Synthetic surveys of monetary policymakers: perceptions, narratives and transparency[1]

Daniel Aromí
IIEP UBA-Conicet, FCE UCA

Daniel Heymann
IIEP UBA-Conicet

Abstract:

We propose a method to generate "synthetic surveys" that reveal 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 demonstrate the value of this tool in three different tasks: measurement of perceived economic conditions, evaluation of transparency in Central Bank communication and extraction 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 outputs show policymakers informational advantage. Anticipatory ability increases as models are prompted to discuss future scenarios and financial conditions. To analyze transparency, we compare the content of each FOMC meeting 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, LLMs produce narratives that show policymakers' views on their responsibilities and their understanding of main forces shaping macroeconomic dynamics.

JEP codes: E58, E47, D84, E71

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

[1] A previous version of this document was circulated under the title "Talk to Fed: a big into FOMC transcripts".

2009, Angeletos & Jennifer 2013, Heymann & Sanguinetti 1998). One set of economic actors for which this assertion is particularly relevant are 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 policymakers' statements, speeches and related documents. 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). In these exercises, 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 their capacity to interpret highly dimensional data and use that knowledge to build specific indicators. Hence, there are gains to be made using more expressive models that possess a greater capacity to acquire and transmit ideas expressed in economic discourse.

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. The 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 each monetary policy meeting. 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 highly likely responses as judged by each fine-tuned model. 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 LLM-generated texts to compute a numeric representation that summarizes these responses in a compact manner.

We test the proposed methodology implementing three different tasks: measurement of perceived economic conditions, 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 resulting indices suggest that fine-tuned LLMs are able to express policymakers' perceptions of economic conditions. Forecast models show these indices possess substantial information regarding macroeconomic and financial dynamics. For example, a one standard deviation increment in the index of perceived economic conditions anticipates, over the next two quarters, an increment of 0.3% in mean cumulative revisions of yearly growth forecasts.

The topic of Central Bank communication transparency has consistently captured 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 accurately by the corresponding minutes. Importantly, we do not find evidence suggesting optimism bias during the most acute period of the housing and financial crisis.

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. These frameworks take the form of stories with multiple features that establish relevant concepts and causal relationships between these concepts (Andre et al. 2023). In other words, models of the actors and their environment.

To learn about policymakers' narratives as manifested during closed door deliberations, LLMs are requested to generate multiple discussions. Starting with a focus on the job of policymakers, we prompt time-stamped LLMs to communicate policymaking objectives. We generate a complementary job description extracting information on policymakers' concerns or worries. Then, turning to their understanding of macroeconomic dynamics, we prompt completions that identify main drivers of economic growth and the principal forcer underlying inflation.

The extracted policymaking narratives indicate sharp contrasts. In terms of stated goals, we verify a balanced focus on two traditional policy objectives: price stability and economic growth. When it comes to manifested concerns, there is no balance. All through our sample period, economic activity constitutes the single major worry. This difference is suggestive of a high level of confidence regarding the ability to attain the price stability goal. Moving beyond aggregate demand, evaluating the evolution of other manifested worries, we find a set of significant although less prominent concerns. The manifested worries include inflation, housing, the financial system and the fiscal deficit. These unveiled patterns constitute new evidence on how policy makers respond to changing economic conditions.

We also find contrasts when we extract synthetic discussions regarding main macroeconomic forces. In the case of economic growth drivers, we find that policymakers' views are relatively stable with a focus on aggregate demand. When inflation drivers are identified, there are important fluctuations in the responses. In addition, in the case of inflation, supply shocks are the most frequently mentioned driver. Another noteworthy finding is that, during our sample period, economic policy is not frequently mentioned as a relevant force underlying inflation dynamics.

Our work makes contributions 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. Previous contributions have emphasived that macroeconomic dynamics are explained by features of economic agents' perceptions of their environment (Lorenzoni 2009, Angeletos and La'O 2013, Heymann, D., & Pascuini 2021). Relatedly, our analysis is linked to contributions that study economic narratives as important elements that frame economic evaluations and can explain collective behavior (Shiller 2019, Binetti et al. 2024, Andre et al. 2023). Contributing to this agenda, our work develops new strategies for the characterization of perceptions and economic narratives. We test these tools and report novel evidence regarding macroeconomic perceptions and narratives.

Taking a more focused perspective, 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) implement two dictionary methods and three LLM techniques to measure sentiment in FOMC transcripts. Compared to these

previous contributions, in the current work we move beyond text classification and develop a generative methodology to extract information from FOMC meetings. The satisfactory performance observed in three different tasks suggests the fitness of the proposed strategy.

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.

The document is organized as follows. In the next section we present the data and provide a detailed description of the methodology. Section 3 reports the exercises in which policymakers' perceptions of economic conditions are extracted using the generative methodology. Section 4 presents the evaluation of communication transparency. Extracted narratives are reported in section 5. Concluding remarks are provided in section 6.

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 acquire and transmit information about linguistic regularities. In the next subsection, we describe the data and the construction of the training dataset. Next, we provide a detailed description of the natural language methodology.

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 construct an 80-document corpus collecting the FOMC transcripts corresponding to the period 2003-2012. 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, each 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 sentences with more than 10 words. In the third stage, from each sentence in this subset, we build two training examples that differ in terms of the selected surrounding text. The first type of training example has an input that consists of the first half of the sentence and is matched to an output consisting of the second half of the sentence plus the following two sentences. The second type of 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 variation in the design of

examples is proposed conjecturing that, in this way, the fine-tuned model will acquire more information from the corresponding FOMC transcript. In this way, for each document, the number of training examples is two times the number of sentences with more than 10 words.

2.2 Methodology

LLMs are widely used tools 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 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 meetingin 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 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…”. Given the stochastic nature of this models, to obtain more information, we sample multiple outputs per trigger phrase.

In the last step, outputs are processed to obtain a numerical representation of 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 aspects of interest and for alternative sample 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 surveys. Before proceeding to a detailed description of each step, it is worth noting that the first stage of the pipeline is fixed, that is, the fine-tuning stage is performed once per meeting, and the resulting fine-tuned models are used for all subsequent tasks. In contrast, the other two stages are performed multiple times with adjustments according to the different use cases. The three steps of our proposed methodology 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 produce an output in the form of sequences of tokens that are likely to complete the phrase. 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 download the model from Huggingface portal.[2] 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 or 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 the token's 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".[3] Through quantization, floating points are represented via less memory demanding low precision formats. Complementarily, fine-tuning 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, following standard practice, LoRA dimensionality and "r" parameters were both set equal to 16.

2.2.2 Text generation:

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 produce semi-structured data.

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 of 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 per trigger phrase 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. We generate multiple outputs in which each token of completion is the result of a lottery determined by probabilities that are recursively computed by the model.[4]

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 other phrases with similar meaning such as "Currently, the state of the economy is marked by

[2] meta-llama/Llama-2-7b-chat-hf.
[3] https://huggingface.co/docs/peft/index.
[4] This is achieved setting "pipeline" function argument "do_sample" to "True" and choosing the number of sampled outputs through argument "num_return_sequences".

…" 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 methodological feature is that we observe instances in which model responses seem to be very sensitive to specific wordings. Given our interest in extracting some general properties regarding economic perceptions and narratives, this can be interpreted as a form of overfitting. The use of multiple phrases is proposed to mitigate the impact of this undesirable property sometimes present in fine-tuned models.

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. For example, we find that adjectives such as "favorable" and "positive" are generated with high probability. In contrast, the LLM trained with the transcript corresponding to March 2008, the during the financial crisis, produces text suggesting deteriorating perceptions. In this case, words such as "worsening", "poor" and "uncertain" are generated with high probability.

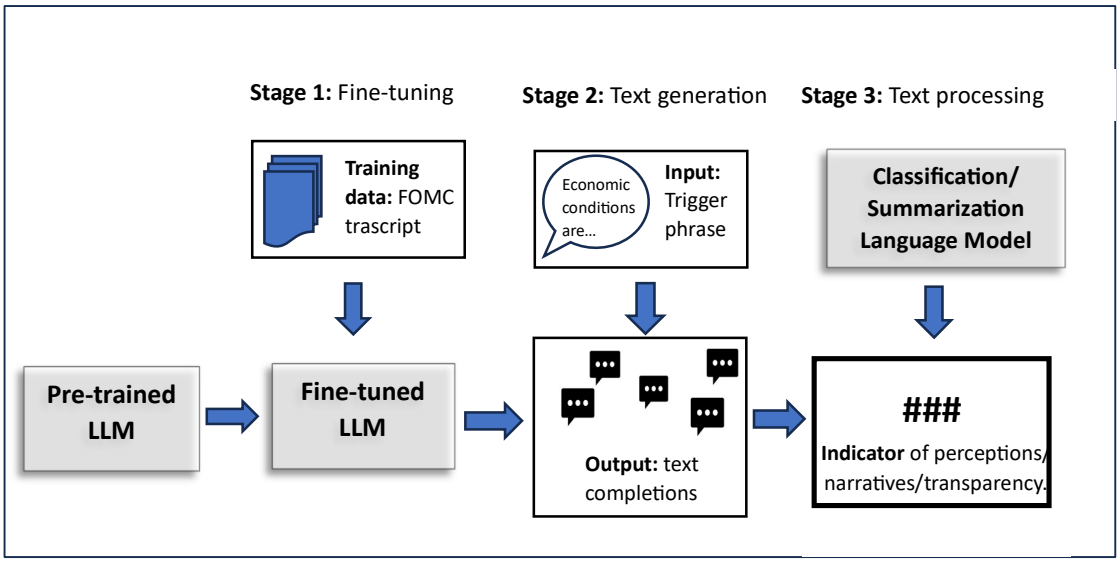
2.2.3 Text processing

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 allows for a compressed representation of evolving policymakers' perspectives 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 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 another relevant feature. To implement this task, we use a fine-tuned version of model "bart-large-mnli". The model was trained via a multi-genre dataset.[5]

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

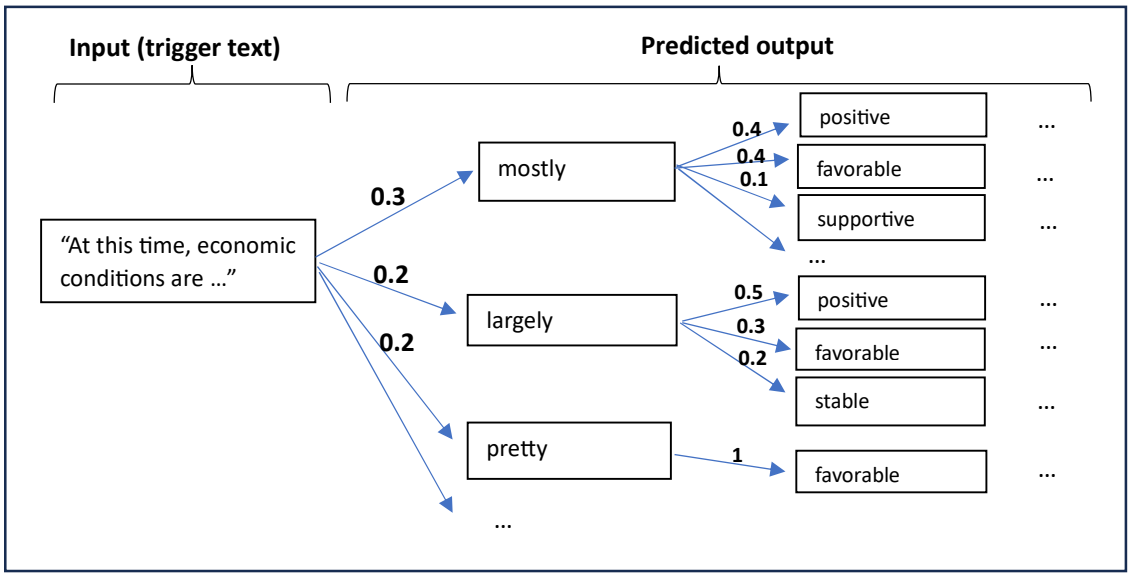
Figure 1: Workflow for the generation of synthetic text



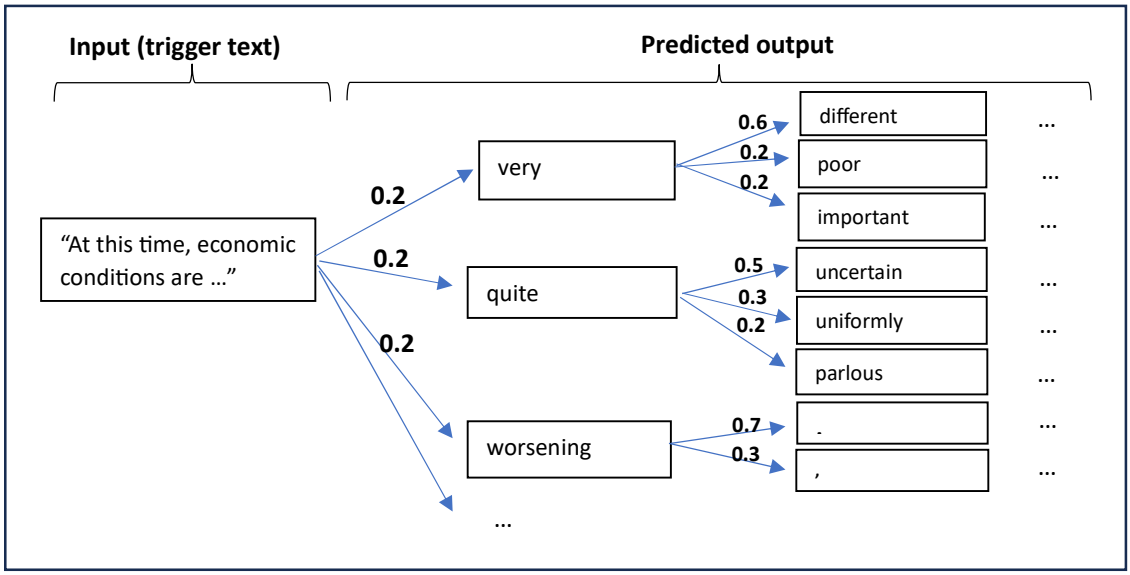
Note: This is a sketch of the pipeline that consists of three steps: fine-tuning, text generation and text processing.

Figure 2: Sample generated texts

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



B. 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 to 25 samples produced using transformers library function "pipeline" and setting the argument "do_sample" to True.

3. 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 is applied to the evaluation of alternative 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 texts from different documents produced by policymakers are 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 through a combination rich text datasets and powerful natural language techniques.

With adjustments according to the specific use case, we apply the methodology detailed in the previous section. To generate text completions that are informative of policymakers' views about the economy, we build eight trigger phrases with similar meaning. As previously discussed, with this feature our objective is to reduce the inaccuracies that might result from fine-tuned LLMs that respond with noise to specific wordings. 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 multiple times. In terms of choosing the length of the completion, we set the maximum number of generated tokens to 25.

To evaluate the proposed methodology in a more exhaustive manner, we consider multiple specifications of the index through alternative lists of trigger phrases. Trigger phrases are changed considering different length (short vs. long), alternative time orientation (present vs. future) and variations in the focus (economic vs. financial). For the short prompts alternative that focus on present economic conditions, we select 8 phrases similar to "Economic conditions are".[6] Each trigger phrase is sampled only 25 times. Under the long prompt alternative, we build 64 prompts, that result from the combination 8 different references to the present time and previously listed 8 alternative references to economic conditions. The set of time orientation 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, ". In this case, given the large set of trigger phrases, each trigger phrase is sampled only 10 times.

We extend the analysis considering two variations of the index of policymakers' perceptions. 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 verb tenses and, in the case of long prompt, we also adjust the time references.[7] Finally, considering the key role played by financial conditions during the sample period, we consider alternative in

[6] The complete list of phrases is: "Economic conditions are", "Economic indicators show", "The economic climate is", "The economic environment points to", "The economy shows", "The economic landscape hints at", "The overall economic scenario shows" and "The state of the economy is"

[7] In the case of long prompts, the group of phrases indicating time orientation is: '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.

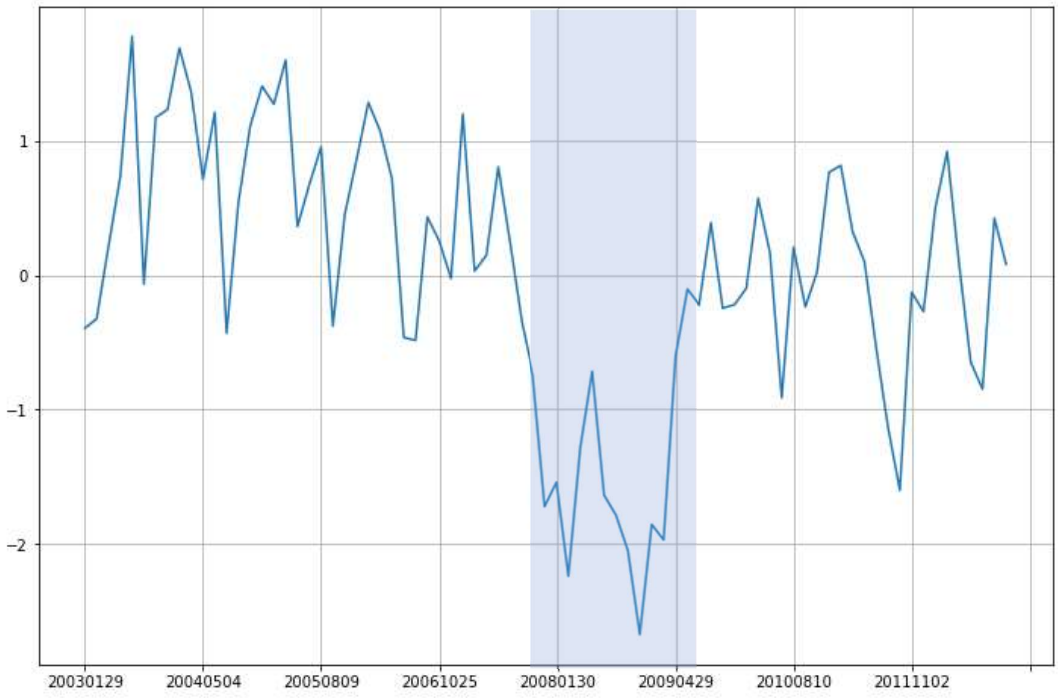
which trigger phrases refers to future financial conditions.[8] In this way, we obtain six alternative specifications for the metric of perceived conditions.

As previously indicated, 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. Finally, the classifications are averaged to produce the respective indicator of perceived economic conditions on a given date.

As a first illustrations of the resulting metrics, Figure 3 shows the average sentiment indicator that is computed after calculating z-scores for the six specifications of the perceived conditions metric. 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 (2011). Moving beyond this descriptive analysis, the information content of the indicators is evaluated formally through forecast exercises detailed below.

[8] For this specification of the index, the 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'.

Figure 3: Index of policymakers' perceptions of economic conditions



Notes: The figure shows the standardized metric of sentiment computed that results from combining the six metrics that were generated using different collections of trigger phrases. Shaded area indicates the Great Recession.

To evaluate the information content of the synthetic sentiment index we assess its ability to anticipate revisions in GDP growth forecasts. We consider the Survey of Professional Forecasters. These projections correspond to quarterly private sector forecasts collected by the Philadelphia Federal Reserve. The 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 $\hat{g}^{q'}_{q,q+4}$ 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: $Rev^{q}_{q+h} = \hat{g}^{q+h}_{q,q+4} - \hat{g}^{q}_{q,q+4}$. 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:

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

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.

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 magnitude of the coefficients increases with the time horizon. This pattern indicates that the information content of the index is not concentrated in the immediate future but is spread over multiple future periods. A comparison of the estimated coefficients suggests that short prompts result in more informative indices. In other words, the addition of time orientation phrases to short phrases seem to introduce noise. From the perspective of extracting economic perceptions, longer phrases might generate overfit. Also, showing LLMs' ability to respond to relatively subtle changes in the orientation of the prompts, the estimated coefficients show that the informativeness of the indices increase as prompts point towards forward looking assessments. Also, a switch from a focus on economic conditions to a focus on financial conditions results in further improvements. The estimated anticipatory ability is economically significant. For example, in the case of short prompts and future financial conditions, an increment of one standard deviation in the index is followed by an average increment of 0.75% in cumulative revisions over the following 3 quarters. To gain a better understanding of this result, we observe that during an important part of 2008 the corresponding index was two standard deviations below zero. Then, according to the estimated forecast model, this implies a cumulative downward revision of 1.5% in yearly GDP growth over the next 3 quarters.

Table 1: Forecasting revisions in growth forecasts – Survey of Professional Forecasters

A. Short prompts

	Target evaluated conditions		
	Current econ.	Future econ.	Future financial
$\hat{\beta}_{+1}$	0.1978*	0.3298***	0.4245***
	(0.116)	(0.118)	(0.128)
$\hat{\beta}_{+2}$	0.3165	0.5053**	0.6646***
	(0.241)	(0.240)	(0.203)
$\hat{\beta}_{+3}$	0.4259	0.6604*	0.7504***
	(0.325)	(0.339)	(0.255)

B. Long prompts

	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	0.3998*	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.

4. 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.

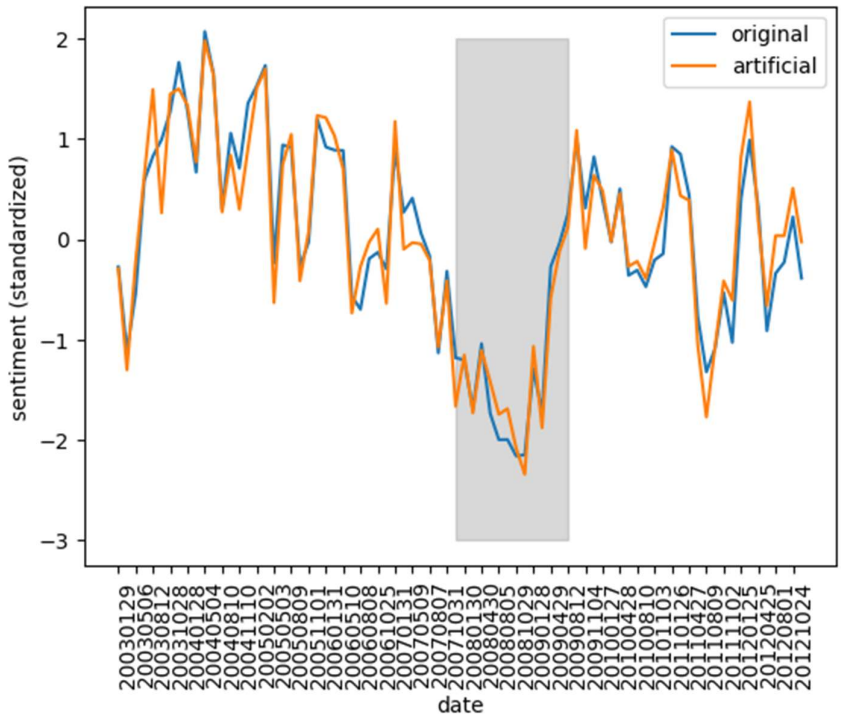
Despite its importance, communication by monetary policymakers is incompletely understood. While there are benefits from building a reputation for truthful telling (Woodford 2005), heterogeneous objective functions might prevent full transparency (Stein 1989). There are no warranties of transparency or truthfulness. This is particularly true in the context of financial crises. In these extreme scenarios, policymakers might have incentives to send reassuring but biased messages. These messages can mitigate 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. In these synthetic minutes we substitute content of FOMC communications by content that mirrors closed doors discussions captured in FOMC transcripts. 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, we can assume that the topic is, to a large extent, set by the first half of the sentences. Hence, our test is focuses on the tone of the message. This is an important feature of communication that can affect public perceptions and aggregate outcomes. Hence, 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. We find very similar trajectories for the indicators. This feature is a first indicator that suggests FOMC minutes transmit a tone that is consistent with the tone of the meeting. In particular, it is worth observing that there is no indication of biased communication during the most severe stages of the financial crisis.

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



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

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 element in the disparity between the tone found in minutes and the tone transmitted by FOMC transcripts. This systematic component might be linked to changing incentives that lead to biases in Central Bank communication. We consider three variables indicating economic conditions: expected economic growth over the next quarters according to Board staff (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 a formal test of truthfulness in Central Bank communication.

Table 2: 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 different economic indicators ($\hat{\beta}$). Heteroskedasticity-autocorrelation robust standard errors reported in parenthesis.

5. Policymaking narratives

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 representations and evaluations that guide decision-making. Narratives can indicate the objectives of the decision maker and a causal representation of the environment (Andre et al. 2023). Our understanding of monetary policy and, more broadly, macroeconomic dynamics can benefit from a more precise and systematic documentation of narratives. Previous contributions have advanced the understanding of macroeconomic narratives and their implications using tabular household surveys (Andre et al. 2023), classifying open ended household and expert surveys (Binetti et al. 2024) and processing corporate meetings' transcripts (Flynn & Sastry 2024). In this section, we demonstrate fine-tuned LLMs can be used to characterize the evolution of policymaking narratives. In this way, we move beyond the focus of the previous sections that was placed on polarity and proceed to identify features of policymakers' mental models. We conjecture that our generative methodology is particularly well-adapted for this task.

In the analysis below we focus on two aspects of policymaking narratives that we consider of particular interest. First, we extract policymakers' views of their responsibilities. We characterize these duties, prompting finetuned models to identify two connected concepts: objectives and concerns. These elements are tightly linked since, in any economic narrative, listed concerns are to an important extent determined by objectives and, in turn, stated goals are framed in the context of perceived concerns. By prompting models to identify these two elements, we intend to produce an informative characterization of what can be termed as a job description.

Second, shifting towards policymakers' understanding of macroeconomic dynamics, we request models to complete phrases that identify the main drivers of economic growth and, relatedly, the main forces behind inflation. In this way, we extract key elements of simplified causal models that underlie policymakers' expectations and decision making. In extended analysis, reported in Appendix C, we also show how finetuned models generate classifications of economic conditions in terms of traditional economic labels (recession, crisis, fragility) and measure emotions in narratives.

5.1 Policymakers' responsibilities: objectives and concerns

Monetary policy objectives have been the focus of attention in academic and policy circles both from a normative and positive perspective (Bernanke 2013, Hakes 1990, Christian 1968). The objectives guiding policy decisions are a latent feature that can be inferred from public

decisions and statements. In the analysis below, we contribute a novel metric of policymakers' goals via the proposed generative methodology. Additionally, to construct a more complete description of policymakers' responsibilities, we prompt models to identify economic concerns. These concerns, or threats, are likely to represent the focus of attention at a given point in time and, in this way, reveal complementary evidence about how policymakers perceive their responsibilities.

To extract information regarding policymakers' goals and concerns that are consistent with FOMC deliberations we adapt the general methodology described in section 2 to this specific use case. For objectives, we design 5 different prompts, or trigger phrases, with meaning similar to: "Our main focus as policymakers is to achieve…".[9] Similarly, 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…". 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 selected labels. Mean probabilities are interpreted as the frequency with which a goal or concern is mentioned in a synthetic survey. Since the classification mode is multi-label, these probabilities do not necessarily add up to one. The labels were selected according to expert judgement after inspecting a subset of generated phrases. In the case of goals, the list of candidate labels is: "price stability", "economic growth" and "unemployment". In the case of concerns, the list of candidate labels is: "economic activity", "inflation", "financial system", "housing" and "fiscal deficit".

The results are reported in figures 5 and 6. When we consider average values for manifested goals over 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. Manifesting a change in discourse that accompanied changing economic conditions, 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 combined results are suggestive of a shift in the prominence of different objectives that might be explained by the weak recovery from the Big Recession.

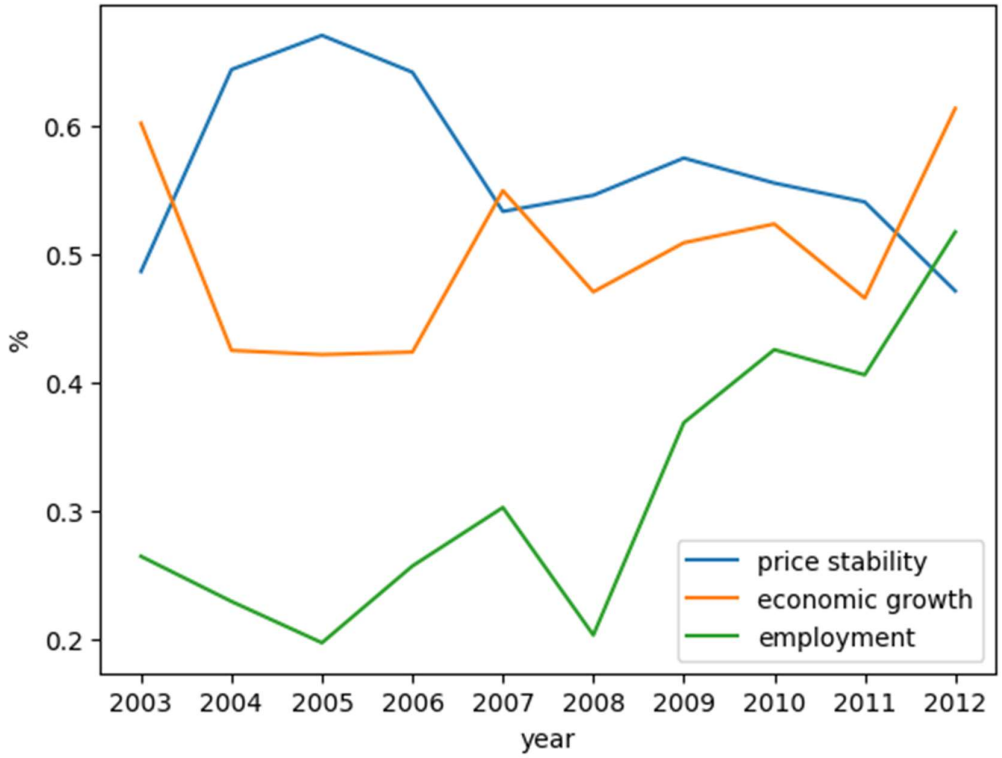
When we consider manifested concerns, we verify an important contrast with the balanced focus on two objectives observed in the case of monetary policy goals. In terms of worries, there is a single topic that clearly emerges as the main source of concern. By far, the most frequently mentioned concern 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 6 shows the frequency of mentions by year. The central position of economic activity is a persistent feature during our sample period. The frequency of mentions of economic activity is above 40% in all years of the sample. Demonstrating the shifting dynamics in the economic

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

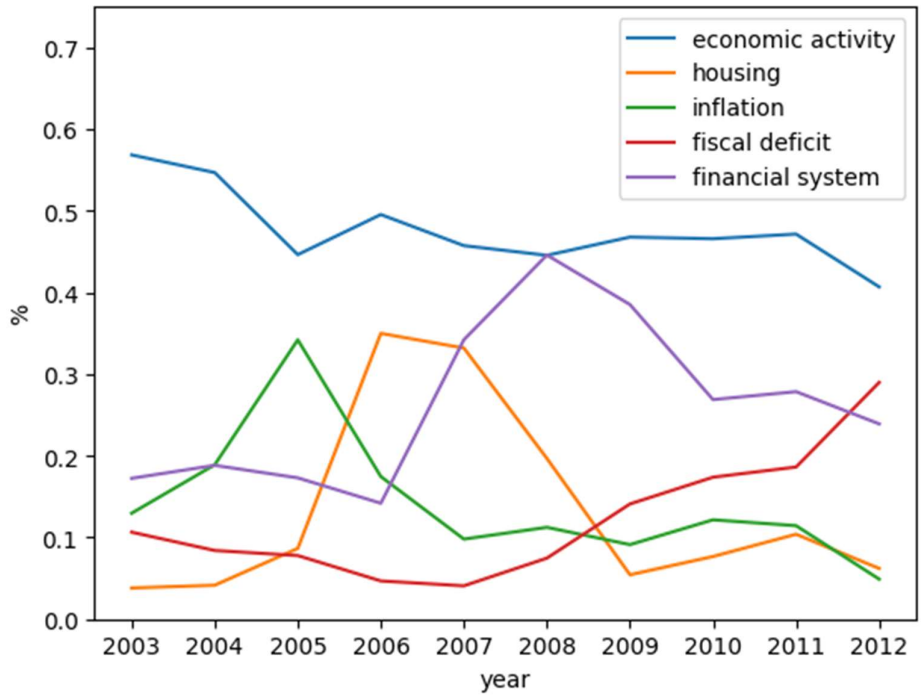
scenario and associated policy challenges, when we consider other relevant concerns, we observe a sequence of spikes. Taking turns, these spikes position one of those topics as the second most important threat. First, we detect worries about inflation that peak in 2005. Next, housing emerges as the second 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 5: Monetary Policy Objectives



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

Figure 6: Source of policy concerns



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

5.2 Macroeconomic drivers

Macroeconomic dynamics are complex. They are determined by multiple interacting forces. Making a precise understanding particularly challenging, these dynamics are shaped by economic agents' anticipation of how these drivers will influence the trajectories in the future. In this 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 drivers in narratives can be useful to describe policymakers' changing views and rationalize monetary policy decisions.

To extract this feature of 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 3 shows the frequency with which alternative 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, a frequently referred driver is economic policy (31%). LLMs generated texts suggest that financial markets, with a frequency of 19%, is perceived as a prominent driver of economic growth.

When we evaluate the frequency of mentions by year we observe, for the most part, a stable scenario. Figure 7 shows the evolution for five representative labels.[10] 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 not spike in the highly uncertain context of the 2008/09 financial crisis.

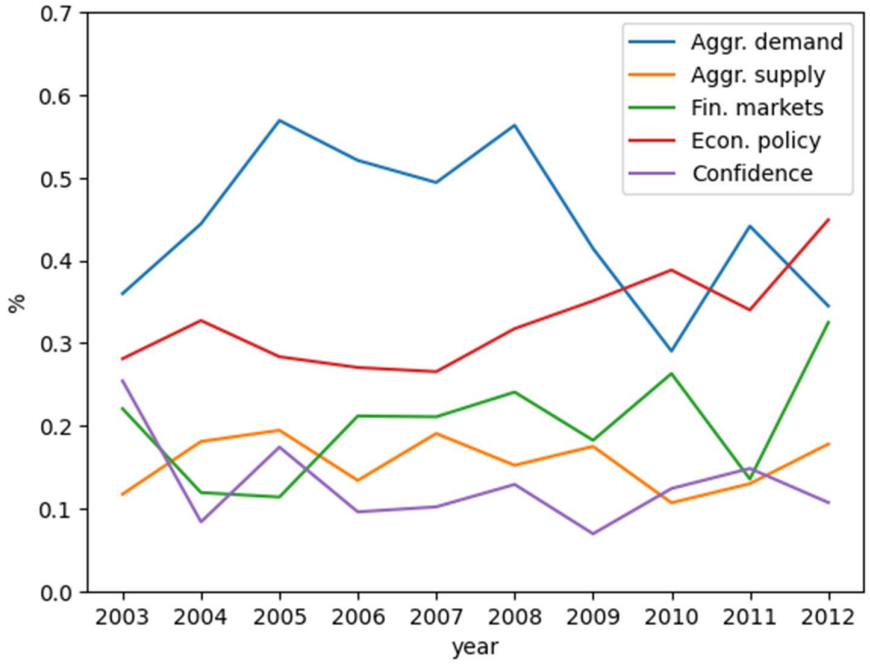
[10] To facilitate the visualization, we choose to eliminate the line corresponding to consumption. This is a component of aggregate demand with a line displaying a very similar trajectory

Table 3: Main growth drivers – Frequency of mentions

Drivers	Frequency
Broad categories	
Aggregate demand	0.456
Aggregate supply	0.154
Demand components	
Consumption	0.340
Investment	0.086
Government expenditures	0.061
Inventories	0.046
Net exports	0.026
Taxes	0.013
Policy	
Economic policy	0.309
Monetary policy	0.097
Fiscal policy	0.097
Miscellaneuos	
Financial markets	0.186
Confidence	0.126
Job creation	0.079
Productivity	0.070
Oil prices	0.041

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 7: 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 4, oil prices is the most frequently mentioned driver with an average frequency of 31%. In this case, the most frequently mentioned driver is a supply side factor. 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 8 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 the 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 the 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 like what we observe for the trajectory of references to past prices.

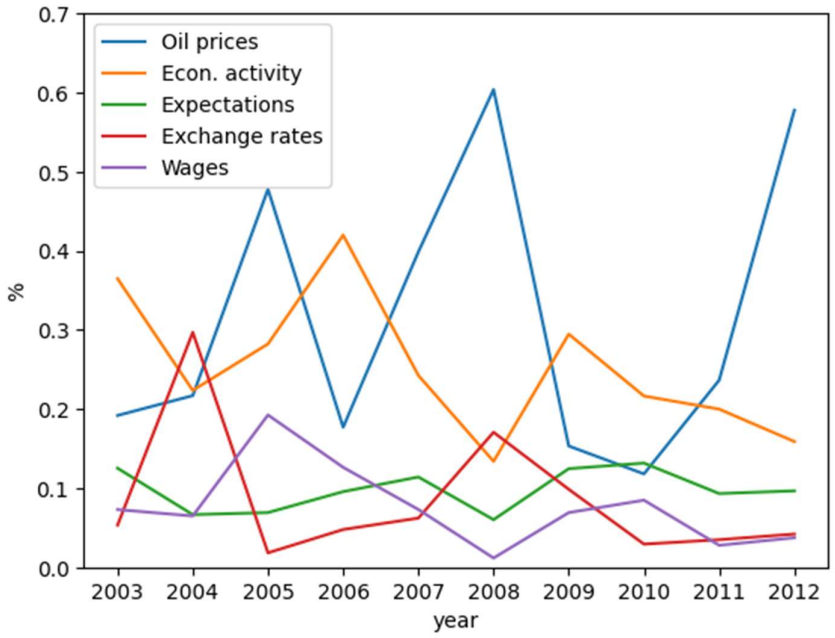
For inflation drivers, our findings can be compared to surveys that have asked about the causes of inflation. Binetti et al. (2024) report that, in a recent tabular survey, households mention government and the federal reserve with a significantly higher frequency than in our synthetic surveys. In contrast, through open-ended surveys, Andre et al. (2023) find households assign more prominence to supply factors and the labor market. When they ask experts, there is a balanced focus on supply factors and economic policy. Finally, in Aromi et al. (2024), a synthetic business survey is carried out implementing a methodology similar to the one proposed in this work. The results show commodity prices, followed by wages, have been the most visible cause of inflation. By the end of their sample period, years 2021 through 2023, supply disruptions and demand also emerge as important drivers. These contrasts are suggestive of differences across economic agents. In particular, compared to households, policymakers and price setters seem to assign less importance to economic policy in their discussions of inflation. Further insights can result from more detailed analyses with matching dates and methodologies.

Table 4: Main inflation drivers – Frequency of mentions

Drivers	Frequency
Cost factors	
Oil prices	0.310
Wages	0.075
Business cycle	
Economic activity	0.252
Aggregate demand	0.111
Policy	
Economic policy	0.120
Monetary policy	0.067
Fiscal deficit	0.012
Miscellameous	
Expectations	0.098
Past inflation	0.091
Exchange rate	0.085
Interest rates	0.041

Notes: Average frequency of mentions of 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 8: Main inflation drivers – Frequency of mentions by year



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

6. Concluding remarks

The use of unstructured text data constitutes a challenging task with 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 produces a synthetic audit of FOMC communication strategy. Finally, in a third task, fine-tuned LLMs generate pieces of text that extracts features of policymaking narratives.

This work positions fine-tuned language models as versatile 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. In the current specification, each LLM is trained by a single document that corresponds to an FOMC meeting. To produce an efficient aggregation of text information corresponding to different points in time, future work can train language models in a sequential manner. Also, we believe that extending this exercise to other collections of documents corresponding to other economic agents can result 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 has potential in the computational simulations of economic behavior and aggregate dynamics. Fine-tuned LLMs could be used to simulate behavior that is consistent with produced text. Under this approach, LLMs can constitute inputs 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.

Andre, P., Haaland, I., Roth, C., & Wohlfart, J. (2023). Narratives about the Macroeconomy. CESifo Working Paper No. 10535.

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.

Aromi, D., Bonel, M.P. & Llada, M. (2024) Listening to the price-setters: Inferring inflation expectations from synthetic surveys, IIEP UBA-Conicet, mimeo.

Aromi, J. D., & Clements, A. (2021). Facial expressions and the business cycle. Economic Modelling, 102, 105563.

Aromi, D. & Heymann, D. (2023) Algorithmic Examination of Monetary Policy Deliberations: a Study on Foresight and Transparency. IIEP UBA-Conicet, mimeo.

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

Binetti, A., Nuzzi, F., & Stantcheva, S. (2024). *People's Understanding of Inflation* (No. w32497). National Bureau of Economic Research.

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).

Flynn, J. P., & Sastry, K. (2024). The macroeconomics of narratives. *Available at SSRN 4140751*.

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

Appendix A: Trigger phrases for policymaking narratives characterization

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"

Appendix B: Extended analysis of information content

B.1 Financial markets and sentiment

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 tables B.1 and B.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 future trajectory economic and financial variables. 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 B.1: Information content of synthetic sentiment indices – Short prompts

	VIX	Stock market returns	Consumer sentiment	Press sentiment
A. Current economic conditions				
$\hat{\beta}_{+1}$	-1.5339	0.6210	0.1974***	0.1535*
	(1.048)	(0.697)	(0.066)	(0.089)
$\hat{\beta}_{+2}$	-3.0345	1.0029	0.1334	0.1896
	(1.901)	(1.323)	(0.113)	(0.116)
$\hat{\beta}_{+4}$	-2.7396	0.6779	0.1096	0.1284
	(2.238)	(2.742)	(0.147)	(0.162)
B. Future economic conditions				
$\hat{\beta}_{+1}$	-2.0157	0.7391	0.2348***	0.2181**
	(1.344)	(0.752)	(0.059)	(0.103)
$\hat{\beta}_{+2}$	-3.8487	1.3844	0.2155**	0.1592
	(2.489)	(1.458)	(0.090)	(0.114)
$\hat{\beta}_{+4}$	-3.1628	1.3490	0.1332	0.1526
	(2.105)	(2.739)	(0.130)	(0.147)
C. Future financial conditions				
$\hat{\beta}_{+1}$	-1.9191*	1.2806*	0.1894***	0.1749*
	(1.090)	(0.724)	(0.049)	(0.100)
$\hat{\beta}_{+2}$	-3.7573*	2.7035**	0.2521***	0.1045
	(2.111)	(1.300)	(0.051)	(0.132)
$\hat{\beta}_{+4}$	-1.5100	3.1262	0.2476**	0.0989
	(1.031)	(2.056)	(0.117)	(0.117)

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.

Table B.2: Information content of synthetic sentiment indices – Long prompts

	VIX	Stock market returns	Consumer sentiment	Press sentiment
D. 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)
E. 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)
F. 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.

B.2 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.3: 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.

Appendix C: Further characterizations of policymaking narratives

C.1 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 C.1 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 C.1, 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.[11] 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 C.1: Classification of economic conditions

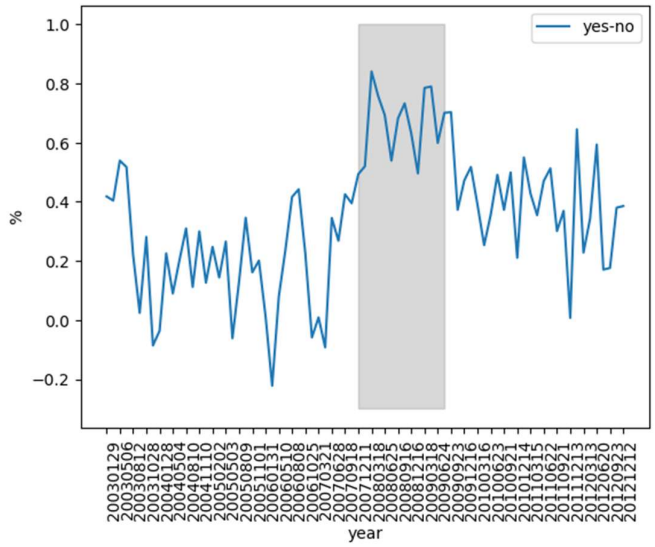
	Category		
Period	Recession	Crisis	Fragility
Pre-recession (before December 2007)	0.199	0.187	0.294
Recession (December 2007- June 2009)	0.658	0.653	0.690
Post-recession (after June 2009)	0.402	0.393	0.507

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.

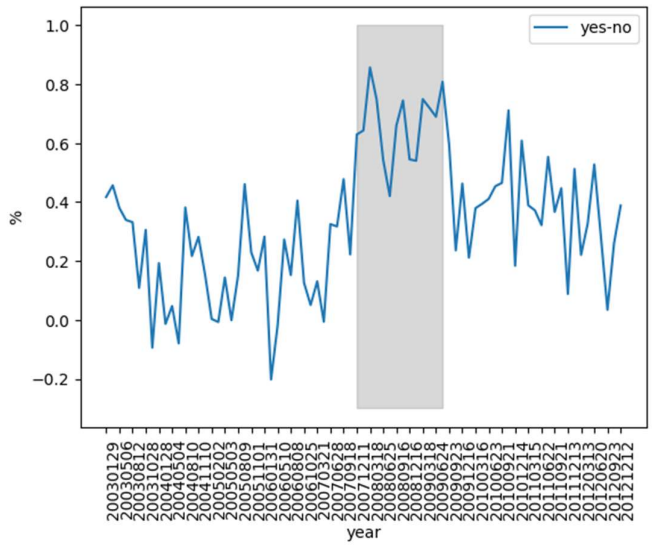
[11] The date of the announcements can be found in: https://www.nber.org/research/business-cycle-dating/business-cycle-dating-committee-announcements

Figure C.1: Categorization of economic conditions

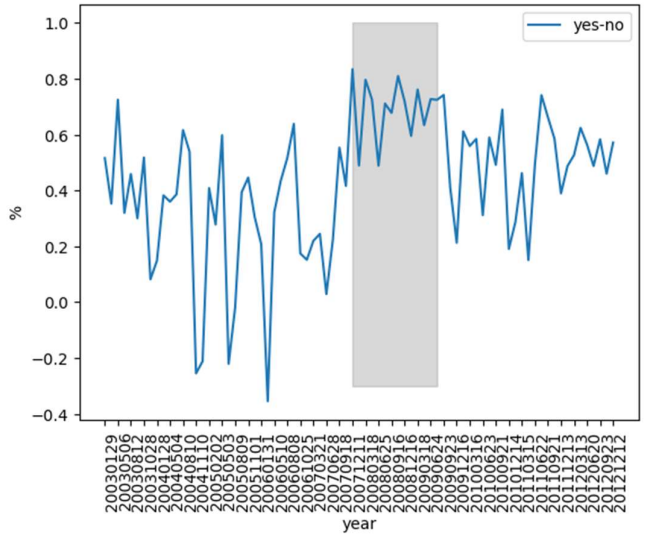
A. Is the economy in a recession?



B. Is the economy in crisis?



C. Is the economy fragile?



C.2 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 C.2 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 C.2 shows 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.

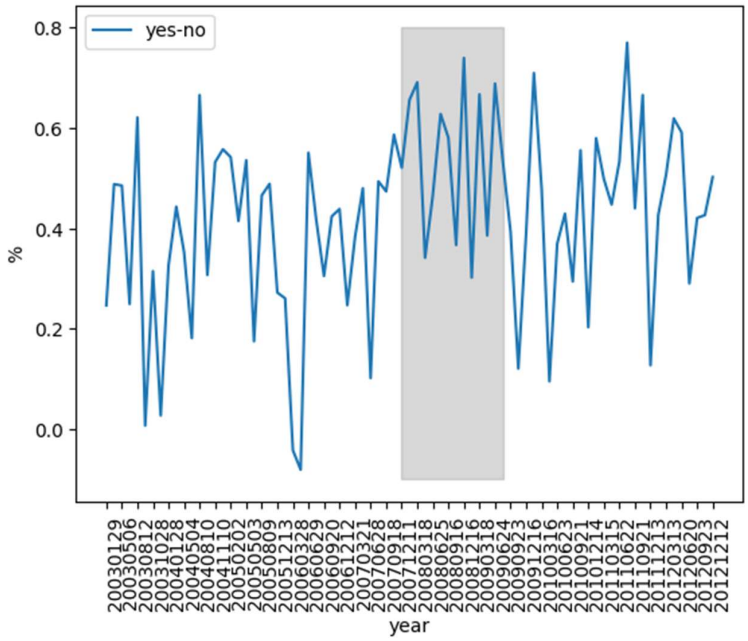
Table C.2: Manifestations of emotions

	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 C.2: Categorization of economic conditions

A. Are you feeling sad?



8

B. Do you experience fear?