Can ChatGPT Decipher Fedspeak?

Anne Lundgaard Hansen and Sophia Kazinnik

Federal Reserve Bank of Richmond

September 29, 2023 | The Fourth New York Fed Conference on Fintech: Artificial Intelligence and Digital Assets

The views expressed in this presentation are solely those of the authors, and do not necessarily reflect the views of the Federal Reserve Bank of Richmond or the Federal Reserve System.

The ChatGPT Revolution

- *Large language models (LLMs)* have received massive attention and are transforming the field of natural language processing (NLP)
  - Academic literature on their applications is growing FAST!

    (Lopez-Lira and Tang, 2023; Jha et al., 2023; Leippold, 2023a,b; Zaremba and Demir, 2023; Dowling and Lucey, 2023; Korinek, 2023)
- This is particularly true for OpenAI's chatbot, *ChatGPT* ⏩ Google Trends
  - Underlying ChatGPT are *Generative Pre-trained Transformer (GPT)* models

Model	Release	# parameters
GPT-1	June 2018	117 million
GPT-2	February 2019	1.5 billion
GPT-3	June 2020	$\sim$175 billion
GPT-4	March 2023	$\sim$ 1 trillion

  - ChatGPT uses GPT-3 (free version) or GPT-4 (subscription)

Do GPT Models Understand Fedspeak?

- GPT models are trained on large sets of data that are not specific to finance and economics
  - How well do they perform on financial texts?
  - … such as the technical and convoluted language used in Federal Reserve communication – *Fedspeak*?
- We explore the ability of OpenAI's GPT models to interpret Fedspeak through two exercises:

  (1) Classification of the policy stance of FOMC announcements relative to a human benchmark

  + their ability to explain the reasoning behind the chosen classifications relative to our research assistant, Bryson

  (2) Automating the narrative approach of Romer and Romer (1989) to identify monetary policy shocks

CLASSIFICATION EXERCISE

Data and Method

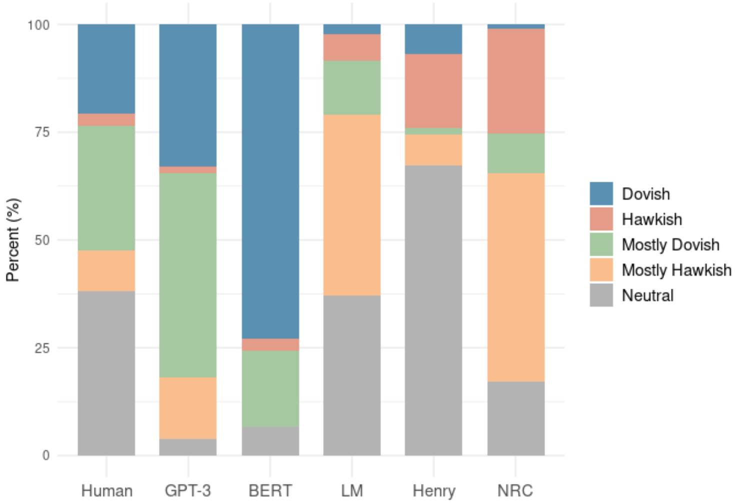
- Our analysis focuses on FOMC statements published between 2010 and 2020
  - FOMC statements are released publicly on FOMC meeting days, 8 times a year
  - They summarize the FOMC's view of the economy and policy deliberations
  - We break down these statements into sets of individual sentences
- Manual annotation:
  - Each sentence is manually classified with respect to their policy stance: "dovish", "mostly dovish", "neutral", "mostly hawkish", "hawkish" ⏩ Definitions
  - To avoid human bias, each sentence is independently reviewed by 3 humans
  - Since manual annotation is costly, we focus on 500 randomly selected sentences
  - ⏩ Summary statstics
- Model-implied annotation:
  - We use the GPT-3 model as is (zero-shot) and with fine-tuning (400/100 training/test data split)
  - We benchmark the performance against state-of-the-art and traditional NLP methods

Benchmark NLP Methods

- BERT (Bidirectional Encoder Representations from Transformers)
  - Introduced by Devlin et al. (2018)
  - Similar to GPT models, BERT is a pre-trained large language model based on the transformer architecture
  - Considered state-of-the-art and therefore widely applied (see, for example, Bertsch et al. (2022) and Huang and Hui Wang (2022) for applications in finance)
- Dictionary-based methods:
  - Dictionary-based methods use pre-defined lexicons containing labeled words or phrases
  - Popular for their simplicity and transparency
  - Performance is limited by their coverage and they struggle with nuances and broader context of the language

Dictionary	# words
Loughran and McDonald (LM, 2011) dictionary	2700
Henry (2008) financial dictionary	190
NRC Word-Emotion Association Lexicon of Mohammad and Turney (2015)	11,251

Zero-Shot Classifications



Zero-Shot Performance Metrics

	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

Note: For each metric, the best performing model is boldfaced

⏩ Performance metrics   ⏩ Fine-tuned learning

Exploring the Reasoning Capabilities of GPT Models

- Mere classification aside, GPT models have the ability to explain why a certain sentence was labeled in a certain way
  - This capability is beyond any existing NLP model
- We test this capability in a case study:
  - We ask both ChatGPT and a human research assistant, Bryson, to classify the sentences and provide explanations for their classifications
  - ChatGPT vs Bryson:

    ChatGPT: We use both GPT-3 and GPT-4 models (zero-shot) ⏩ Prompt

    Bryson: 24-year-old Federal Reserve research assistant, holds a BSc
  - Focus on a few selected sentences (one from each classification)

Summary of Results

- GPT models generally present a logic that successfully justifies their classifications
- The explanations are very similar to those provided by Bryson, especially for GPT-4
- 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
- GPT models not only outperform existing NLP methods, but offer a reasoning capability and logic that existing methods are not capable at providing
  - Potential for a wider array of applications than previous seen in the NLP literature

⏩ dovish sentence ⏩ mostly dovish sentence ⏩ Neutral sentence

⏩ Mostly hawkish sentence ⏩ Hawkish sentence

THE NARRATIVE APPROACH

Narrative Approach in Monetary Economics

- Narrative approach (Friedman and Schwartz, 1963): Use qualitative sources to identify policy changes that were not motivated by factors affecting output
  - Requires identifying policy makers' motivation behind policy decisions
- Romer and Romer (R&R, 1989): Implemented by manually reading through FOMC meeting material to identify monetary policy shocks
  - Reading through huge amount of text $\sim$ 50-100 pages per meeting, 8-12 meetings per year, since 1946(!)
- R&R (2023) provide an update to and revision of the 1989 paper
  - *"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."*
- Given the encouraging results of GPT models for interpreting FOMC statements, we hypothesize that they may also have the capacity to accurately implement the narrative approach

Implementing the Narrative Approach Using GPT

- Goal: To reproduce R&R (1989, 2023) by letting GPT models read and interpret meeting materials
- Method:
  - Use the detailed transcripts from 1946–2017 and shorter minutes from 2017–2023, where transcripts are not available
  - Construct a prompt based on R&R's instructions and query the above documents
    ⏩ Prompt
  - All documents are re-queried 10 times to avoid spurious results
  - Use the GPT-4 model with zero-shot learning

Results: Can ChatGPT Identify the R&R Policy Shocks?

- GPT-4 identifies most of the shocks discovered by R&R
- Three of the shocks (marked by $*$) are classified as *maybe* by GPT-4
- For the meetings where GPT-4 does not identify a shock but R&R do, we asked the model to explain why it didn't identify a shock
  - Absence of critical components in the definition of shocks; in most cases, missing evidence for the belief that the economy was operating at full potential

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

⏩ Example: June 2022

Results: Does ChatGPT Identify Additional Policy Shocks?

- GPT-4 identifies six new shocks
  - As per R&R, consecutive dates are not interpreted as multiple shocks, but one shock dated at the earliest identified point in time
- Discussions with David Romer:
  - *"In our identification of shocks, we try to distinguish the Fed attempting to prevent inflation from rising (which we wouldn't count as a shock) and it attempting to reduce inflation (which we would). ChatGPT doesn't appear to have picked up on that."*
  - ChatGPT points to key quotes using terms like "combat" and "control" inflation, rather than "reduce" or "lower"

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		

Discussion

- Are experts and researchers now obsolete? Absolutely not!
- The assessments of GPT models are a function of prompts and data
    - Thoughtful and well-constructed prompts are key to insightful model output
    - We relied heavily on the work of Romer & Romer to construct the prompt
- Qualified researchers are still needed to figure out what can be learned from the data, what type of data can be extracted, and in what shape

CONCLUSION

Concluding Remarks

- GPT models demonstrate huge potential:
  - The GPT-3 model demonstrates strong performance in classifying Fedspeak sentences, especially when fine-tuned
  - GPT-4 offers improvement over GPT-3 with more cases of agreement with human classifications and explanations
  - GPT-4 is able to identify monetary policy shocks with essentially zero training (zero shot)
- Despite impressive performance, GPT models are not infallible and can't fully replace human evaluators
- Future work: Open-source local LLMs
  - GPT models are criticized for lack of privacy, transparency, and reproducibility
  - Local LLMs facilitate many interesting and important applications involving data sources are confidential in nature

References

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

Devlin, J., Chang, M.W., Lee, K., Toutanova, K.. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:181004805 2018;.

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

Friedman, M., Schwartz, A.J.. A Monetary History of the US 1867-1960. Princeton University Press, 1963.

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

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

Jha, M., Qian, J., Weber, M., Yang, B.. ChatGPT and Corporate Policies. Chicago Booth Research Paper 2023;23-15.

Korinek, A.. Language Models and Cognitive Automation for Economic Research. Technical Report; National Bureau of Economic Research; 2023.

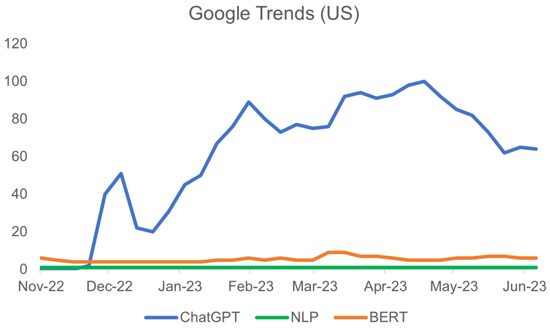
Leippold, M.. Sentiment Spin: Attacking Financial Sentiment with GPT-3. SSRN Working Paper 2023a;.

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

Lopez-Lira and Tang, . Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. SSRN Working Paper 2023;.

Loughran, T., McDonald, B.. When Is a Liability Not a Liability? Textual Analysis, Dictionaries,

Google Trends



⏩ Go Back

Policy Stance Categories

Classification	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.

⏩ Go Back

Summary Statistics

- The mode classification is “neutral”
- The sample is skewed towards “dovish” sentences
- Average disagreement is symmetrically U-shaped with less disagreement about “neutral” sentences and more about “dovish” and “hawkish” sentences

	Total	Dovish	Mostly Dovish	Neutral	Mostly 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. N (>1 step) and N (>2 steps) are the number of sentences for which maximal disagreement exceed 1 and 2 steps.

⏩ Go Back

Performance Measures

- Numeric error metrics: MAE and RMSE
- Accuracy – the proportion of correct predictions
- Balanced Accuracy – the average of true positive and true negative rates
  - Useful metric when classes are imbalanced
- Kappa – the agreement between predictions and actual values
  - Value ranges from -1 (total disagreement) to 1 (perfect agreement), with 0 indicating agreement due to chance alone
- F1 Score – accuracy computed by the harmonic mean of precision and recall
  - Precision: True positive predictions to all positive predictions
  - Recall: True positive predictions to the sum of true positive and false negative predictions
  - Value ranges from 0 (precision or recall are zero) to 1 (perfect precision and recall)

⏩ Go Back

Fine-Tuned Performance Metrics

	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

Note: For each metric, the best performing model is boldfaced. Performance is tested on test sample of 100 sentences only.

Go Back

Reasoning Exercise 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 continued to strengthen but decided, for the time being, to wait for some further evidence of continued progress toward its objectives.

5. The committee expects to begin implementing its balance sheet normalization program relatively soon, provided that the economy evolves broadly as anticipated; this program is described in the June 2017 addendum to the committee's policy normalization principles and plans.

⏩ Go Back

Bryson vs ChatGPT: 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.

⏩ Go Back

Bryson vs ChatGPT: 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.

⏩ Go Back

Bryson vs ChatGPT: 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.

⏩ Go Back

Bryson vs ChatGPT: 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.

⏩ Go Back

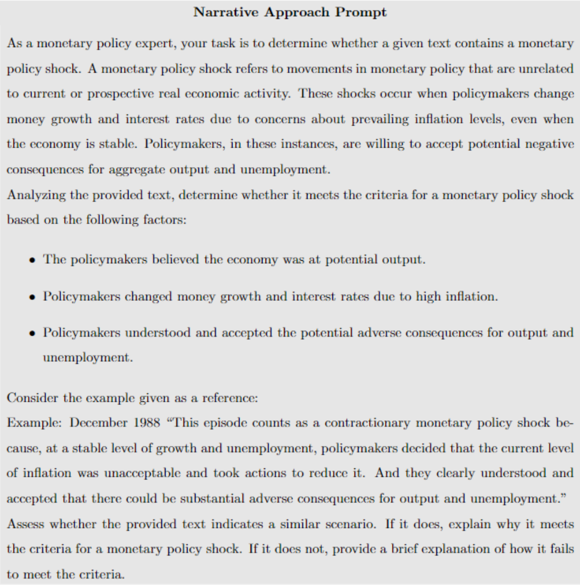
Bryson vs ChatGPT: Sentence 5

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

⏩ Go Back

R&R Prompt



Narrative Approach Prompt

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

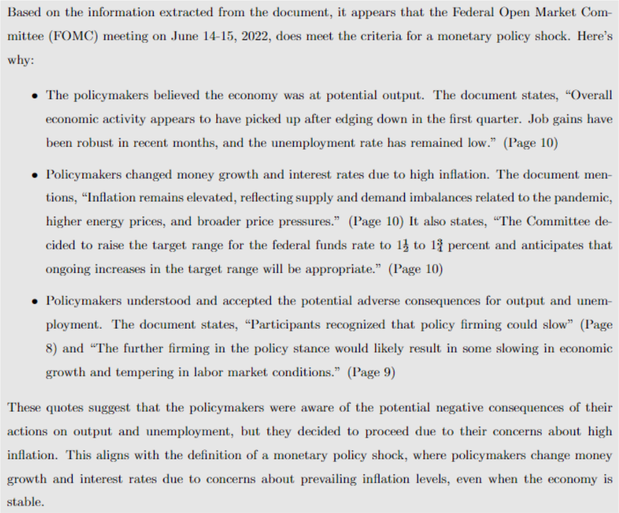
Consider the example given as a reference:

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.

⏩ Go Back

Example: An Identified Shock (June 2022)



Based on the information extracted from the document, it appears that the Federal Open Market Committee (FOMC) meeting on June 14-15, 2022, does meet the criteria for a monetary policy shock. Here's why:

- The policymakers believed the economy was at potential output. The document states, "Overall economic activity appears to have picked up after edging down in the first quarter. Job gains have been robust in recent months, and the unemployment rate has remained low." (Page 10)
- Policymakers changed money growth and interest rates due to high inflation. The document mentions, "Inflation remains elevated, reflecting supply and demand imbalances related to the pandemic, higher energy prices, and broader price pressures." (Page 10) It also states, "The Committee decided to raise the target range for the federal funds rate to $1\frac{1}{2}$ to $1\frac{3}{4}$ percent and anticipates that ongoing increases in the target range will be appropriate." (Page 10)
- Policymakers understood and accepted the potential adverse consequences for output and unemployment. The document states, "Participants recognized that policy firming could slow" (Page 8) and "The further firming in the policy stance would likely result in some slowing in economic growth and tempering in labor market conditions." (Page 9)

These quotes suggest that the policymakers were aware of the potential negative consequences of their actions on output and unemployment, but they decided to proceed due to their concerns about high inflation. This aligns with the definition of a monetary policy shock, where policymakers change money growth and interest rates due to concerns about prevailing inflation levels, even when the economy is stable.

⏩ Go Back