RedNIE
Red Nacional de Investigadores en Economía

Talk to Fed: a Big Dive into FOMC Transcripts

Daniel Aromí (IIEP UBA-Conicet/FCE UBA)

Daniel Heymann (IIEP UBA-Conicet/FCE UBA)

DOCUMENTO DE TRABAJO N° 323

Mayo de 2024

Citar como:
Aromí, Daniel, Daniel Heymann (2024). Talk to Fed: a Big Dive into FOMC Transcripts. Documento de trabajo RedNIE N°323.

Talk to Fed: A Big Dive into FOMC Transcripts

Daniel Aromí
IIEP UBA-Conicet, FCE UCA

Daniel Heymann
IIEP UBA-Conicet, UdeSA

April 2024

Abstract:

We propose a method to generate "synthetic surveys" that shed light on policymakers' perceptions and narratives. This exercise is implemented using 80 time-stamped Large Language Models (LLMs) fine-tuned with FOMC meetings' transcripts. Given a text input, fine-tuned models identify highly likely responses for the corresponding FOMC meeting. We evaluate this tool in three different tasks: sentiment analysis, evaluation of transparency in Central Bank communication and characterization of policymaking narratives. Our analysis covers the housing bubble and the subsequent Great Recession (2003-2012). For the first task, LLMs are prompted to generate phrases that describe economic conditions. The resulting output is verified to transmit policymakers' information regarding macroeconomic and financial dynamics. To analyze transparency, we compare the content of each FOMC minutes to content generated synthetically through the corresponding fine-tuned LLM. The evaluation suggests the tone of each meeting is transmitted adequately by the corresponding minutes. In the third task, we show LLMs produce insightful depictions of evolving policymaking narratives. This analysis reveals relevant narratives' features such as goals, perceived threats, identified macroeconomic drivers, categorizations of the state of the economy and manifestations of emotional states.

Keywords: monetary policy, large language models, narratives, transparency.

1. Introduction

Understanding macroeconomic events requires assessing the perceptions and ideas of economic actors at the time decisions are made and aggregate dynamics unfold (Lorenzoni 2009, Angeletos & Jennifer 2013, Heymann & Sanguinetti 1998). One type of key economic actor for which this exercise is particularly relevant corresponds to monetary policymakers. The interest in policymakers' views is reflected in previous literature that has focused on the information policymakers possess (Romer & Romer 2008, Hansen 2019), their preferences (Shapiro & Wilson 2022, Malmendier &Nagel 2021) and the way in which they communicate their views and activities (Woodford 2005, Gorodnichenko et al. 2023).

One way in which this agenda has been advanced involves processing the text in statements, speeches and related documents generated by policymakers. This type of unstructured data has been analyzed using dictionary methods (Apel et al 2022, Shapiro & Wilson 2022, Cieslak et al. 2023), topic models (Lucca & Trebbi 2009, Acosta 2023) or text classification models (Gorodnichenko et al. 2023). Text is processed to learn about features such as sentiment, policy stance and frequency of topics in the respective documents. While fruitful, these methods are relatively rigid in terms of its capacity to interpret high dimensionality data and to use that knowledge to provide useful indicators. Hence, there are gains to be made from exploiting more expressive models that possess a greater capacity to acquire and transmit ideas expressed in deliberations.

In this work, we propose a novel method that exploits techniques from natural language processing. Our presumption is that the expressiveness of LLMs can allow for more exhaustive and specific information extraction. Ther proposed method has three stages: LLM fine-tuning, text generation and text processing. First, we fine-tune LLMs so that the models learn the patterns of language used during monetary policy deliberations. In other words, in this first stage, we produce a latent representation of perceptions and ideas expressed during each meeting. In the second stage, we use selected trigger phrases to prompt responses that, as a result of the fine-tuning process, each time-stamped model judges highly likely. For example, to learn about how the state of the economy was perceived by policy makers during FOMC meetings, we can ask fine-tuned LLMs to complete phrases such as "Currently, economic conditions are…". In the third stage, completing the information extraction task, we process the LLMs' generated texts to compute a numeric representation that summarizes these responses in a compact manner.

We test the proposed methodology implementing three different tasks: sentiment analysis, evaluation of transparency in public communications and characterization of policymaking narratives. For the first task, LLMs are prompted to generate phrases that discuss economic conditions. Selected trigger phrases include: "Currently, economic conditions are" and "In the next months, the financial environment is expected to". The evaluation of the resulting indices suggests that fine-tuned LLMs are able to express policymakers' perceptions of economic conditions. Estimated forecast models show policymakers possess substantial information regarding macroeconomic and financial dynamics. For example, a one standard deviation increment in the index of perceived economic conditions is associated with an average 0.3% upward revision in yearly growth forecast revisions over the next two quarters.

The topic of Central Bank communication transparency has consistently garnered significant attention (Woodford 2005, Fischer et al. 2023, Stein 1989, Acosta 2023). To evaluate transparency, in the second task, we compare the content of the minutes corresponding to each FOMC meeting to the content generated synthetically through the corresponding fine-tuned LLM. More specifically, we prompt time-stamped LLMs to complete the beginning of each sentence of the original corresponding minutes. The evaluation suggests the tone of each meeting is transmitted adequately by the corresponding minutes. In particular, we do not observe that the tone communicated by the minutes during the most acute period of the housing and financial crisis was more positive than the tone generated by LLMs that were trained "listening" to deliberations during the respective meeting.

In the third task, we focus on policymaking narratives. Following Shiller (2019), narratives are viewed as relevant frameworks that determine how economic settings are interpreted and guide decision-making (Shiller 2019). In this task, LLMs are requested to generate a diverse set of manifestations that characterize policymakers' narratives. Starting with a broad perspective, we prompt time-stamped LLMs to communicate policymaking objectives and worries. Then, turning to more specific assessments, we collect statements about main drivers of economic growth and inflation. Complementary, we extract pseudo real-time classifications of economic conditions in terms of traditional labels such as "recession" and "crisis". Finally, we characterize narratives' disposition by asking LLMs about emotions manifested during policymaking discussions.

In terms of stated goals, we verify a balanced focus on two traditional policy objectives: price stability and economic growth. More specifically, when prompted to mention goals, fine-tuned LLMs' refer to price stability about 55% of the time. This frequency is slightly higher than the frequency with which LLMs refer to economic growth. In contrast to these balanced manifestations, when it comes to threats, all along our sample period, economic activity constitutes the single major concern. Beyond this main preoccupation, we uncover a sequence of important but less prominent worries that were manifested in different times of our sample period. This sequence of concerns starts with inflation in 2005, continues with housing in 2006, is followed by a persistent concern about the financial system and, during the last sample years, turns to the fiscal deficit.

Turning to more specific enquiries, we prompt LLMs to discuss main drivers of economic growth and inflation. We find that policymakers show stable views regarding the main forces driving economic activity. The most frequently mentioned force is aggregate demand with a frequency of 45%. Demonstrating a notable asymmetry, this figure triples the frequency with which aggregate supply is mentioned. In the same line, consumption is the second most mentioned driver with a frequency of 34%. Other frequently mentioned drivers are economic policy (30%) and financial markets (18%).

In the case of inflation, oil price is the most frequently mentioned driver with an average frequency of 31%. This high average value is a consequence of multiple spikes that arise from relatively low initial levels. According to LLM-generated responses, the second most prominent driver of inflation is given by economic activity with a frequency of 25%. Beyond these two forces, we find a diverse set of drivers with a frequency of approximately 10%. This group includes aggregate demand, expectations, past inflation, the output gap, exchange rates and wages. On the other hand, in the case of inflation, we do not find economic policy as a prominent driver. Monetary policy is mentioned only 7% of the time and the fiscal deficit is rarely signaled (1%).

We also evaluate the use labels in monetary policy narratives. We identify how business cycle conditions are categorized at different points in time. Regarding the occurrence of a recession, we verify that LLMs' responses allow for accurate real-time assessments of business cycle peaks and troughs. Complementarily, LLMs generated content shows how the Big Recession was characterized as a crisis very early on and how this event led to a persistent perception of fragility. In the final evaluation, we assess if policymaking narratives are charged with emotional content. According to generated responses, the Big Recession led to narratives charged with an increased perception of sadness and fear. In the case of sadness, we observe that the increment is not reversed in the subsequent years.

Our contribution is related to several strands of the literature. From an ample perspective, our focus on economic actors' perceptions and narratives is motivated by the understanding of the limitations of full information rational expectations models (Lorenzoni 2009, Angeletos and La'O 2013, Heymann, D., & Pascuini 2021). Additionally, our analysis is motivated by the interest in narratives as features of the economy that frame economic evaluations and behavior (Shiller 2019). Contributing to this agenda, our work contributes new strategies for the characterization of perceptions and economic narratives. We test this tools and the methodology results in novel insights regarding policymakers perceptions and narratives..

This work is related to analyses that have measured monetary policymakers' expectations, opinions and policy stance (Bauer and Swanson 2023, Romer & Romer 2008, Sharpe et al. 2023, Malmendier & Nagel 2021, Cieslak et al. 2023, Gorodnichenko et al. 2023). Some of these analyses have implemented LLMs' techniques to extract information from text. Gorodnichenko et al. (2023) use deep learning models to infer the tone from audios of FOMC press conferences. This paper also uses LLMs to measure sentiment in FOMCs documents. Aromi & Heymann (2023) implements three LLM techniques to measure sentiment in FOMC transcripts. Compared to these previous contributions, the current work goes beyond classification tasks and, to the best of our knowledge, we are the first to present a methodology in which time-stamped fine-tuned models are trained to generate text that reflects discussions of specific FOMC meetings. The resulting output is shown to be advantageous in three different tasks.

This study is also related to transparency in Central Bank communication (Woodford 2005, Fischer et al. 2023, Stein 1989). In this respect, the most closely related work is the analysis of Acosta (2023) which finds that, in terms of topic frequency, FOMC minutes reflect deliberations in a transparent manner. Taking a different approach, our analysis focuses on the tone transmitted by FOMC minutes versus the tone inferred from meetings' transcripts. We conclude that the tone transmitted by the minutes is consistent with that reflected in meetings' transcripts. Also

The document is organized as follows. In the next section we present the data and methodology. Section 3 details the results. Concluding remarks are provided in section 4.

2. Data and methodology

Our approach combines a rich corpus with powerful language modeling techniques. The corpus covers detailed records of monetary policy deliberations. Language modeling techniques are used to learn about linguistic regularities in each transcript and use that knowledge to produce text in response to selected trigger phrases.

2.1 Data and construction of the training dataset

Our main dataset is the collection of FOMC transcripts that are provided by the Board of the Federal Reserve System with a 5-year delay. Regular FOMC meetings are carried out eight times a year. Typically, each of these meetings consist of two days in which FOMC members, Fed staff and other authorities of the Federal Reserve System discuss economic conditions, policy options and decide which policies are implemented and how these decisions are communicated. We collect the transcripts of these meetings beginning in the year 2003 and ending in year 2012 to construct an 80-document corpus. These documents contain on average 2713 sentences and 50105 words.

Each of these documents is processed to construct a training dataset that is later used to fine-tune the corresponding language model. In our application, the models are trained to complete phrases; hence, the dataset consists of pairs of phrases from FOMC transcripts in which the beginning of a phrase (input) is matched with the corresponding text that follows in the document (expected output). More in detail, the construction of each training dataset involves three steps. In the first stage, the document is split into sentences. Then, we select the sentences with more than 10 words. In the third stage, from each sentence in this subset, we build training examples. One training example has an input that consists of the first half of the sentence and is matched to an output in the form of a string of characters formed by the second half of the sentence plus the following two sentences. For the second training example, the input is given by the first half of the sentence preceded by the previous sentence in the document and the corresponding output is the second half of the sentence followed by the next sentence in the document. This second type of training example displays shorter outputs, but provides more context to inform the desired output. More than one example is proposed conjecturing that, in this way, the fine-tuned model will acquire more information from the corresponding FOMC transcript. In this way, the number of training examples is two times the number of sentences with more than 10 words.

2.2 Methodology

LLMs are tools widely used in natural language processing. These models receive a piece of text as input and complete a task returning a second piece of text as output. With the appropriate training, a very diverse set of tasks can be performed by this type of tools. For example, we can mention tasks such as information retrieval, text editing and customer service. In this work, we propose to use these models to as time-stamped interactive representations of perceptions and narratives of policymakers. In this way, we can prompt these models to produce text that is informative of the ideas that characterize specific FOMC meetings.

To extract information regarding one aspect of interest at a selected point in time, three steps need to be completed. The first step is model fine-tuning. Models are trained to generate text completions that are judged to be highly likely during a specific FOMC meeting. We perform this step 80 times, one for each FOMC regular meeting during our sample period.

The second stage is text generation. In this step a series of trigger phrases are provided as inputs to fine-tuned LLMs. The models respond with text completions that, according to the learned patterns, are the most consistent with the deliberations during the selected FOMC meeting. For example, if we are interested in perceptions of economic conditions, we use trigger phrases such as: “At this time, economic conditions are…”. To obtain more informative outputs, in this stage we sample more than output per trigger phrase.

In the last step, the outputs are processed to generate an indicator that quantifies the information generated by the model. For this task, we use a natural language inference (NLI), a flexible technique that can be used classify a text in terms of diverse dimensions such as sentiment, topic or other desired features.

Applying this 3-step procedure for different relevant aspects and alternative meetings, we build a collection of time series that reflect evolving policymaking perceptions and narratives. Figure 1 summarizes the three stages followed to produce synthetic text. It is worth noting that the first stage is fixed, that is, the fine-tuning stage is performed once and the resulting fine-tuned models are used for all subsequent tasks. In contrast, the other two stages are adjusted as a function of the use case. This three steps are described in more detail below:

2.2.1 LLMs fine-tuning:

In this step, a pretrained LLM is fine-tuned to carry out a text completion task. The objective is to learn to produce text that is most similar to what was observed during a selected FOMC meeting. More specifically, the language model is trained to process a selected sequence of tokens in the form of an incomplete phrase, or input, and identify the most likely sequences of tokens that completes the phrase, or output. For this purpose, we use the training dataset described in the previous section.

The pretrained model used in this work is Llama 2 7b chat. This is an open-source model offered by Meta. We downloaded the model from Huggingface portal.[1] This model is a high dimensionality nonlinear autoregressive model which, given a sequence of tokens computes the probability of the next token. The model implements the popular transformer network architecture which features multiple layers and channels and an attention mechanism. Through the attention mechanism, the representation of small pieces of text, tokens, is adjusted recursively as a function of the context. More formally, tokens are represented by vectors that, in subsequent layers of the network, are adjusted via a weighted function of its most recent representation and the representation of the neighboring tokens (Vaswani et al. 2017).

Model fine-tuning is carried out using libraries that, as in the case of Llama2, are also accessed through Huggingface portal. In particular, we use the traditional "Transformers" library and implement efficient quantization and low rank adapters (QloRA) techniques through library "PEFT".[2] Through quantization, floating points are represented via less memory demanding low precision formats. Complementarily, fine-tunig with low rank adapters (or LoRA) is a convenient method in which the number of trainable parameters is a small fraction of total model parameters. Under this strategy, the original model is extended to incorporate new weights that transmit information through supplementary low dimension channels. These low rank weights, or adapters, are adjusted during fine-tuning while the original high dimensionality weights are kept frozen. In this way, we avoid prohibitively high computational cost of traditional fine-tuning while performance does not seem to be affected in a noticeable manner (Dettmers et al. 2023). In each instance of model fine-tuning we train for 4 epochs and LoRA dimensionality and "r" parameters were both set to 16.

2.2.2 Text generation:

[1] meta-llama/Llama-2-7b-chat-hf.
[2] https://huggingface.co/docs/peft/index.

In this step we interact with fine-tuned LLMs to extract information in an exercise that might be described as a synthetic survey. For example, to learn about perceived economic conditions we ask the model to complete phrases such as “At this moment, economic conditions are…”. In this way, we propose a generative strategy to extract information from unstructured information.

LLMs generate output estimating the probability of a string of characters conditional on an input in the form of a string of characters. In typical applications if these tools, a single response, the one that is judged more likely, is generated. Given our use case, the characterization of policymakers’ perceptions and narratives, generating a single response would result in significant loss of information. That is, beyond the information provided by the most likely response, there is information that can be acquired from other highly likely answers. That is why in the text generation stage we sample multiple answers. More specifically, we generate multiple outputs in which the each token of each output is the result of a lottery determined by probabilities that are recursively computed by the model.[3]

In most of the cases, we design multiple prompts that pose a similar question with different wordings. That is, if we are interested in perceived economic conditions we use the previously mentioned prompt “At this moment, economic conditions are…” and, in addition, we generate text using similar trigger phrases such as “Currently, the state of the economy is marked by …” and "As of now, economic indicators show…". We construct the set of alternative wordings with the assistance of chatbots (ChatGPT and Bing) as generators of similar examples. The reason for this design feature is that we found that model responses are sometimes very sensitive to specific wordings. This is probably the result of the relatively small time-stamped training dataset that we are able to construct from a single FOMC meeting. Understanding that this large sensitivity to specific wordings might be an undesirable property in our exercise, when possible, we generate text using multiple trigger phrases that have the same meaning.

As an early example of the outputs, figure 2 outlines text produced by different fine-tuned models when prompted to discuss economic conditions. In these examples, we can verify a significant contrast between the text generated by LLMs fine-tuned with different transcripts. The LLM fine-tuned with the transcript from March 2006 produces text that points to a positive economic scenario. In contrast, the LLM trained with the transcript corresponding to March 2008, the during the financial crisis, produces text suggesting deteriorating perceptions.

2.2.3 Text summarization

In this step, we start with a collection of generated text that describes perceptions or narratives corresponding to a specific FOMC meeting. We are interested in generating a quantitative representation of these responses that summarizes the phrases. This numeric representation is used to analyze the evolution of policymakers’ views during our sample period.

To carry out this task we use natural language inference. This is technique in which a model identifies if a first sentence, or premise, implies or contradicts a second sentence, or hypothesis. For example, given the premise “At this time, oil prices have a large impact on consumer prices.” the task might involve identifying entailment or contradiction for the hypothesis “A main driver of inflation is given by energy prices”. In this case, the expected output from the model is a high probability assigned to entailment. In contrast, if the second

[3] This is achieved setting “do_sample” argument to “True” and the choosing the number of sampled outputs through argument “num_return_sequences”.

sentence, or hypothesis, is “This is an example of positive sentiment” the probability assigned to “implication” should be close to zero. In our use case, the first sentence corresponds to text generated by a fine-tuned LLM and the second sentence is used to classify the first sentence, in terms of sentiment, topic or other relevant feature. To implement this task, we use fine-tuned model “bart-large-mnli” a model trained via a multi-genre dataset.[4]

[4] The model is available through Huggingface: https://huggingface.co/facebook/bart-large-mnli.

Figure 1: Workflow for the generation of synthetic text

Pipeline: fine-tuning, text generation, summary of generated text.

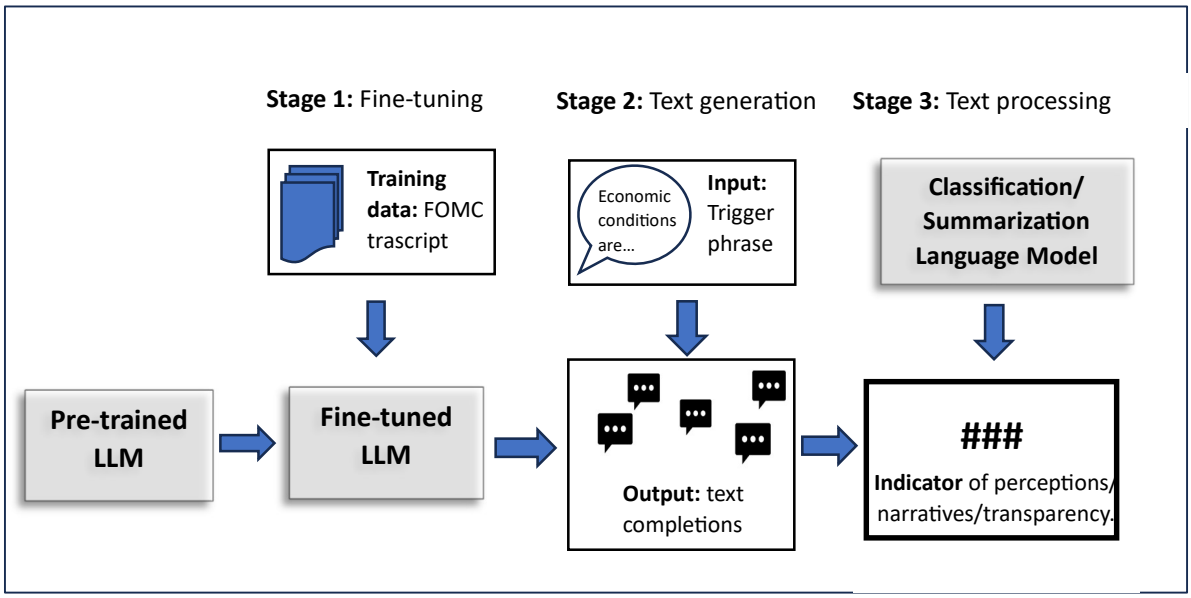
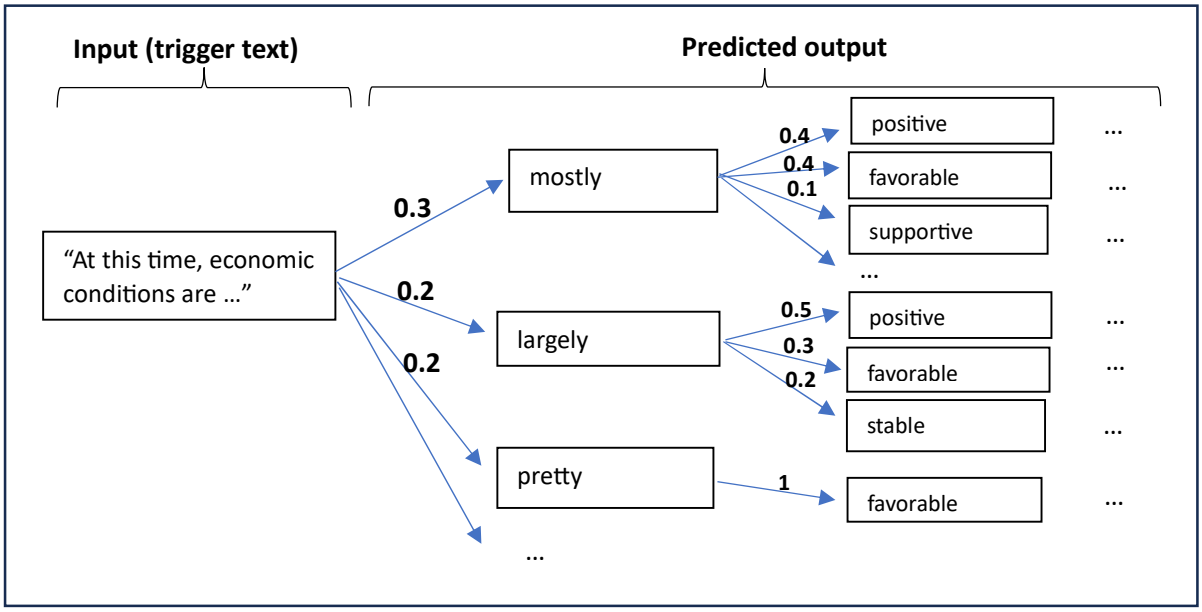
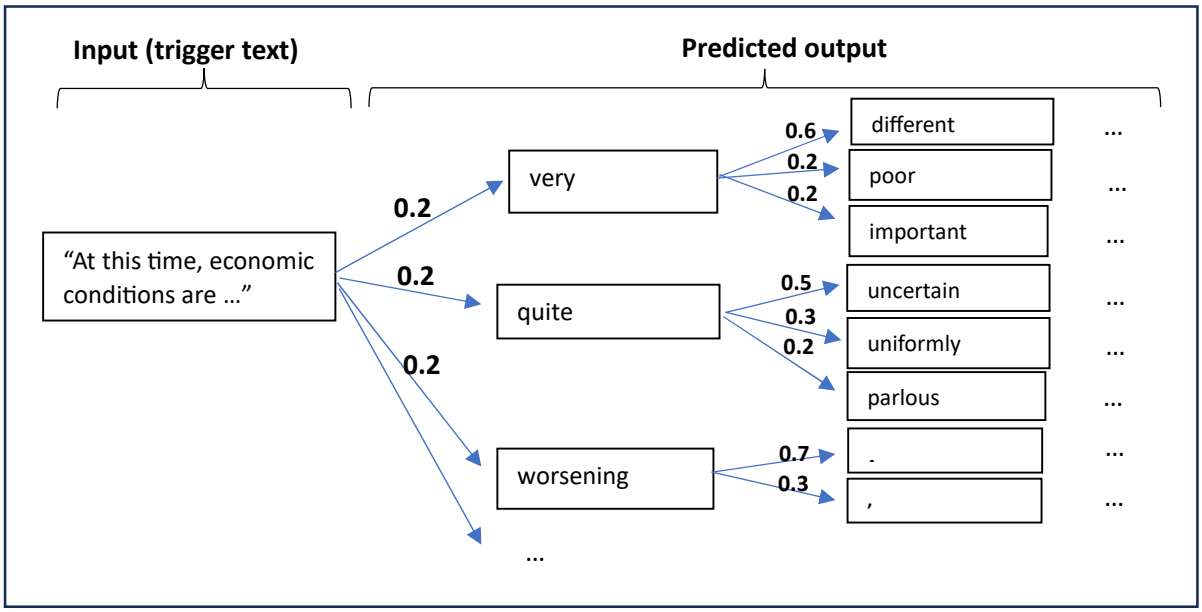


Figure 2: Sample generated texts

A. Model fine-tuned using the transcript corresponding to FOMC meeting of 03/28/2006



A. Model fine-tuned using the transcript corresponding to FOMC meeting of 03/18/2008



Note: Frequency of generated text given the input "At this time economic conditions are…". The number in the arrows indicate the frequency of the indicated word/symbol. In both fine-tuned models, reported frequencies correspond of 25 samples produced setting the using transformers library function "pipeline" and setting the argument "do_sample" to True.

3. Results

In this section, we detail three different exercises through which the proposed methodology is evaluated. First, we show how fine-tuned LLMs completions can be used to elicit policymakers' perceptions of economic conditions. Next, we implement a test of communication transparency in which we show how LLMs can be used to test if FOMC meetings' tone is properly transmitted through meetings' minutes. Finally, we characterize policymaking narratives implementing a diverse set of text completion tasks.

3.1 Indicators of perceived economic conditions

Policymakers allocate valuable resources to the analysis of economic conditions. These analyses result in an informed understanding of the state of the economy that explains policy decisions. It is worth noting that policymakers' perceptions constitute a latent state. The evolution of policymakers' perceptions and their information content need to be inferred from decisions and communications.

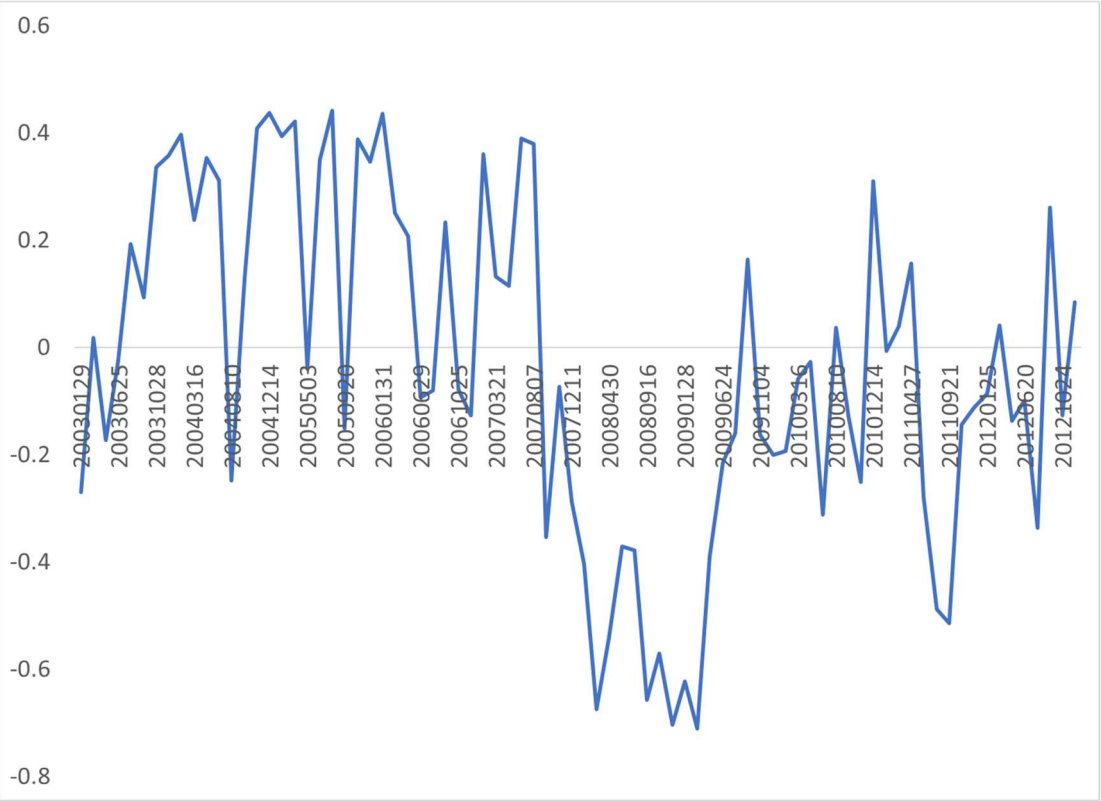
The existing literature has proposed approximations in which the text from different documents produced by policymakers is classified in terms of relevant dimensions such as sentiment, topic, uncertainty and hawkishness vs. dovishness (Hansen & McMahon 2016, Apel et al. 2022, Gorodnichenko et al. 2023, Cieslak et al. 2023). We propose a novel approach to assess policymakers' perceptions of economic conditions. In this exercise, we exploit LLMs' capacity to generate language that is representative of the training corpus. In our case, we train LLMs with meetings' transcripts and use the acquired skill to generate text that communicates policymakers' perceptions about the state of the economy.

With some special features, that are described below, we apply the methodology detailed in the previous section. To generate text completions that are informative of policymakers' views about the economy, we build multiple trigger phrases with similar meaning. As previously discussed, with this feature our objective is to reduce the noise that might result from fine-tuned LLMs that respond excessively to particular phrases. More precisely, we build 64 prompts, or trigger phrases, that result from the combination 8 different references to the present time and 8 alternative references to economic conditions. The first set of 8 phrases is given by: "At this time, ", "Currently, ", "At present, ", "In the current period, ", "As of now, ", "At this point in time, ", "At this moment, " and "Presently, ". References to economic conditions are given by: "economic conditions are", "the economic climate is", "the state of the economy is", "the economy shows", "the economic landscape hints at", "'economic indicators show", "the economic environment points to" and "the overall economic scenario reveals". Additionally, to capture more information regarding economic perceptions during the meetings, instead of selecting the most likely completion, the completion of each trigger phrase is sampled 10 times. In terms of choosing the length of the completion, we set the maximum number of generated tokens to 25. In this way, for each FOMC meeting, we generate 640 phrases that communicate perceptions of economic conditions.

Once the text completions have been generated, each completed phrase is classified in terms of sentiment using NLI. The sentiment of each phrase is equal to the probability assigned to positive sentiment minus the probability assigned to negative sentiment. The scores are averaged to produce the indicator of perceived economic conditions on a given date. The resulting sentiment index is reported in figure 3. The main feature of the figure is the prominent and persistent drop in sentiment that starts in the second half of 2007. To an important extent, the lowest values coincide with the Big Recession (December 2007-June

2009). Another distinguishable pattern is the relatively subdued sentiment manifested in the years that follow this downturn. Additionally, some declines in sentiment can be linked to adverse events such as the start of the War in Iraq in early 2003, Hurricane Katrina (August 2005) and the Debt Ceiling crisis of 2011. Moving beyond this descriptive analysis, the information content of this indicator is evaluated formally through forecast exercises detailed below.

Figure 3: Policymaking perceptions of current economic conditions



Notes: The figure shows the standardized metric of sentiment computed from the text completions that provide time-stamped perceptions of economic conditions.

To evaluate the information content of the synthetic sentiment index we first assess its ability to anticipate revisions in GDP growth forecasts. We consider the Survey of Professional Forecasters. This is a quarterly private sector forecasts collected by the Philadelphia Federal Reserve. These forecasts are collected and released by the middle of each calendar quarter. To avoid forward looking biased estimates, the sentiment metric of each quarter corresponds to the first FOMC meeting. This meeting takes place a couple of weeks before survey data of the corresponding quarter is collected. We test if the synthetic metric of sentiment anticipates revisions in the forecast released after the corresponding meeting. Let $\hat{g}_{q,q+4}^{q'}$ represent a forecast released in quarter q' that targets cumulative GDP growth from quarter q through quarter $q+4$. Then, our metric of forecast revisions is given by: $Rev_{q+h}^{q} = \hat{g}_{q,q+4}^{q+h} - \hat{g}_{q,q+4}^{q}$. That is, the cumulative change in the yearly growth forecast that is released in quarter q and adjusted in quarter $q+h$.

We estimate simple forecasting models in which the sentiment index predicts growth forecast revisions. The empirical model is given by:

$$Rev_{q+h}^{q} = \alpha + \beta_{+h} sent_q + u_{q+h}$$

Where the sentiment metric for the first meeting of quarter q is $sent_q$, u_{q+h} is a noise term and h indicates the forecast horizon. The parameter of interest is β_{+h}. An estimated value different from zero would indicate that the sentiment index contains information that anticipates changes in private sector forecasts that are released (and revised) in future dates. A positive estimated value is consistent with advantageously informed policymakers.

We extend the analysis considering two variations of the index of economic conditions. First, we shift the time orientation. Instead of discussing current conditions we prompt models to discuss future economic conditions. To achieve forward looking prompts, we modify both the time references and verb tenses.[5] In addition, considering the key role played by financial conditions during the sample period, we modify the text of the trigger phrases to prompt references to financial, instead of economic, conditions.[6]

Table 1 shows the estimated coefficient of the sentiment metrics for different specifications of the indicator synthetic policymakers' views. Overall, the results suggest that the synthetic survey is able to extract valuable information that is incorporated gradually by private sector forecasters during the subsequent quarters. The associations are economically significant. Suggesting gains in information content as we modify the target evaluation through prompt modifications, estimated coefficients are larger as we move toward forward looking assessments and focus on financial conditions.

Table 1: Forecasting revisions in cumulative growth forecast

	Target evaluated conditions		
	Current econ.	Future econ.	Future financial
$\hat{\beta}_{+1}$	0.1634	0.2664**	0.2857***
	(0.115)	(0.119)	(0.107)
$\hat{\beta}_{+2}$	0.2655*	03998*	0.4188**
	(0.232)	(0.241)	(0.197)
$\hat{\beta}_{+3}$	0.3364	0.5278	0.5861**
	(0.312)	(0.323)	(0.296)

Notes: The table reports standardized estimated coefficients for the corresponding tone metric ($\hat{\beta}$). Heteroskedasticity-autocorrelation robust standard errors reported in parenthesis.

[5] In this case, the group of phrases indicating time orientation is: time: 'Looking ahead to the coming months, ', 'Over the next quarters, ', 'In the near future, ', 'On the horizon, ', 'In the immediate future, ', 'As we move forward, ', 'In the short term, ' and 'Advancing into the future, '. Accordingly, we also changed the tense of the second part of each trigger phrase.

[6] For this specification of the index, the second group of phrases pointing to financial conditions is given by: 'financial conditions will be', 'the financial climate will be', 'the state of the financial system is expected to', 'the financial system is expected to', 'the financial landscape is likely to', 'financial indicators are anticipated to', 'the financial environment is estimated to' and 'the overall financial scenario will show'

To generate complementary evaluations of the information content of synthetic sentiment indices we estimate forecasting models corresponding to four economic indicators. We consider simple forecasting models in which the sentiment index is incorporated to an autoregressive model. The empirical model is given by:

$$y_{t+h} = \alpha + \beta y_t + \beta_{+h} sent_t + u_{t+h}$$

Where y_t is an economic variable, the sentiment metric is $sent_t$ and u_{t+h} is a noise term. The time index t corresponds to the second day of the corresponding FOMC meeting and h indicates de forecast horizon. The four economic variables are: stock market expected volatility (VIX), stock market cumulative returns (S&P 500), consumer sentiment (Univ. of Michigan) and press sentiment (WSJ headlines). The value of the VIX and S&P 500 cumulative are computed considering the day in which the corresponding meeting ends. No lagged term is used in the case of stock returns. Consumer sentiment corresponds to the latest released value as of the last day of the meeting. To smooth high frequency fluctuations, WSJ sentiment corresponds to the average value for the 28-day period that ends on the last day of the corresponding meeting.

The results shown in table 2 are consistent with the findings reported in the case of growth forecast revisions. Fine-tuned LLMs seem to be able to capture policymakers' perceptions of economic conditions. These extracted perceptions contain information regarding the trajectory economic and financial variables during the following months. As in the previous evaluation, the evidence on the information content of the indices is strongest in the case of the indicator that of financial conditions in the future.

Table 2: Information content of synthetic sentiment indices

	VIX	Stock market returns	Consumer sentiment	Press sentiment
A. Current economic conditions				
$\hat{\beta}_{+1}$	-2.3031	0.7310	0.2665***	0.1600**
	(1.517)	(0.784)	(0.065)	(0.082)
$\hat{\beta}_{+2}$	-3.3772*	1.0528	0.2765***	0.1276
	(1.972)	(1.391)	(0.095)	(0.110)
$\hat{\beta}_{+4}$	-4.6988*	1.0832	0.1674	0.1158
	(2.800)	(2.869)	(0.116)	(0.170)
B. Future economic conditions				
$\hat{\beta}_{+1}$	-1.5323	0.7510	0.2179***	0.2315*
	(1.265)	(0.741)	(0.053)	(0.084)
$\hat{\beta}_{+2}$	-3.0467	1.3796	0.1605	0.2252**
	(1.972)	(1.382)	(0.100)	(0.095)
$\hat{\beta}_{+4}$	-3.8028	1.0814	0.1315	0.1200
	(2.653)	(2.767)	(0.128)	(0.160)
C. Future financial conditions				
$\hat{\beta}_{+1}$	-1.2871	1.2541*	0.1892***	0.1638*
	(0.952)	(0.656)	(0.043)	(0.090)
$\hat{\beta}_{+2}$	-2.8014*	2.4014**	0.2403***	0.1965*
	(1.672)	(1.150)	(0.072)	(0.109)
$\hat{\beta}_{+4}$	-1.8802	2.4572	0.1843	0.1518
	(1.460)	(2.056)	(0.138)	(0.179)

Notes: The table reports standardized estimated coefficients for the corresponding tone metric ($\hat{\beta}$). WSJ sentiment is computed for economic headlines using NLI positivity and negativity scores. Heteroskedasticity-autocorrelation robust standard errors reported in parenthesis.

3.2 Auditing communication strategies

Communication is a strategic decision. This is particularly clear in the case of policymaking. Through communication, policymakers can transmit information about economic events, can build reputation and commit to certain policy actions. Additionally, communication can play a central role in coordinating behavior and, hence, setting a path for the economy.

Despite its importance, communication by monetary policymakers is incompletely understood. While there are benefits from building a reputation for truthful telling (Woodford 2005), heterogeneous objective functions might prevent full transparency (Stein 1989). There are no

warranties of transparency or truthfulness. This is particularly true in the context financial crises. In extreme scenarios, policymakers might have incentives to send biased reassuring messages to avoid disorderly exits such as self-reinforcing liquidity crises (Brunnermeier and Pedersen 2009). Our sample period covers a deep crisis in which policymakers might have perceived benefits associated to sending reassuring messages. Our methodology, offers an opportunity to produce a synthetic audit of monetary policy communication.

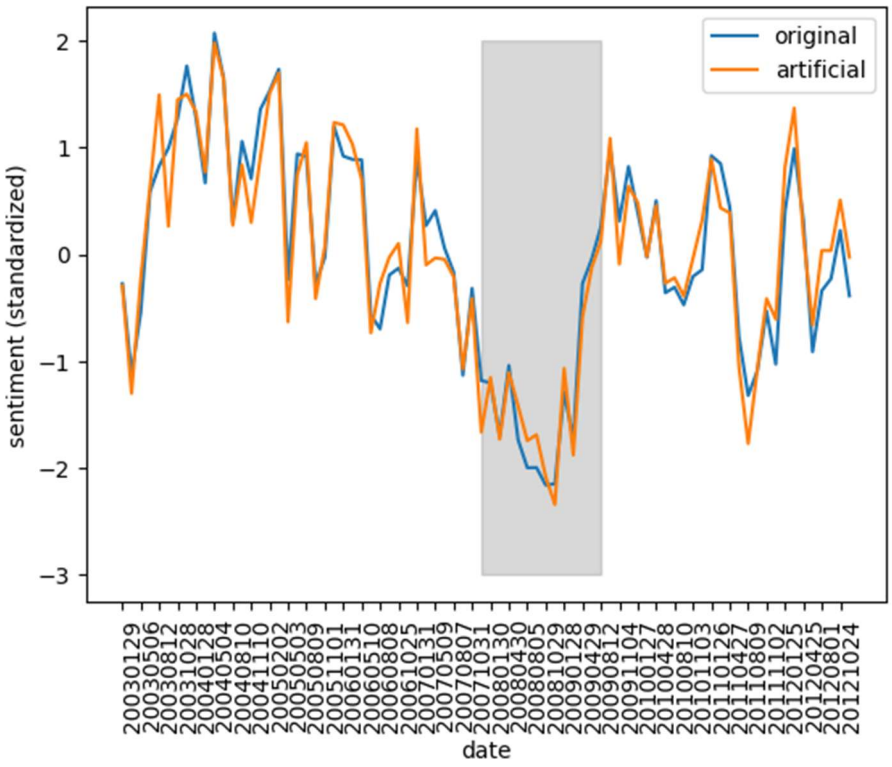
We compare the tone of FOMC meeting minutes to the tone of minutes that are partially generated via fine-tuned LLMs. More in detail, given an FOMC meeting, we generate partially synthetic minutes in which the first half of each sentence corresponds to a sentence of the original minutes and the second half of each sentence is generated by the corresponding fine-tuned LLM. To generate the text that completes each sentence, we use the first half of that sentence as a prompt of a text completion task. As in the previous task, to produce a more informative analysis, we generate five partially synthetic documents carrying out multiple text completions of each selected sentence. Once the alternative texts are generated, we compute the sentiment of each output using NLI. We compare the average sentiment of these sentences to the average sentiment of the sentences in the original minutes. Before proceeding to the results, it is worth noting that this evaluation focuses on a specific feature of FOMC communication. Since the first part of each sentence is taken from the original minutes and we can suppose that the topic is, to a large extent, set by the first half of the sentences. Our analysis is complementary to Acosta (2023) that evaluates transparency of FOMC minutes in terms of the frequency with which each topic is discussed.

Figure 4 shows the two indices of minutes sentiment after standardization. The very similar trajectories of the indicators suggest that FOMC minutes transmit a tone that is consistent with the tone of the meeting. While this visual inspection seems very compelling, we also carry out formal tests. Considering that changes in economic conditions might lead to shifts in incentives, we evaluate if deviations between these two metrics of sentiment are associated to economic or financial conditions. Let Dif_t indicate the difference between the two standardized metrics of minutes sentiment (original minus synthetic) and x_t represent an indicator of economic conditions. Then, we implement our evaluation estimating the following simple model:

$$Dif_t = \alpha + \beta x_t + u_t$$

An estimated value of β different from zero would suggest that there exists a systematic term in the disparity between minutes and transcripts tone. This systematic component might be linked to changing incentives in Central Bank communication. We consider three variables indicating economic conditions: expected economic growth over the next quarters from the Greenbook/Tealbook, expected stock market volatility (VIX) and FOMC sentiment as reflected in the synthetically generated text dealing with current economic conditions. The results, reported in table 3, indicate that the null hypothesis of no association cannot be rejected. This finding serves as further evidence of truthfulness in Central Bank communication.

Figure 4: Standardized sentiment indices: Original vs. LLM Generated



Notes: The figure shows the standardized metric of sentiment computed from FOMC minutes (original and artificial).

Table 3: The association between sentiment difference and economic indicators

	Growth Forecast	VIX	FOMC Sentiment
$\hat{\beta}$	0.0182	0.0011	-0.0013
st. errors	(0.021)	(0.002)	(0.084)
$adj.R^2$	-0.005	-0.011	-0.013
N	80	80	80

Notes: The table reports standardized estimated coefficients for the corresponding tone metric ($\hat{\beta}$). Heteroskedasticity-autocorrelation robust standard errors reported in parenthesis.

3.3 Policymaking narratives

In this section we use fine-tuned LLMs to extract features of policymaking narratives. In this way we generate a detailed characterization of how the economy and policymaking is understood by participants of FOMCs' meetings. In this exercise, we adapt the text generation tasks to extract features of policymaking narratives. That is, we move beyond the previous sections in which the focus was placed on how generated texts map to metrics of polarity or sentiment.

According to Shiller (2019), a narrative is a "particular form of a story, or of stories, suggesting the important elements and their significance to the receiver". These stories constitute key elements that provide a framework for representations and evaluations that guide decision-

making. Our understanding of monetary policy and, more broadly, macroeconomic dynamics can benefit from a more precise and systematic documentation of narratives.

To extract features of policymaking narratives, we start by requesting descriptions of the role and worries of policymakers. Next, with a focus on how the functioning of the economy is understood, we ask LLMs to identify main drivers of economic growth and inflation. Then we produce categorizations of the performance of the economy in terms of traditional labels (recession, crisis, fragility). Finally, we use LLMs to evaluate if, and when, policymaking discussions feature emotionally charged narratives.

3.2.1 Monetary policy goals

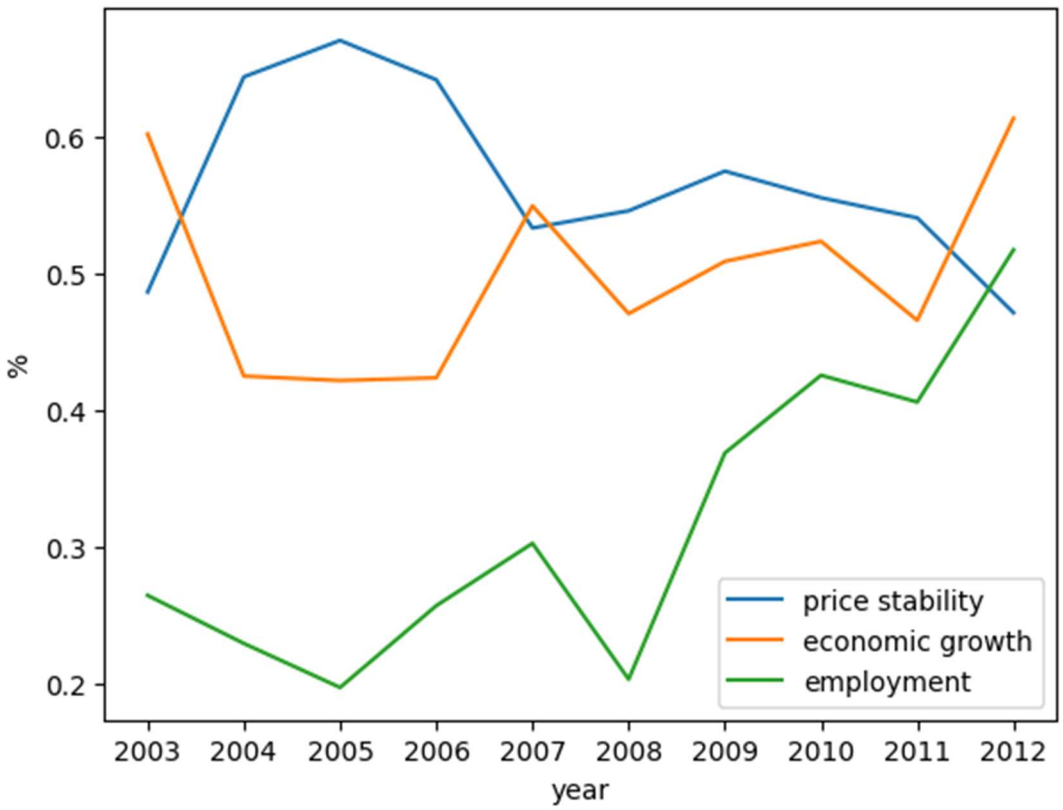
The objectives of monetary policy have been subject to analysis in academic and policy circles both from a normative and positive perspective (Bernanke 2013, Hakes 1990, Christian 1968). We move beyond the analysis of decisions and public statements and assess which are the goals that can be inferred from closed door policy deliberations.

To extract information regarding policymakers' goals that are consistent with FOMC deliberations we adapt the general methodology described in section 2 to this specific use case. We design 5 different prompts, or trigger phrases, with meaning similar to: "Our main focus as policymakers is to achieve…".[7] For each trigger phrase and fine-tuned LLM we sample 25 text completions with a maximum number of 25 tokens. The generated text is then classified through a multi-label implementation of NLI. In this classification task, for each generated piece of text, the model assigns probabilities to a list of possible labels. Since the classification mode is multi-label, these probabilities do not necessarily add up to one. The appropriateness of each label is evaluated independently by the model. The list of candidate labels is: "price stability", "economic growth" and "unemployment. These labels are selected through expert judgement after inspecting frequently mentioned concepts in a sample of text completions. NLI classification provides probabilities that, after averaging across outputs, are interpreted as the frequency with which a goal is mentioned in a synthetic survey.

The results are reported in figure 3. When we consider average values for the 10-year period, we observe a balanced focus on two traditional main objectives: price stability and economic growth. When prompted to mention goals, fine-tuned LLMs' responses refer to price stability, on average, 57% of the time. This frequency is higher but close to the average frequency with which LLMs refer to economic growth (50%). Beyond these average values, we observe some variation during the sample period. During years 2004 through 2006 the price stability goal was significantly more prominent than economic growth. This large gap is largely gone during the subsequent years that cover the crisis and the following years. Furthermore, during the last 4 years of the sample period, we detect a significant increment in references to unemployment, a third goal that is different but closely linked to economic growth. These results are suggestive of a shift in the prominence of different objectives that might be explained by the weak recovery from the Big Recession.

[7] The complete lists of phrases used in this task, and the other NLI classification tasks described below, can be found in appendix A.

Figure 3: Monetary Policy Objectives



Notes: average mentions of alternative monetary policy objectives in LLM generated text. Frequencies correspond to NLI classifications.

3.2.2 Sources of policy concerns

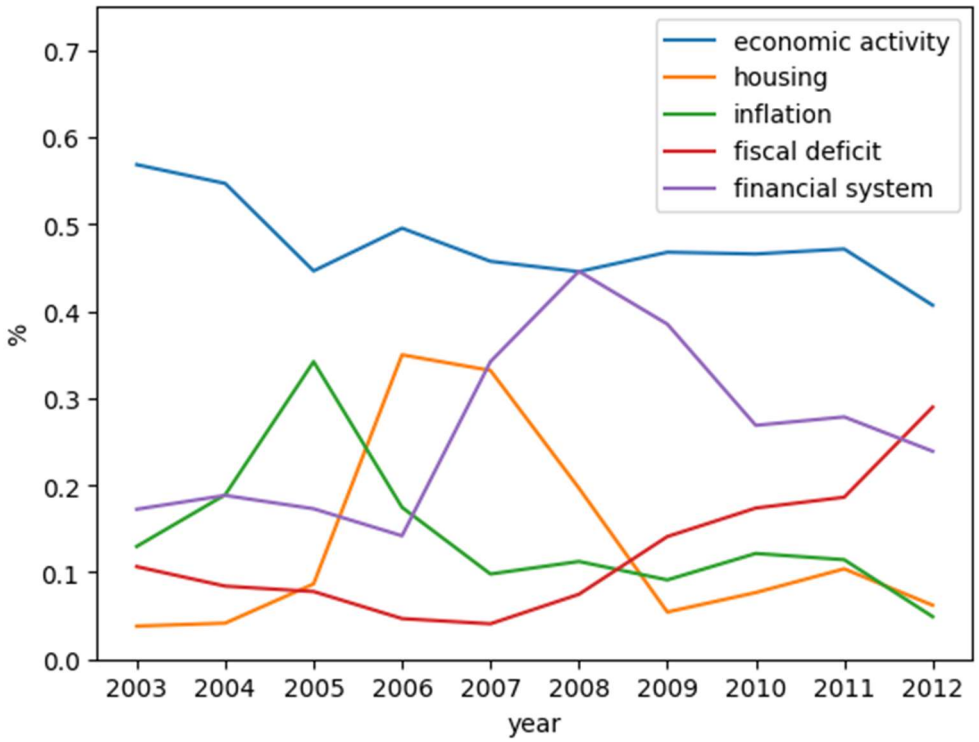
One key feature of monetary policy narrative is the assessment of the most relevant concerns that might interfere with the achievement of policy objectives. These perceived threats are likely to represent the focus of attention at a given point in time. The manifested worries can be conjectured to emerge from the combination of two elements: policymakers' goals and the way in which the functioning of the economy is understood at a given point in time.

To extract information regarding the major worries of policymakers, we first generate text using 5 trigger phrases with meaning similar meaning to: "The biggest threat for the economy is". Then, we implement multi-label NLI classification using labels that emerge from expert judgement after inspection of a random sample of generated pieces of text. NLI classification provides probabilities that, after averaging across generated outputs, are interpreted as the frequency with which a concern is mentioned in a synthetic survey.

In contrast to the case of monetary policy goals, when we evaluate threats, there is a single topic that clearly emerges as the main source of concern. The most frequently mentioned concern is economic activity. This topic is mentioned as a source of major threats in 48% of the synthetic survey answers. Beyond this main theme, we identify four other frequently mentioned sources of concern. The financial system is mentioned in 26% of the answers. Next, we find inflation (14%), housing (13%) and the fiscal deficit (13%).

Figure 4 shows the frequency of mentions by year. The central position of economic activity is a persistent feature of our analysis. The frequency of mentions of economic activity is above 40% in all years of the sample. Demonstrating the shifting dynamics in the economic scenario and associated policy challenges, when we consider other relevant concerns, we observe a sequence of spikes. Taking turns, these spikes position one of those topics as the second most important threat. First, we detect worries about inflation that peak in 2005. Next, housing emerges as the second biggest concern by 2006. This is followed by a period of five years in which the financial system takes the second spot. By year 2008, during the most acute period of the financial crisis, this concern is mentioned in more than 40% of the generated answers. Finally, in the final year of the sample period, the fiscal deficit emerges as the second biggest concern.

Figure 4: Source of policy concerns



Notes: Average mentions of alternative monetary policy objectives in LLM generated text. Frequencies correspond to NLI classifications.

3.2.3 Macroeconomic drivers

Macroeconomic dynamics are complex. They are determined by multiple interacting forces. In addition, dynamics are shaped by economic agents' anticipation of how these drivers will influence the trajectories in the future. In this challenging context, to avoid the curse of dimensionality, policymakers need to identify the most important forces. The analysis of the economic environment is constructed focusing on these prominent drivers. The evolution identified economic driver in narratives can be useful to describe policymakers' views and rationalize monetary policy decisions.

To extract this feature characterizing monetary policy narratives, we prompt LLMs to discuss main macroeconomic drivers. As in the previous exercise, to generate a more informative collection of texts, we use a list of trigger phrases with similar meaning and, for each of these trigger phrases, we sample 25 outputs. The generated text is later classified using multi-label NLI jointly with a list of candidate labels that were selected using subjective judgement after inspecting a sample of generated text.

Table 4 shows the frequency with which a collection of prominent growth drivers are mentioned. The most frequently mentioned force is aggregate demand. According to NLI classification, this force is mentioned with a frequency of 46%. Demonstrating a notable asymmetry, this figure almost triples the frequency with which aggregate supply is mentioned. In the same line, consumption, the largest component of aggregate demand, is the second most mentioned driver with a frequency of 34%. In third position, other driver frequently referred driver is economic policy (31%).[8] LLMs generated text suggest that financial markets constitute another prominent driver with a frequency of 19%.

When we evaluate the frequency of mentions by year we observe, for the most part, a stable scenario. Figure 5 shows the evolution for five representative labels.[9] Aggregate demand appears as the main driver for most of the sample period. The most prominent change is observed following the Great Recession. This shift involves a drop in references to aggregate demand accompanied by an increase in the relevance economic policy. Financial markets rank as an important driver specially since 2006.

One notable demonstration of stability is given by the yearly frequency of references to aggregate supply. This figure is consistently between 10% and 20%. It is also worth noting that confidence is another example of relatively stable frequency of references and, contrary to what could have been conjectured, the frequency of referrals to this driver does spike in the context of the 2008/09 financial crisis.

[8] In untabulated extended analysis we find that, according to NLI classifications, both monetary policy and fiscal policy are mentioned with a frequency of 9%.

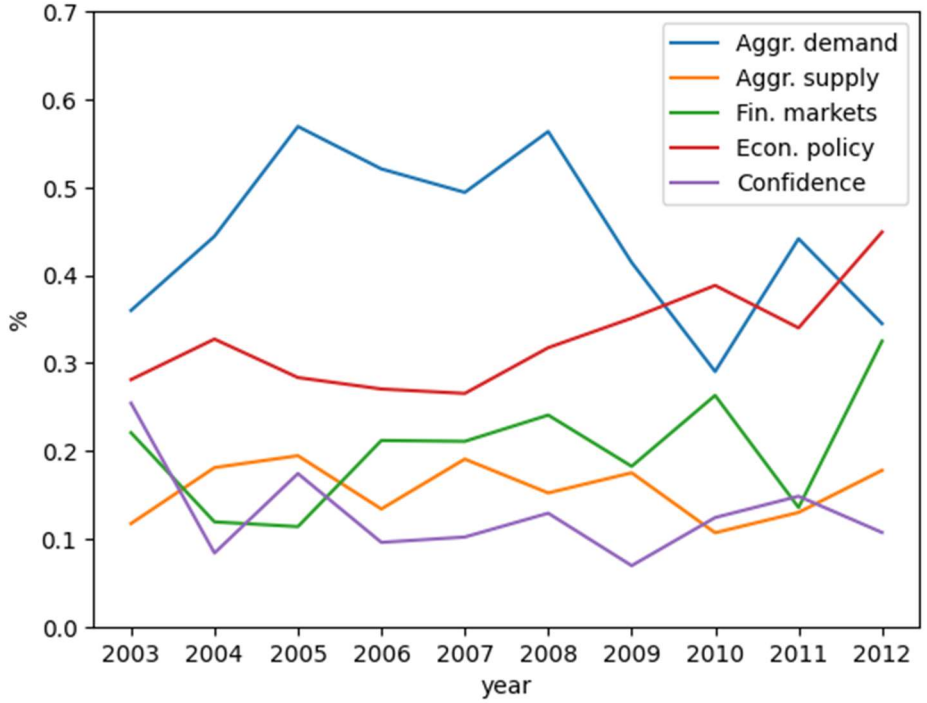
[9] To facilitate the visualization we choose to eliminate the line corresponding to consumption. This is a component of aggregate demand with a line displaying a very similar trajectory

Table 4: Main growth drivers – Frequency of mentions

Drivers	Frequency
Aggregate Demand	0.456
Consumption	0.340
Economic policy	0.309
Financial Markets	0.186
Aggregate supply	0.154
Confidence	0.126
Investment	0.086
Job Creation	0.079
Productivity	0.070
Inventories	0.046
Oil prices	0.041
Net exports	0.026

Notes: Average frequency of mentions of main drivers of economic growth in LLM generated text. Frequencies correspond to NLI classifications. Labels are selected after manual inspection of sample synthetic responses. Sample period 2003-2012.

Figure 5: Main growth drivers – Frequency of mentions by year



Notes: Average mentions of growth drivers in LLM generated text. Frequencies correspond to NLI classifications.

In the case of inflation, as shown in table 5, oil prices is the most frequently mentioned driver with an average frequency of 31%. According to generated outputs, the second most prominent driver of inflation is economic activity with a frequency of 25%. Next, we find a diverse set of drivers with a frequency close to 10%. This heterogeneous group includes aggregate demand, expectations, past inflation, exchange rates and wages.

It is worth noting that in this case, in contrast to what we observe in the analysis of growth drivers, we do not find economic policy as a prominent driver. Monetary policy is mentioned only 7% of the time and the fiscal deficit is rarely signaled (1%). This result suggests that under the macroeconomic regime in force during our sample period, monetary policymakers do not typically view their actions as having an immediate and considerable impact on inflation.

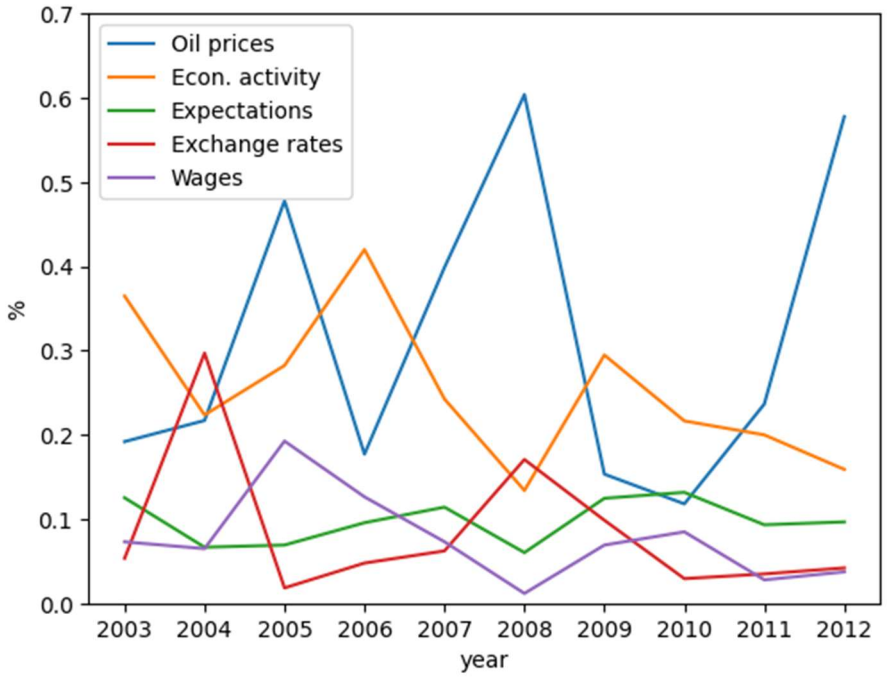
We also analyze the frequency of mentions of forces driving inflation by year. Figure 6 shows the trajectories for a sample of 5 relevant drivers. We verify that in the case of oil prices, the high average frequency is a consequence of multiple spikes that arise from relatively low initial levels. These spikes in years 2005, 2008 and 2012 are seen to coincide with instances in which oil prices increased in a significant manner. Exchanges rates were the most frequently mentioned driver in year 2004. Economic activity is the most mentioned driver in year 2006 and in the years that immediately follow the financial crisis. Expectations constitute a driver with a notably stable trajectory in the range of 10%. This is similar to what we observe for the trajectory of references to past prices.

Table 5: Main inflation drivers – Frequency of mentions by year

Drivers	Frequency
Oil prices	0.310
Economic activity	0.252
Aggregate demand	0.110
Expectations	0.098
Past inflation	0.091
Exchange rates	0.085
Wages	0.075
Monetary policy	0.067
Fiscal deficit	0.012

Notes: Average frequency of mentions of main inflation drivers in LLM generated text. Frequencies correspond to NLI classifications. Labels are selected after manual inspection of sample synthetic responses. Sample period 2003-2012.

Figure 6: Main inflation drivers – Frequency of mentions by year



Notes: Average mentions of inflation drivers in LLM generated text. Frequencies correspond to NLI classifications.

3.2.4 Classification of economic conditions

The state of the economy is many times represented in terms of categories such as recession, fragility and crisis. This feature of policymaking narratives is relevant since it they are likely to influence policy evaluations and decision making (Shiller 2019, Mullainathan 2002). For example, if the economy is viewed in a recessionary state policymakers might conjecture different behavioral patterns by economic agents. In addition, in this scenario, they might respond more aggressively to any adverse unemployment news. In response to a state labeled as a crisis, policymakers might be more open to extreme monetary and fiscal interventions. Finally, if the economy were considered in a persistent state of fragility, there would be more willingness to sustain unconventional policies during an extended period.

To assess categorization of economic conditions in policymaking narratives we design prompts such as "Are we currently in a recession? Currently we are…". As in the previous exercise, to extract more information, we generate 25 text completions of each trigger phrase. The completion of this phrase, and similar phrases, by fine-tuned LLMs are later processed through NLI to identify if the generated answer is an affirmative or negative answer. The model assigns probabilities to these two labels. Then, we average answers by FOMC meeting to generate a time series of the synthetic surveys.

Figure 7 reports the resulting classifications for three categories: recession, crisis and fragility. In each case, the indicator is equal to the difference between the average probability assigned to a positive answer minus the average probability assigned to a negative answer. Despite some high frequency volatility, a common pattern can be detected. In each case we observe a visible increment by the end of 2007. Then, during the recession, the indices remain at elevated levels. After the recession the indicators drop but average values stay significantly higher than those observed before the crisis. Table 6 reports the average values of these indicators before, during and after the Big Recession. The same common patterns are again very noticeable.

Some more specific observations help us visualize the insights resulting from these evaluations. When we focus on the assessed likelihood of a recession, reported in panel A of figure 7, we find that establishing a threshold somewhere between 0.6 or 0.7 the synthetic answers generate surprisingly precise classifications. That is, synthetic responses result in accurate pseudo real-time assessments of business cycle peaks and troughs that are reported about a year later by the NBER's Business Cycle Dating Committee.[10] It is also of interest to note that LLMs generated content show how the Big Recession was characterized as a crisis very early on. Finally, providing a rationale to the persistent of unconventional monetary policy measures, we can verify how The Big Recession is followed by a persistent perception economic fragility in policymaking narratives.

Table 6: Classification of economic conditions

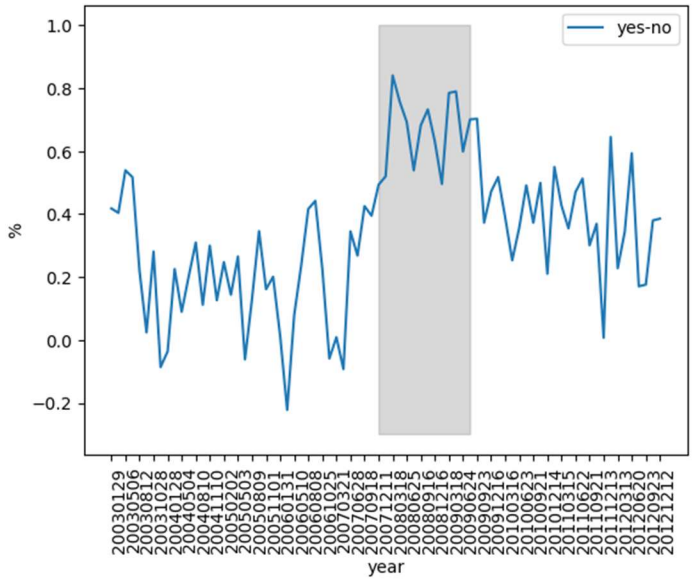
	Category		
Period	Recession	Crisis	Fragility
Pre-recession (before December 2007)	0.199	0.187	0.294
Recession (December 2007- June 2009)	0.658	0.653	0.690
Post-recession (after June 2009)	0.402	0.393	0.507

Notes: The table reports, for each category and each period, the difference between the average probability of a positive answer and the average probability of a negative answer. The probability is assigned using NLI classification.

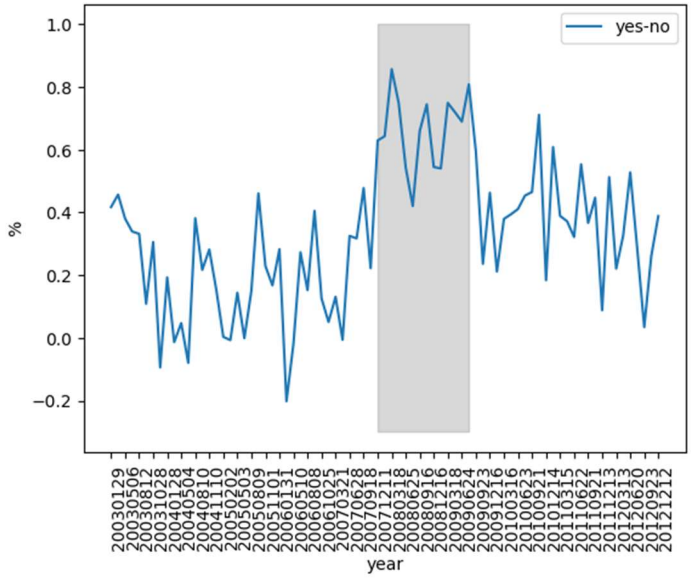
[10] The date of the announcements can be found in: https://www.nber.org/research/business-cycle-dating/business-cycle-dating-committee-announcements

Figure 7: Categorization of economic conditions

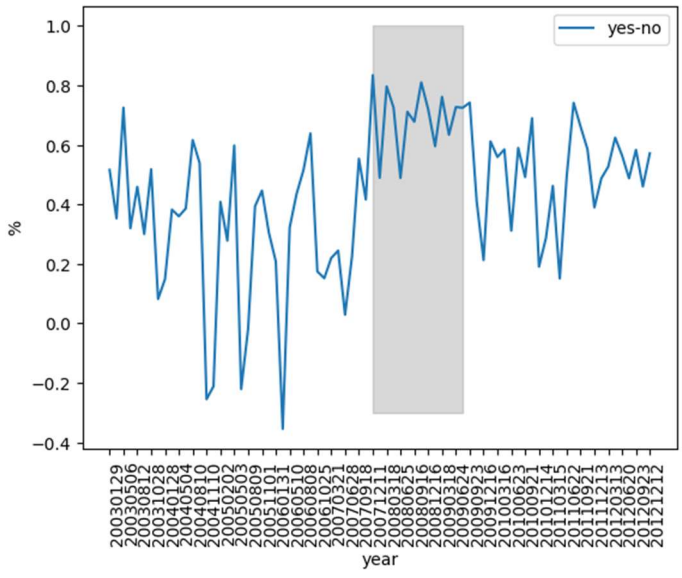
A. Is the economy in a recession?



B. Is the economy in crisis?



C. Is the economy fragile?



3.2.5 Emotions in narratives

In this section, we interact with fine-tuned LLMs to document the presence of emotions in monetary policy narratives. Emotionally charged narratives can provide information on the assessments of policymakers regarding economic conditions and, at the same time, can influence decision making. The manifestation of emotions by policymakers and the evaluation of their informational content has been previously implemented using voice recordings (Gorodnichenko et al. 2023) and facial expressions (Aromi and Clements 2021). We extend this literature using fine-tuned models to generate text that signals the extent to which policymaking deliberations are charged with emotional content.

We consider two emotions: sadness and fear. These are two negative emotions that are conjectured to be more noticeable during periods of economic difficulties. To extract information, we design two sets of trigger phrases with meanings that are similar to: "Are you depressed? I feel …" and "Are you feeling a sense of angst? I …". The underlying conjecture is that fine-tuned LLMs' completions of these phrases would provide evidence of the extent to which narratives are charged with any of these emotions. As in the previous analysis, we implement NLI classification of the generated text. In this case the classification involves assigning probabilities to the presence or absence of the respective emotion in the text.

Table 7 reports average values for three periods. According to generated responses, the Big Recession is associated to narratives charged with an increased perception of sadness. We observe a large increment that is only partially reversed during the post-recession period. In the case of fear, we observe a more moderate increment during the economic downturn. This increment is completely reversed during the post-recession period. In more detail, figure 8 show the evolution of the index reporting the computed metrics for each sample FOMC meeting. Despite the high frequency volatility, the Big Recession can be distinguished as a period in which narratives were characterized by negative emotions.

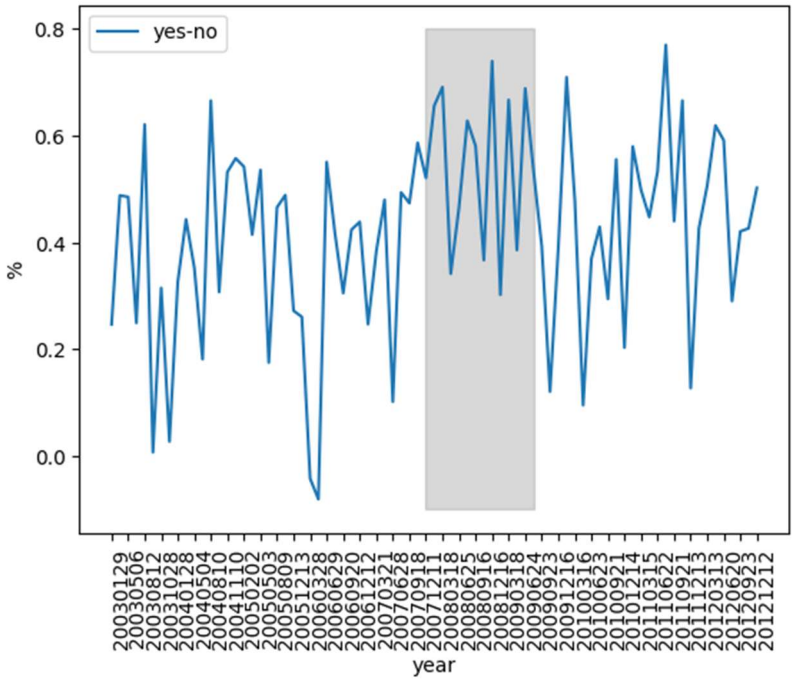
Table 7: Manifestations of emotions

	Emotion	
Period	Sadness	Fear
Pre-recession (before December 2007)	0.362	0.337
Recession (December 2007- June 2009)	0.541	0.440
Post-recession (after June 2009)	0.444	0.323

Notes: The table reports, for each category and each period, the difference between the average probability of a positive answer and the average probability of a negative answer. The probability is assigned using NLI classification.

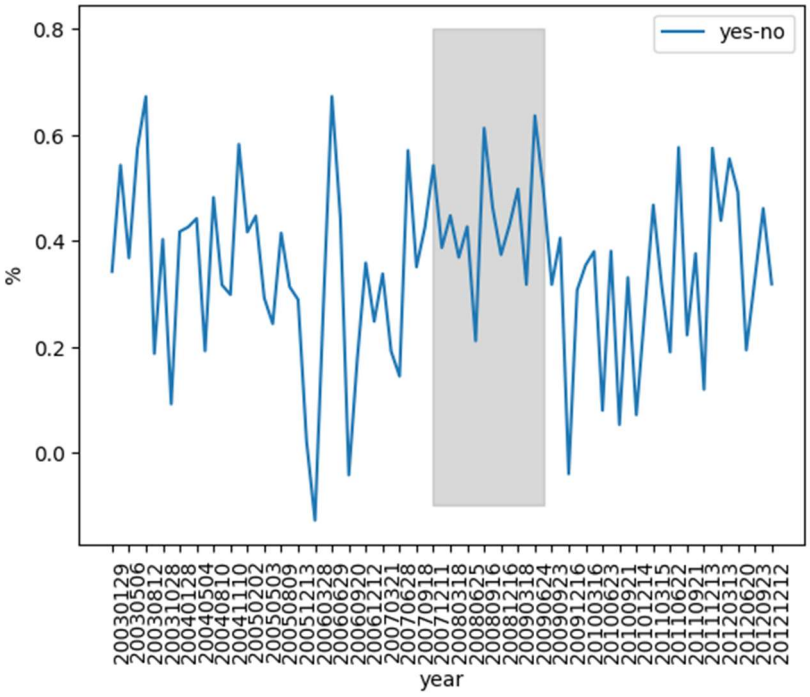
Figure 8: Categorization of economic conditions

A. Are you feeling sad?



8

B. Do you experience fear?



4. Concluding remarks

The use of unstructured text data constitutes a challenging task with, potentially, important rewards. In this work, we propose a novel methodology to extract information from text using generative language models. The methodology is applied to the case of FOMC deliberations. The information extraction approach involves, in a first stage, learning about linguistic patterns through LLM fine-tuning. In this way, an interactive representation is built. In a second step,

LLMs are prompted to generate pieces of text that are informative of policymakers' views at different points in time. Three tasks are implemented to evaluate the proposed approach. We find that time-stamped LLMs are able to generate text that is informative of policymakers' assessments of economic conditions. Also, the methodology is able to produce a synthetic audit of FOMC communication strategy. Finally, in a third task, fine-tuned LLMs generate pieces of text that illustrate features of policymaking narratives.

This work positions fine-tuned language models as latent representations of evaluations and the worldview of economic agents. There are several directions in which this work can be extended. First, there is space for evaluations of variations in the methodological specification of our exercise. These variations include the selected language model, the design of the training dataset and the parametrization of the text generation task. Also, we could consider alternative ways in which training text information is aggregated through time to generate fine-tuned models. In the current specification, each LLM is trained by a single document that correspond to a single FOMC meeting. Language models trained by a sequence of consecutive transcripts could allow for more efficient extraction of information.

Also, we believe that extending this exercise to other collections of documents corresponding to other economic agents can results in insightful descriptions of the evolution of economic perceptions and narratives.

Considering a bigger departure from the present exercise, that is, moving beyond the extraction of perceptions and opinions, this tool can be used to carry out computational simulations of economic behavior and aggregate dynamics. Fine-tuned LLMs could be used to simulate behavior that is consistent with produced text. Going one step further, we can envision exercises in which LLMs constitute an input in simulations of learning processes and decision making in dynamic interactive settings.

Bibliography:

Acosta, M. (2023). A new measure of central bank transparency and implications for the effectiveness of monetary policy. *International Journal of Central Banking*, *19*(3), 49-97.

Angeletos, G. M., & Jennifer, L. O. (2013). Sentiments. Econometrica: Journal of the Econometric Society, 81(2), 739-779.

Apel, M., Blix Grimaldi, M., & Hull, I. (2022). How much information do monetary policy committees disclose? Evidence from the FOMC's minutes and transcripts. *Journal of Money, Credit and Banking*, *54*(5), 1459-1490.

Bernanke, B. S. (2013). A century of US central banking: Goals, frameworks, accountability. Journal of Economic Perspectives, 27(4), 3-16.

Brunnermeier, M. K., & Pedersen, L. H. (2009). Market liquidity and funding liquidity. *The review of financial studies*, *22*(6), 2201-2238.

Christian, J. W. (1968). A further analysis of the objectives of American monetary policy. The Journal of Finance, 23(3), 465-477.

Cieslak, A., Hansen, S., McMahon, M., & Xiao, S. (2023). *Policymakers' Uncertainty* (No. w31849). National Bureau of Economic Research.

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient fine-tuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Fischer, E., McCaughrin, R., Prazad, S., & Vandergon, M. (2023). Fed Transparency and Policy Expectation Errors: A Text Analysis Approach. *FRB of New York Staff Report*, (1081).

Gorodnichenko, Y., Pham, T., & Talavera, O. (2023). The voice of monetary policy. *American Economic Review*, *113*(2), 548-584.

Hakes, D. R. (1990). The objectives and priorities of monetary policy under different Federal Reserve chairmen. Journal of Money, Credit and Banking, 22(3), 327-337.

Hansen, S., & McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics*, *99*, S114-S133.

Hansen, S., McMahon, M., & Tong, M. (2019). The long-run information effect of central bank communication. *Journal of Monetary Economics*, *108*, 185-202.

Heymann, D., & Pascuini, P. (2021). On the (In)consistency of RE modeling, Industrial and Corporate Change, Volume 30, Issue 2, Pages 347–356.

Heymann, D., & Sanguinetti, P. (1998). Business cycles from misperceived trends. *ECONOMIC NOTES-SIENA*, 205-232.

Lorenzoni, G. (2009). A theory of demand shocks. American economic review, 99(5), 2050-2084.

Lucca, D. O., & Trebbi, F. (2009). *Measuring central bank communication: an automated approach with application to FOMC statements* (No. w15367). National Bureau of Economic Research.

Malmendier, U., Nagel, S., & Yan, Z. (2021). The making of hawks and doves. *Journal of Monetary Economics*, *117*, 19-42.

Mullainathan, S. (2002). *Thinking through categories* (pp. 1031-1053). Working Paper, Harvard University.

Romer, C. D., & Romer, D. H. (2008). The FOMC versus the staff: where can monetary policymakers add value?. *American Economic Review*, *98*(2), 230-235.

Shapiro, A. H., & Wilson, D. J. (2022). Taking the fed at its word: A new approach to estimating central bank objectives using text analysis. *The Review of Economic Studies*, *89*(5), 2768-2805.

Sharpe, S. A., Sinha, N. R., & Hollrah, C. A. (2023). The power of narrative sentiment in economic forecasts. *International Journal of Forecasting*, *39*(3), 1097-1121.

Stein, J. C. (1989). Cheap talk and the Fed: A theory of imprecise policy announcements. *The American Economic Review*, 32-42.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Woodford, M. (2005). Central bank communication and policy effectiveness. NBER Working Paper Nr. .11898

Annex A: Trigger phrases used to characterize policymaking narrative

Goals:
"The main objective of our policies is to",
"The primary purpose of our policy decisions is to",
"When making policy decisions, our central goal is to",
"The central mission of our policy is to",
"Our main focus as policymakers is to achieve"

Perceived threats:
"The biggest threat for the economy is",
"The major economic worry is",
"In terms of the economy, the most relevant risk is",
"A looming danger to the economy is",
"The central economic vulnerability lies in"

Growth drivers:
“The main factor determining economic growth at this time is',
“Currently, economic activity is driven by”,
“The driving force behind economic activity right now is”,
“At present, the key determinant of economic growth at this moment is”,
“At this moment, the trajectory of GDP is primarily shaped by”

Inflation drivers:
“The main factor determining inflation at this time is”,
“Currently, inflation is driven by”,
“The driving force behind inflation right now is”,
“At present, the key determinant of inflation at this moment is”,
“At this moment, inflation is primarily shaped by”

Recession (similar for crisis):
"Do you believe we're experiencing an economic downturn?",
"Is there a possibility that we're in the midst of a recession?",
"Do you reckon the economy is going through a recessionary phase?",
"Is it your opinion that we might be facing a recession at the moment?",
"Do you think our economy is currently in a state of recession?",
"Are we in a recessionary period in your view?",
"Do you perceive any signs of a recession in our economy?",
"Is there any indication to you that we might be in a recession?",
"Is the economy undergoing a recession right now?",
"In your estimation, are we currently experiencing an economic downturn?",
"Are we currently in a recession?",
"Do you think we are in a recession?"

Fragile economy:
"Is the economy in a fragile position? I think that",
"Do you think the economy is weak? I believe that",
"Does the economy seem vulnerable to you? I feel that",
"In your opinion, is the economy in a delicate state? From my point of view,",

"Do you believe the economy is in a precarious condition? My stance is that"

Fear:
"Are you afraid? I feel",
"Do you feel fear? Personally, I ",
"Are you scared? I experience",
"Are you feeling frightened? In my case, I ",
"Are you feeling a sense of angst? I"

Sadness:
"Are you depressed? I feel",
"Do you feel sadness? Personally, I ",
"Are you down? I experience",
"Are you feeling downcast? In my case, I ",
"Are you feeling a sense of sorrow? I"

Annex B: Information content of indices that classify transcripts' sentences

For comparison purposes we evaluate the information content of indices that use NLI to classify sentences in FOMC transcripts. The classification of each sentence is equal to the difference of the probability assigned to positive sentiment and the probability assigned to negative sentiment. The estimated model is the same specification that was used in the analysis of the main text.

Table B.1: Information content of indices that classify sentiment in transcripts

	VIX	Stock market returns	Consumer sentiment	Press sentiment
$\hat{\beta}_{+1}$	-1.8426*	0.6921	0.2655***	0.2262***
	(1.003)	(0.660)	(0.058)	(0.083)
$\hat{\beta}_{+2}$	-3.3909	1.3046	0.2201**	0.1880*
	(2.229)	(1.349)	(0.085)	(0.105)
$\hat{\beta}_{+4}$	-3.0060*	1.9603	0.2622**	0.2167**
	(1.573)	(2.347)	(0.126)	(0.100)
$\hat{\beta}_{+8}$	-5.7761***	3.7062	0.2926*	0.2099
	(1.799)	(4.055)	(0.172)	(0.168)

Notes: The table reports standardized estimated coefficients for the sentiment index that results from inferring sentiment in transcripts sentences using NLI. WSJ sentiment is computed for economic headlines using NLI positivity and negativity scores. Heteroskedasticity-autocorrelation robust standard errors reported in parenthesis.