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. There are 85 negative words and 105 positive words in this lexicon.

¹⁵We use the SENTENCE-TRANSFORMERS/PARAPHRASE-MPNET-BASE-V2 model in our analysis.

¹⁶Take the term “transitory”, for example. This term might be classified as positive, neutral, or not classified at all, depending on the dictionary in question. However, in the realm of Fed communication, the term has been used to describe high inflation rates as temporary suggesting a positive sentiment.

- NRC Word-Emotion Association Lexicon of Mohammad and Turney (2015): NRC is a linguistic resource created by the National Research Council of Canada to capture the sentiment of everyday language. There are 6,468 words labeled as positive and 4,783 words labeled as negative in this lexicon.

In general, these dictionaries are popular because they offer a simple and transparent method for sentiment analysis. They typically come with predefined lists of words and corresponding sentiment labels, which makes them easy to use. However, their effectiveness can be limited by their coverage and their ability to handle nuances and context.

3.3 Results

We begin by evaluating how well GPT-3 classifies the FOMC announcements using zero-shot learning based on our full sample of 500 labeled sentences.¹⁷ We then report results from a case study that explores the reasoning capabilities of GPT models.¹⁸

3.3.1 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 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.

[Insert Figure 1 about here]

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

¹⁷The base model used in this analysis is TEXT-DAVINCI-003.

¹⁸We also consider the performance of GPT-3 on a subset of 100 sentences after fine-tuning the model with the remaining 400 sentences. These results are reported in the Appendix.

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;¹⁹ (iii) F1 score, the harmonic mean of true positive prediction proportions, precision and recall; and (iv) balanced accuracy, the average of true positive predictions among all positive cases and true negative predictions among all negative cases.

[Insert Table 3 about here]

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

3.3.2 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 model, leaving 100 sentences for cross validation. Table 4 reports the performance metrics of all methods on this test sample.

[Insert Table 4 about here]

¹⁹Kappa values range from -1 (total disagreement) to 1 (perfect agreement), with 0 indicating agreement due to chance alone.

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 further: 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.

3.4 Case Study of Explanations: Bryson vs. ChatGPT

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. This section tests this capability in a case study of selected sentences.

We ask both ChatGPT and a human research assistant, Bryson, to classify the sentences and provide explanations for their classifications. 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:

Explanation Prompt

Imagine you are a research assistant working for the Fed. You have a 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 contin-

ued 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:

[Insert Table 5 about here]

For the first sentence, Bryson and GPT-4 agree that the policy stance is dovish due to discussion of 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.

[Insert Table 6 about here]

For the second sentence, both algorithms agree with Bryson on 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”).

[Insert Table 7 about here]

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

[Insert Table 8 about here]

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

[Insert Table 9 about here]

Finally, the fifth 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 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 with more cases of agreement with Bryson. For example, GPT-3 fails to capture the classification of the “dovish” and “hawkish” sentence.

4 The Narrative Approach with the GPT-4 Model

Given the ability of GPT models to provide reasoning and explanations that are on par with those of humans, we hypothesize that they may also have the capacity to accurately implement the narrative approach to macroeconomic identification.²⁰ The narrative approach, pioneered by Friedman and Schwartz (1963), and extended and formalized by Romer and Romer (1989), sets out to uncover monetary policy shocks from rigorous analysis of qualitative data, e.g., transcripts and minutes of FOMC meetings. The manual treatment of these texts is resource and time consuming.²¹ With this constraint in mind, we set out to reproduce their method with the help of the GPT-4 model.

4.1 Method

In their seminal paper, Romer and Romer (1989) argue that in order to capture the impact of monetary policy actions, one needs to first pinpoint the motivation for these actions to account for endogeneity and omitted variable bias. The authors further argue that rigorous

²⁰We thank Yuriy Gorodnichenko for this idea.

²¹As Romer and Romer (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!).*”

narrative analysis is key to capturing monetary policy shocks, and they provide a set of criteria that allow for reproduction.

Romer and Romer (2023) provide an update to this framework. They postulate that the effectiveness of a narrative analysis largely depends on the following few key factors. First, the source of the narrative should be reliable, meaning it should be contemporaneous, consistent over a long period of time, and accurate. Second, it is essential to have a clear idea of what to look for in the narrative source. Romer and Romer (2023) propose to look for periods when monetary policymakers altered money growth and interest rates in a way unrelated to current or future economic activity, hence termed as policy "shocks". Last, it is crucial to approach the narrative source dispassionately and consistently, and document the narrative evidence carefully.²²

We follow these guidelines in our setup. First, we first gather the relevant documents. Similar to Romer and Romer (1989) and Romer and Romer (2023), we use the detailed transcripts from 1946 to 2017, and shorter minutes for years 2017 through 2023, where transcripts are not available.²³ The minutes provide a timely and concise summary of the significant policy issues discussed during each meeting of the committee. The transcripts offer a detailed record of the FOMC meeting proceedings.²⁴

Next, we put together a prompt that reflects the principles summarized in Romer and Romer (2023). Since most shocks identified by 1989; 2023 are contractionary in nature, we focus only on these and leave the identification of expansionary shocks for future work.

Narrative Approach Prompt

As a monetary policy expert, your task is to determine whether a given text contains a monetary

²²These criteria are summarized in Table 1, Romer and Romer (2023).

²³Transcripts are not available for the full sample because the Federal Reserve releases transcripts of FOMC meetings with a five-year lag.

²⁴Initially, the transcripts were produced as a reference for the Secretariat when preparing meeting minutes and were not fully edited for accuracy. The most comprehensive version of the transcripts comes from 1994 onward, as the Secretariat began to produce them shortly after each meeting from an audio recording of the proceedings. These records are lightly edited for clarity, and meeting participants are given a chance to review the transcript for accuracy.

policy shock. A monetary policy shock refers to movements in monetary policy that are unrelated to current or prospective real economic activity. These shocks occur when policymakers change money growth and interest rates due to concerns about prevailing inflation levels, even when the economy is stable. Policymakers, in these instances, are willing to accept potential negative consequences for aggregate output and unemployment.

Analyzing the provided text, determine whether it meets the criteria for a monetary policy shock based on the following factors:

- The policymakers believed the economy was at potential output.
- Policymakers changed money growth and interest rates due to high inflation.
- Policymakers understood and accepted the potential adverse consequences for output and unemployment.

Consider the example given as a reference:

Example: December 1988 “This episode counts as a contractionary monetary policy shock because, at a stable level of growth and unemployment, policymakers decided that the current level of inflation was unacceptable and took actions to reduce it. And they clearly understood and accepted that there could be substantial adverse consequences for output and unemployment.”

Assess whether the provided text indicates a similar scenario. If it does, explain why it meets the criteria for a monetary policy shock. If it does not, provide a brief explanation of how it fails to meet the criteria.

This prompt is constructed in a way as to have the following features. First, it clearly outlines the criteria for identifying a monetary policy shock: policymakers 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. The prompt asks for a detailed explanation of why the provided text does or does not meet the criteria for a monetary policy shock. In sum, the prompt is constructed in a way that aims to align with the criteria for narrative analysis outlined in Romer and Romer (2023).

We assess each document in our sample one by one, in no particular order, so as to

follow the method of the original paper as closely as possible. To accommodate extensive documents and overcome existing token limits imposed by OpenAI, we take advantage of a PDF plugin designed for this purpose. This enables us to process larger volumes of content without compromising our approach. We employ the combination of the ChatGPT user interface and the advanced GPT-4 model, augmented with the specialized PDF file plugin. Consequently, the model generates responses based on the analyzed content. We query each document ten times to improve reliability of responses, as the model structure of ChatGPT is not deterministic (e.g., Reiss (2023)). For illustrative purposes and further understanding, we have included several examples of the aforementioned answers in the Appendix section of our paper.

4.2 Results

We first test whether the GPT-4 model can identify the same monetary policy shocks as the ones identified in Romer and Romer (1989) and Romer and Romer (2023) (R&R, hereafter). Then, we discuss additional shocks identified by GPT-4 model, and which were not present in R&R (1989) or R&R (2023).²⁵

4.2.1 Comparison with R&R Shocks

Table 10 lists the shocks identified by the R&R studies alongside with those identified by the GPT-4 model.

[Insert Table 10 about here]

Both R&R studies concurred 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. For instance, R&R (2023) identified additional shocks in September 1958, May 1981, December 1988, and June 2022, not found in R&R (1989). Conversely, the 1989 study found a shock in September 1955 not recognized in the R&R (2023) study.

²⁵We list all responses that capture a shock in a separate online appendix available at Hansen's website (direct link).

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 the 1955, 1974, and 1981 being classified as *maybe* being monetary policy shocks.²⁶

The shocks on which R&R (2023) and GPT-4 agree all share the following elements: discussions of policymakers 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 policymakers' 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 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 about 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. R&R (1989) concludes that there is a shock in September 1955. The model doesn't identify a shock during this month, but it does 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

²⁶This happens when the model finds *some* evidence for the shock, but not all R&R criteria are satisfied.

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 11:

[Insert Table 11 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. Each text is queried ten times to avoid potential issues with nondeterministic outcomes of the GPT-4 model.

4.3 Discussion

Does this result 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. With the help of LLMs, we are now more capable at processing text data.

However, it is important to reiterate that these tools are not infallible. First, it is important to keep in mind that GPT-4 assessments are a function of both the prompts and the examined documents. The prompt acts as guiding signals for the model, telling it what type of information or response to look for. If a prompt is designed to prioritize certain criteria or guide the model towards specific features, it can produce more nuanced responses. These nuanced responses can reflect deeper, more intricate understandings or aspects of the topic. However, this also means that if the prompt is too narrow or overly focused, it could potentially overlook broader or more general information that may also be important. Furthermore, the construction of the prompt might prioritize certain criteria or guide the model to seek particular features that might be more nuanced than instructed. Therefore, a non-match occurs when the model does not find any relevant or significant information in the examined documents corresponding to the prompt, as the level of nuance and detail in the documents might not necessarily be homogeneous over time. A thoughtful and well-constructed prompt, coupled with comprehensive and relevant documents, can lead to more accurate and insightful outputs from the model, and in this case should not be used for blind classification.

Second, there is a trade-off between using the application programming interface (API) – in which the randomness of the model can be controlled, but the token window is limited – and the ChatGPT interface – which makes the process more reproducible, and allows plugin access, but in which one cannot reliably set the randomness. GPT models are designed to make probabilistic rather than deterministic predictions based on input data. As such, model output can change in random directions and asking the same question twice, may result in different answers. It can also change its responses based on context, if the randomness setting is not adjusted to zero. The probabilistic nature of ChatGPT allows it to generate

diverse and nuanced responses to inputs, which can make conversations more engaging and interesting. However, this also means that the same input can generate different responses each time ChatGPT is used.

Can the narrative approach be delegated? R&R (2023) raise this question, answering yes, albeit carefully (“*at some point, in some cases, if it is done very careful*”). Given our findings, we argue that this answer could be modified to “*probably now, in a lot of cases, and still very carefully*”. The reason for this update is the following. Given the ability of LLMs to follow instructions and perform well with virtually no training (i.e., zero-shot or few-shot learning), the need for training is definitely reduced. But, qualified researchers are still needed to figure out what can be learned from the source, what type of information can be extracted, and in what shape. While we were able to put together a prompt that would extract 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 Concluding Remarks

The analysis presented in this paper shows that GPT models demonstrate a strong performance in classifying FOMC language, especially when fine-tuned. Furthermore, GPT models can explain its classifications using arguments similar to a human benchmark (Bryson). Based on the case study, the GPT-4 model 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.

We also show that qualitative analysis, traditionally a domain of human expertise, now has the potential to be enhanced and scaled up by AI. The implications of this 'renaissance' are profound, as it fundamentally transforms our understanding of what qualitative analysis (on AI) can achieve. It opens up possibilities for a new era of research and knowledge discovery 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. Thus, 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*.
- AKYÜREK, E., D. SCHUURMANS, J. ANDREAS, T. MA, AND D. ZHOU (2022): “What Learning Algorithm Is In-Context Learning? Investigations With Linear Models,” *arXiv preprint arXiv:2211.15661*.
- 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*.
- BISWAS, S. S. (2023): “Role of Chat GPT in Public Health,” *Annals of Biomedical Engineering*.
- 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*.
- CHOI, J. H., K. E. HICKMAN, A. MONAHAN, AND D. SCHWARCZ (2023): “ChatGPT Goes to Law School,” *Minnesota Legal Studies Research Paper*.
- 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.
- FRIEDER, S., L. PINCHETTI, R.-R. GRIFFITHS, T. SALVATORI, T. LUKASIEWICZ, P. C. PETERSEN, A. CHEVALIER, AND J. BERNER (2023): “Mathematical Capabilities of ChatGPT,” *arXiv:2301.13867v1*.
- 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*.
- KORINEK, A. (2023): “Language Models and Cognitive Automation for Economic Research,” Tech. rep., National Bureau of Economic Research.
- LEIPPOLD, M. (2023a): “Sentiment Spin: Attacking Financial Sentiment with GPT-3,” *SSRN Working Paper*.
- (2023b): “Thus Spoke GPT-3: Interviewing a Large-Language Model on Climate Finance,” *Finance Research Letters*, 53, 103617.
- 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,” .
- PICAULT, M. AND T. RENAULT (2017): “Words Are Not All Created Equal: A New Measure of ECB Communication,” *Journal of International Money and Finance*, 79, 136–156.

- 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.
- SHAPIRO, A. H. AND D. J. WILSON (2022): “Taking the Fed at its Word: A New Approach to Estimating Central Bank Objectives Using Text Analysis,” *The Review of Economic Studies*, 89, 2768–2805.
- ZAREMBA, A. AND E. DEMIR (2023): “ChatGPT: Unlocking the Future of NLP in Finance,” *Available at SSRN 4323643*.
- 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

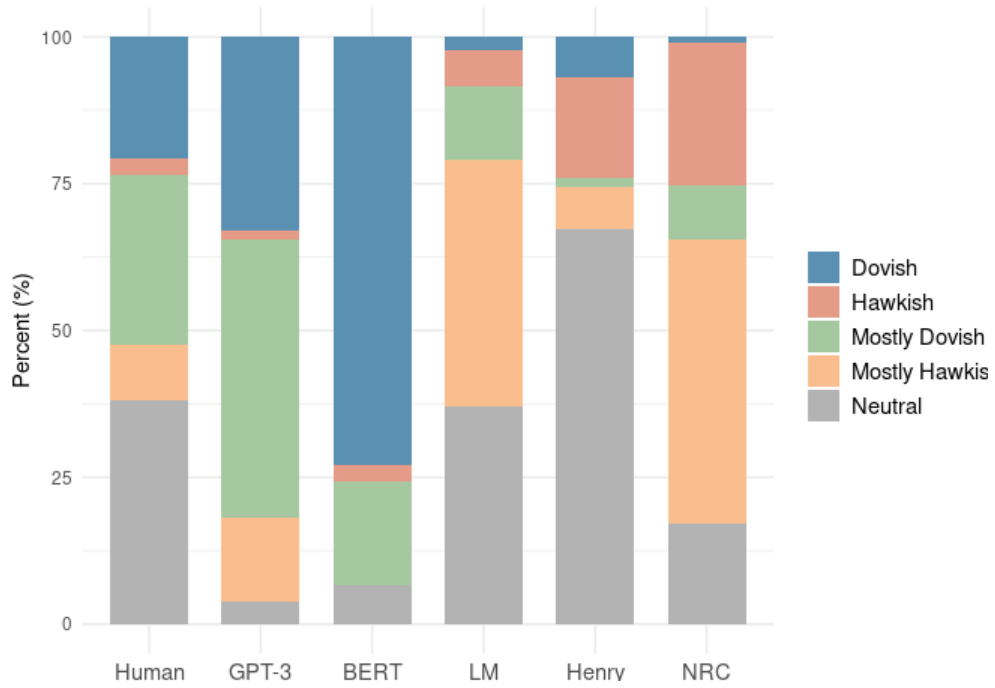


Figure 1: Distribution of Labels by Method

Tables

Table 1: Category Definitions

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	M. Dovish	Neutral	M. Hawkish	Hawkish
Count	500	104	144	191	47	14
Avg. Disagreement	0.47	0.67	0.52	0.31	0.51	0.67
N (>1 step)	264	104	60	67	19	14
N (>2 steps)	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.

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.

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.

Table 5: Explanations of classifications of **Sentence 1** as provided by the human research assistant (Bryson) and the GPT-3 and GPT-4 models.

Sentence 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.*

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.

Table 6: Explanations of classifications of **Sentence 2** as provided by the human research assistant (Bryson) and the GPT-3 and GPT-4 models.

Sentence 2: *Labor market conditions have shown some improvement in recent months, on balance, but the unemployment rate remains elevated.*

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.

Table 7: Explanations of classifications of **Sentence 3** as provided by the human research assistant (Bryson) and the GPT-3 and GPT-4 models.

Sentence 3: *The committee will closely monitor market conditions and is prepared to adjust its plans as appropriate.*

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.

Table 8: Explanations of classifications of **Sentence 4** as provided by the human research assistant (Bryson) and the GPT-3 and GPT-4 models.

Sentence 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.*

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.

Table 9: Explanations of classifications of **Sentence 5** as provided by the human research assistant (Bryson) and the GPT-3 and GPT-4 models.

Sentence 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.*

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.

Table 10: Contractionary monetary policy shocks from 1946–2023 as identified by Romer and Romer (R&R, 1989, 2023) and the GPT-4 model. For shocks denoted with *, the GPT-4 model concludes that there *maybe* was a policy shock.

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

Table 11: Contractionary shocks from 1946–2023 identified by the GPT-4 model only. Each column represent a shock; the rows list the associated FOMC meeting dates for which the GPT-4 model identified policy shocks.

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		

A 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.

B 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.²⁷ Second, we run a BERT-based SetFit model, a framework for few-shot fine-tuning of Sentence Transformers.²⁸

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.

²⁷We use TEXT-EMBEDDING-ADA-002 model to embed our data.

²⁸SetFit Model Reference

C Narrative Approach: Sample Responses

Below is an example response of a “no shock” classification:

Based on the information extracted from the document, it does not appear that the Federal Reserve’s March 16, 2010 meeting fits the criteria for a monetary policy shock. Here’s why:

1. The policymakers believed the economy was at potential output: The document does not provide clear evidence that policymakers believed the economy was at potential output. There are mentions of the economy being weak and unemployment being high, which suggests that they did not believe the economy was at its potential.
2. Policymakers changed money growth and interest rates due to high inflation: There is some discussion about the possibility of raising interest rates and concerns about inflation. However, it’s not clear that these changes were made due to high inflation. In fact, one excerpt suggests that inflation risks were seen as low at the time, and another suggests that inflation was likely to be subdued for some time.
3. Policymakers understood and accepted the potential adverse consequences for output and unemployment: While there is some discussion about the potential adverse consequences of policy actions, it’s not clear that these were accepted as a necessary trade-off for controlling inflation. Some policymakers expressed concerns about the high rate of unemployment and the weak state of the economy.

In conclusion, while there were discussions about potential changes to monetary policy and concerns about inflation, the document does not provide clear evidence that this was a case of a monetary policy shock as defined by your criteria.

Below is an example response of a “shock” classification:

Based on the information extracted from the document, it appears that the Federal Reserve’s meeting on March 26, 1957, does meet the criteria for a monetary policy shock. Here’s why:

1. The policymakers believed the economy was at potential output: The document mentions that there was a “slackened momentum of cyclical advance after some 30 months of sustained rising activity” (Page 12). This suggests that the policymakers believed the economy was operating at or near its potential output.
2. Policymakers changed money growth and interest rates due to high inflation: The document indicates that the policymakers were concerned about inflation. It states, “a policy course was set, directed to resisting inflationary pressures as they intensified” (Page 12). This shows that the policymakers were willing to change the money growth and interest rates to combat inflation.
3. Policymakers understood and accepted the potential adverse consequences for output and unemployment: The document suggests that the policymakers were aware of the potential negative consequences of their actions. For instance, one excerpt states, “Mr. 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” (Page 27). This indicates that they understood the potential adverse effects of their policies on output and unemployment and were willing to accept these consequences to control inflation.

Therefore, the scenario described in the document is similar to the example given in the summary. The policymakers believed the economy was at potential output, changed money growth and interest rates due to concerns about inflation, and accepted the potential negative consequences for output and unemployment.

Below is an example response of where the text does not meet all of the criteria of a shock, only some:

Based on the information extracted from the document, it does not appear that the meeting minutes from August 23, 1955, describe a scenario that meets all the criteria for a monetary policy shock. Here's why:

1. The policymakers believed the economy was at potential output. The document does not provide clear evidence that policymakers believed the economy was at potential output. While there are mentions of strong demand for industrial products and active real estate markets, there's no explicit statement about the economy operating at its potential.
2. Policymakers changed money growth and interest rates due to high inflation. The document does mention concerns about inflation. For instance, on page 8, it states, "We are faced with another period of increases in prices," and on page 7, it mentions the "wage cost push" and the "psychology that that creates." However, it does not clearly indicate that policymakers changed money growth and interest rates specifically due to these inflation concerns.
3. Policymakers understood and accepted the potential adverse consequences for output and unemployment. The document does not provide clear evidence that policymakers understood and accepted potential adverse consequences for output and unemployment. While there is a discussion about the potential for increased borrowing and the impact on banks, there's no explicit acknowledgment of potential negative effects on output and unemployment.

In conclusion, while the document does discuss concerns about inflation and the potential need for monetary restraint, it does not clearly meet all the criteria for a monetary policy shock as defined in your question. The document does not provide clear evidence that policymakers believed the economy was at potential output, changed money growth and interest rates due to high inflation, and understood and accepted the potential adverse consequences for output and unemployment.