Central Bank Mandates and Monetary Policy Stances: through the Lens of Federal Reserve Speeches*

Christoph Bertsch[1], Isaiah Hull[2], Robin L. Lumsdaine[†3], and Xin Zhang[1,4]

[1]Sveriges Riksbank
[2]BI Norwegian Business School; CogniFrame
[3]Kogod School of Business, American University; Erasmus University Rotterdam; National Bureau of Economic Research (NBER); Tinbergen Institute; Center for Financial Stability
[4]BIS Innovation Hub Nordic Centre

May 2023

Abstract

The Federal Reserve has an institutional mandate to pursue price stability and maximum sustainable employment; however, it remains unclear whether it can also pursue secondary objectives, such as financial stability, economic equality, or climate risk mitigation. The academic literature has largely argued that it should not. We characterize the Fed's interpretation of its mandate using state-of-the-art methods from natural language processing, including a collection of large language models (LLMs) that we modify for enhanced performance on central bank texts. We apply these methods and models to a comprehensive corpus of Fed speeches delivered between 1960 and 2021. We find that the Fed perceives financial stability to be the most important policy concern that is not directly enumerated in its mandate, especially in times when the debt-to-GDP ratio is high, but does not generally treat it as a separate policy objective. From a policy perspective, it has, in fact, frequently discussed the use of monetary policy to achieve financial stability and this discussion appears to have consequences. In particular, its discussion of both financial stability and financial crises predicts both monetary policy decisions and movements in asset prices, even after rigorously controlling for macroeconomic and financial variables.

Keywords: Natural Language Processing, Machine Learning, Central Bank Communication, Financial Stability.

JEL classification: C55, E42, E5, E61, G28.

*We are grateful for comments from Francesco Bianchi, Ricardo Correa, Michael Ehrmann, Juri Marcucci, Tho Pham, and seminar participants at George Washington University, Linköping University, Sveriges Riksbank, Universidad Carlos III, AI Sweden: NLP Seminar Series, the Applied Machine Learning, Economics and Data Science (AMLEDS) seminar, the Transdisciplinary Econometric & Data Science seminar, the 2022 Society for Financial Econometrics (SoFiE) Annual Meeting, the 2022 CARMA Conference, the 2022 Open Data Science Conference, the 2022 Data Science Global Summit, and the 2023 Royal Economic Society Annual Meeting. Nicole Brynjolffson provided excellent assistance in collecting and organizing some of the speeches used in this paper as part of the Center for Financial Stability's Summer Internship Program. This work was conducted while Lumsdaine was on partial secondment to the Office of Financial Research, US Department of the Treasury, under an Intergovernmental Personnel Agreement (IPA). We thank Edvin Ahlander for excellent research assistance. The opinions expressed in this article are the sole responsibility of the authors and should not be interpreted as reflecting the views of the BIS, Sveriges Riksbank, the US Department of the Treasury, or any of the other institutions they represent. The authors have no relevant conflicts of interest to declare. All remaining errors are our own.

†Corresponding author. Email: Robin.Lumsdaine@american.edu. Tel: (202) 885-1964. Kogod School of Business, 4400 Massachusetts Avenue, NW Washington, DC 20016.

A central bank's mandate is a legal directive from the government that specifies its responsibilities as an institution. Nearly all central banks have a mandate that contains an explicit reference to price stability (BIS, 2009). The next most common objective is financial stability, followed by various requirements to target economic growth, employment, or welfare. The standard interpretation of the Federal Reserve's "dual mandate" is that it requires the pursuit of both price stability and maximum sustainable employment.

While there is broad agreement about which objectives are specified in the Fed's mandate, disagreement remains over how it should be executed and what, if anything, it prohibits. In this paper, we focus on an important source of disagreement: namely, can and should the Fed pursue objectives that are not explicitly enumerated in its dual mandate, such as financial stability, economic equality, or climate risk mitigation? Setting the legal question aside, the academic literature has largely concluded that the Fed should not, at least not with respect to financial stability (Vollmer, 2021; Schularick et al., 2021). We attempt to characterize the Fed's interpretation of its mandate over its institutional history, and examine the implications for monetary policy and the impact on financial markets.

In contrast to academics, Federal Reserve officials appear to be less convinced that pursuing secondary objectives, such as financial stability, is in conflict with the dual mandate. Eric Rosengren, former President of the Federal Reserve Bank of Boston, for example, has indicated on several occasions that financial stability does and should influence monetary policy. In a 2016 speech during his tenure on the FOMC, Rosengren claimed that: "...financial stability concerns could be a consideration in how long policymakers wait before resuming the gradual removal of monetary accommodation."

Other members of the FOMC have argued for the pursuit of financial stability more explicitly during their tenure. For example, Lael Brainard, in a 2014 speech, claimed that monetary policy should constitute a "second line of defense" against financial instability and suggested that the Fed would be faster to use monetary policy to pursue financial stability than some other central banks because "... [the Fed's] regulatory perimeter is narrower, the capital markets are more important, and the macroprudential toolkit is not as extensive."[1]

Although Brainard argues that dual mandate and financial stabilization concerns typically coincide, she also concedes that the two may conflict under certain circumstances. A clear example of this was the failures of Silicon Valley Bank and Signature Bank, during a period of high inflation in the Spring of 2023. In response to these adverse developments in the financial sector, Austan D. Goolsbee, President of the Federal Reserve Bank of Chicago, acknowledged that "[in] moments like this, of financial stress, the right monetary approach calls for prudence and patience–for assessing the potential impact of financial stress on the real economy."[2]

Outside of the Federal Reserve System, other central bank officials downplayed the tension and argued for a swift response to signs of financial instability. For example, Christine Lagarde, the President of the European Central Bank (ECB), flatly rejected the notion that responding to financial

[1] Lael Brainard in a speech to the Hutchins Center on Fiscal and Monetary Policy, The Brookings Institution, Washington, D.C., December 3, 2014

[2] Austan D. Goolsbee in a speech to the Economic Club of Chicago Forum Luncheon, Chicago, IL, April 11, 2023.

instability could conflict with the attainment of low and stable inflation: "...there is no trade-off between price stability and financial stability."[3]

While anecdotal evidence suggests that Federal Reserve officials may pursue goals beyond the dual mandate, it is unclear how widespread this belief is and whether it affects monetary policy or influences financial markets. To examine these questions, we compile the largest collection (corpus) of Federal Reserve speeches and focus on those delivered between 1960 and 2020.[4] The length of the corpus permits us to measure changes in the Fed's position over time and in response to events, such as financial crises and recessions. And the variation across districts and across Federal Reserve officials enables us to examine heterogeneity and control for speaker and district-specific idiosyncracies.

In addition to assembling a novel speech corpus, we also employ a collection of modified large language models (LLMs) to allow for the precise measurement of concepts that are inaccessible using more familiar approaches in economics. Dictionary-based methods, for example, count the frequency with which certain words appear in a document and are often used in economics to measure latent features of text; they are best employed, however, when the researcher has a strong prior about the words used to describe a topic and when the text data are not inherently informative (Gentzkow et al., 2019). Ideal applications for dictionary-based methods in economics include the measurement of Economic Policy Uncertainty (Baker et al., 2016) and sentiment in central bank texts (Loughran and McDonald, 2011; Apel and Grimaldi, 2014; Apel et al., 2022; Correa et al., 2021).

We are interested in measuring concepts relevant to policymakers that are embodied in sequences of text, such as sentences or paragraphs, rather than word frequencies. For example, in some exercises, we want to evaluate whether a speech given by a Federal Reserve official provides evidence that he or she believes that monetary policy should be used to achieve financial stability. In other cases, we want to determine whether a statement about financial stability discusses a past financial crisis, an ongoing financial crisis, or concerns about the possibility of a future crisis.[5] Each distinct concept would require its own dictionary, which would rely on us having a strong prior about how to measure such concepts through the frequency of word use and without examining words in the context of their usage. In contrast, the approach we take makes use of the flexibility and precision of large language models, which can achieve state-of-the-art performance on a variety of language tasks after slight modification.

More specifically, we use a collection of LLMs called transformer models (Vaswani et al., 2017; Devlin et al., 2019; Liu et al., 2019), which we fine-tune for improved performance on central bank texts. These models process text sequences and are pre-trained on general language tasks, such as masked language learning, which involves predicting omitted words in a sentence. This pre-training protocol enables the automatic generation of extensive training sets from large text corpora, such as the entirety of Wikipedia. Combined with other beneficial characteristics of transformer models

[3]Christine Lagarde in a hearing of the Committee on Economic and Monetary Affairs of the European Parliament, Brussels, March 20, 2023.

[4]The corpus extends back to the founding of the Federal Reserve System in 1914; however, we focus on speeches from 1960 onward, since speech frequency is much higher in this period.

[5]Byrne et al. (2022) show that temporal focus is an important text feature that has largely been absent from existing work on central bank communication. They quantify the temporal dimension of text to study the communication transmission mechanism and the information deficit.

(discussed in the Appendix), this approach allows for the development of models that are orders of magnitude larger and have a greater ability to retain information about natural language. This base or "foundation" model is then fine-tuned to perform different language tasks, such as instruction following, text summarization, or text classification, often with proficiency equal to, or even surpassing, human performance. Generative versions of LLMs, like the GPT-3.5 and GPT-4 models that underpin ChatGPT, can write text that is indistinguishable from human-produced work and are even capable of passing rigorous licensing exams in medicine and law (OpenAI, 2023).

Employing these LLMs, we first attempt to distinguish between content in the Federal Reserve speech corpus that discusses the dual mandate and content that discusses other topics. To do this, we extend the pre-training of a type of LLM called a "sentence transformer" (Devlin et al., 2019; Reimers and Gurevych, 2019) using the Semantic Scholar Open Corpus (Lo et al., 2020) to enhance its understanding of language used in the context of central banking. A sentence transformer maps a text sequence, be it a sentence or a paragraph, to an embedding, which is a dense vector representation of the text. Embeddings that represent related text co-locate in the embedding space.

After extending the pre-training process, we further refine the model to yield embeddings that are comparable using cosine similarity scores, where a higher score indicates that two speech excerpts contain similar content. We then employ these model-generated embeddings to compare a natural language description of the dual mandate with the content of each paragraph in each speech. When the cosine similarity scores are high, we categorize the paragraph as containing content related to the dual mandate. Note that we do not use the term "dual mandate" to identify references to the Fed's mandate, since it was not codified into law until the late 1970s.[6] Furthermore, the term itself was not commonly used by the Fed until the mid-1990s.

Classifying paragraphs as being related to the dual mandate also allows us to identify those paragraphs that are not. We apply an LLM that performs extractive question answering to each of these "non-dual mandate" paragraphs and identify the excerpt in each paragraph that corresponds to the speaker's most significant concern. Parsing the output, we find that the most frequently discussed concerns relate to financial stability and the financial sector. While the Fed also discusses other topics, such as economic inequality and climate risk, these do not appear to account for a substantial share of the speech content, even in the latter half of the sample.

Sorting paragraphs into "dual mandate" and "non-dual mandate" groups provides us with a coarse division of the speech corpus. We also construct a more granular partition that places each paragraph into a class based on the economic or financial topic it discusses. This is achieved through the use of a separate pre-trained LLM and an approach known as zero-shot classification (Pushp and Srivastava, 2017), which generates a probability distribution over a set of classification labels without first requiring training specific to that classification task. We use the labels "financial stability," "output," "inflation," and "labor market" and generate classification scores for all speech paragraphs in our sample. Relative to topic modeling, this approach offers three advantages in our context. First, it does not suffer from look-ahead bias, since the model is pre-trained on an auxiliary corpus, but not

[6] The Fed's "dual mandate" is described in The Federal Reserve Reform Act of 1977 and the Full Employment and Balanced Growth Act of 1978.

on the Fed speech corpus. Second, it allows us to select the classes of interest, rather than allowing them to emerge as latent features. And third, it permits us to make use of a model that has been extensively pre-trained to have fluency in the language in which it is performing tasks.

Since we are interested in the Fed's interpretation of its mandate, including the admissibility of secondary objectives, we first focus on paragraphs that have high financial stability classification scores. We then employ the sentence transformer described earlier to determine whether the speaker advocates for the use of monetary policy to achieve financial stability. We repeat this exercise to determine whether the speaker advocates for banking regulation to achieve financial stability. Around the time of the Great Recession, we observe a growing endorsement for the use of both methods to achieve financial stability; however, the support for the use of monetary policy begins to rise earlier – even prior to the Great Recession – and remains elevated until the end of our sample period. In contrast, support for the use of banking regulation to achieve financial stability appears to be much more episodic, and transitory.

We next attempt to determine what influences the Fed's support for the use of bank regulation or monetary policy to achieve financial stability, making use of the two aforementioned text features: advocacy for monetary policy as a means to achieve financial stability and advocacy for banking regulation towards the same objective. In separate empirical exercises, we regress these two variables on text features for inflation, employment, financial stability, bank liquidity, and bank capital. We also include macroeconomic and financial controls, housing controls, time and district fixed effects, and an indicator for financial crises. In both cases, we find statistically significant associations at the 1% level between the discussion of the components of the dual mandate (inflation and employment) and advocacy for the use of either monetary policy or banking regulation to achieve financial stability.

Given the extensive set of controls and fixed effects, these results are best interpreted as capturing the Fed's perception of its mandate, rather than its response to economic and financial shocks. In this respect, our findings suggest that the Fed primarily views financial stability in the context of the dual mandate and banking regulation, rather than as a separate component of its mandate, contributing to the recent discussion in Peek et al. (2016), van Dieijen and Lumsdaine (2019) and Istrefi et al. (2021).

We also show that concerns about bank capital appear to be associated with advocacy for the use of banking regulation (but not monetary policy) to achieve financial stability. In contrast, concerns about financial crises tend to be accompanied by calls for both the use of monetary policy and banking regulation. These results suggest that the Fed does not always see monetary policy as the correct instrument for achieving financial stability. Rather, it holds a more nuanced view that is closer to what Lael Brainard articulated in her 2014 speech.

We have now established that the Fed connects discussion of financial stability with discussion of its policy tools. We also show the appearance of such content in Fed speeches is associated with policy decisions and movements in asset prices.[7] In particular, a one standard deviation increase in the discussion of financial crises is associated with a 26 basis point reduction in the rate of return on

[7] This contributes to a growing literature on the subject.(Clarida et al., 2000; Orphanides, 2003; Sims and Zha, 2006; Bianchi, 2012; Bianchi et al., 2022).

safe assets, even after the inclusion of macroeconomic and financial controls. Additionally, discussion of financial stability is associated with decreases in equity returns and risky asset returns, which suggests that it has an impact on market participants through the risk-taking channel of monetary policy (Jiménez et al., 2014; Dell'Ariccia et al., 2017). Discussion of financial stability also appears to be an important determinant of the asset valuation-monetary policy regimes identified in Bianchi et al. (2022).

In addition to the aforementioned empirical exercises, we also evaluate the suitability of LLMs for tasks in economics and, in particular, central bank communication. In one such exercise, we aggregate the text features into time series and apply endogenous structural break tests, allowing for multiple breaks at unknown dates (Bai, 1997; Bai and Perron, 1998). We find that breaks coincide with noteworthy events in the Fed's institutional history, such as peaks in the federal funds rate and inflation, the passage of banking regulation such as the Gramm-Leach-Bliley Act in 1999, changes in the Fed chairperson, the Global Financial Crisis and policy events that followed, providing support for the methods and models we apply in this paper.

Finally, we extend our main findings by examining the impact of communication during the long Zero Lower Bound (ZLB) episode that followed the Great Recession, which was speculated to have amplified the importance of communication in policymakers' toolkits, see Hansen and McMahon (2016). A popular way to assess changes in monetary policy is the Taylor rule framework (Taylor, 1993, 1999). Conceptually, deviations from a monetary policy rule may be driven by factors that are either inside or outside of the Fed's mandate. Making use of the text features described earlier, we estimate an augmented Taylor rule to determine whether they have explanatory power during this period.

Because the target federal funds rate (TFFR) is effectively zero after the Great Recession, we use the shadow rate when estimating the augmented Taylor rule. We find that a one standard deviation increase in textual content that is not related to the dual mandate is associated with a 19 basis point reduction in the shadow rate. This effect is even more pronounced for financial stability discussion, where an equivalent increase is associated with a 31 basis point reduction. This is consistent with our descriptive finding that content that is not about the mandate content tends to discuss financial stability. It also accords with our finding that discussion of financial stability signals a more accommodative stance when there is weakness in labor markets.

Our paper uses state-of-the art machine learning tools for natural language processing to evaluate key aspects of Federal Reserve policy, namely the extent of financial stability considerations, the advocacy for using either monetary policy or banking regulation in pursuit of these objectives, and the implications for policy decisions and market participants. Understanding these dynamics will help policymakers to make better-informed decisions, formulate policies that are in line with the central bank's objectives, and communicate them effectively, thereby enhancing the consistency and transparency of central bank decision-making processes.

The paper is organized as follows. Section 1 discusses the data we use and Section 2 introduces the natural language processing (NLP) methods. Thereafter, Section 3 introduces the text features extracted using LLMs and presents a set of descriptive findings. Section 4 discusses the econometric results. Finally, Section 5 concludes.

1 Data

We use three different types of data. The first is a collection of speeches given by Federal Reserve Bank presidents and members of the Board of Governors, which we use to measure features of Federal Reserve officials' communication, including discussion of financial stability and its institutional mandate. The second is a corpus of journal articles and working papers related to central banking, which we use to refine the natural language processing (NLP) models, allowing us to extract higher quality features from the Fed speeches. The third is a set of macroeconomic and financial variables, which are used as either controls or the dependent variable in different regression exercises. For the sake of consistency and to ensure coverage over the long sample period, we take most of these variables from the Macrohistory Database, introduced by Jordà et al. (2016).

1.1 Federal Reserve Speeches

Our primary source of text data is a novel collection of speeches given by presidents of Federal Reserve Banks and members of the Board of Governors of the Federal Reserve System. It includes 6,851 speeches and 129 speakers and spans the period between 1914 and July of 2021.[8] While the coverage of speeches given is not complete over the sample period, it is, as far as we are aware, the most comprehensive collection assembled.

Speech frequency increases over the sample period. This is a consequence of two factors. First, the Fed increased the frequency of its public communication over the sample. And second, works that were produced more recently are more likely to have been digitized and incorporated into an online collection.[9] There are fewer than 20 speeches in the corpus for most years between 1914 and the early 1960s. However, after 1960, the number of speeches given and available annually tends to rise over time. The number also jumps again in the late 1990s.[10] We will focus primarily on the period between 1960 and 2021. This period includes 6,195 speeches delivered by 99 speakers.

1.2 Journal Articles and Working Papers

The second text corpus consists of journal articles and working papers that were drawn from the Semantic Scholar Open Corpus (S2ORC), introduced in Lo et al. (2020). The latest version of the dataset contains metadata for 136M papers, including titles, abstracts, and citations. Within this corpus, we focus on the 2.3M journal articles and working papers that were identified by S2ORC as being from the field of economics.

[8] We build on the sample by van Dieijen and Lumsdaine (2019), containing speeches of the Board of Governors from 1997-2016, which we extend over a longer time period and augment with speeches of the Federal Reserve Bank presidents.

[9] See Figure A.1 of the Appendix for a plot of the speech counts by year

[10] In the late 1990s under Chairman Greenspan, there was a gradual shift to more transparency, which is not only reflected in a large increase in the frequency of speeches by presidents of Federal Reserve Banks and members of the Board of Governors, but also in other communication, such as the introduction in 1994 of FOMC statements in conjunction with all rate changes and later on in 2000 for all FOMC meetings.

We next identify the subset of articles that discuss topics related to macroeconomics, monetary economics, and financial markets. We then filter those articles using two additional criteria. First, the article must have an abstract available, because abstracts will be used in the training process. And second, it must be published in a journal or working paper series that has at least 500 entries in the S2ORC database. The second criterion is intended to filter out articles from obscure journals. The final sample comprises 328,370 articles.

From this final sample, we construct two datasets. The first consists of the full article abstract texts. This is used to extend and fine-tune the pre-training of the LLMs we use in this paper, which are already pre-trained on large text corpora, including the full English language Wikipedia.[11]

To construct the second dataset, we start by selecting abstracts that reference central banks, which reduces the sample to 29,781 articles. We then divide each abstract into sentences and identify two types of sentence pairs: 1) sentences in the same abstract; and 2) sentences in different abstracts. We randomly select an equal number of both types of pairs, yielding a total sample size of 194,227, in order to emulate the dataset construction process for the next sentence prediction (NSP) task. We use the text data and this task to pre-train an LLM called BERT (Vaswani et al., 2017), as described below.[12]

Table A1 in the Appendix provides article counts for a subset of the journals and working paper series in the sentence pair corpus. The most common working paper repository is the Social Science Research Network (SSRN), which accounts for almost 14% of the 29,781 articles. The most common journal is the *Journal of Banking and Finance*. In total, there are 283 journals and working paper series included in the corpus.

1.3 Macroeconomic and Financial Data

In addition to the text data, we also use macroeconomic and financial data in regression exercises. These include bond returns, CPI inflation, financial crisis classifications, the debt-to-GDP ratio, return on equity, house prices, the loan-to-deposit ratio, returns to risky assets, returns to safe assets, the interest rate, loan origination volume, and the output gap. With the exception of the output gap, all variables are taken from the Macrohistory Database (Jordà et al., 2016). The output gap is measured as the percentage difference between actual and potential (real) GDP. Actual GDP is taken from the Bureau of Economic Analysis (BEA) and potential GDP is measured by the Congressional Budget Office (CBO). All variables are measured at an annual frequency and span the period between 1960 and 2020. All variables used are listed in Table A2 in the Appendix.

2 Methods

Our objective is to examine the Fed's interpretation of its own mandate. We do this by extracting text features from Federal Reserve speeches and examining their variation across time, district, and

[11] See Section 2.2.1 for a description of the models and pre-training process.

[12] This task involves using the current sentence in a sequence to predict the next sentence. Rather than train on the NSP task, however, we instead fine-tune our models using the approach in Reimers and Gurevych (2019), because our objective is to measure the similarity between sentences, not to identify the precise next sentence.

speaker. Many of the more informative features involve sequences of words, such as sentences and paragraphs, rather than individual words. As such, we use LLMs that are capable of modeling sequences, building on recently introduced variants of transformer models (Vaswani et al., 2017). This section provides a brief overview of the NLP tasks we perform, the LLMs we use, and the types of features we extract. We also discuss the pre-training and fine-tuning process for these models, but relegate the technical details about the NLP models to Section A.2 of the Appendix.

2.1 The NLP Task

Many natural language processing tasks can only be performed using a sequence-to-sequence (S2S) model, which maps one sequence of symbols to another sequence of symbols. Others do not require an S2S model, but achieve better performance when one is employed. Machine translation tasks, for instance, yield low quality results when performed at the word level. An entire sentence typically needs to be processed and interpreted before a suitable translation can be generated. This is, in part, because words have different meanings in different contexts. See A.1 in the Appendix for an overview of S2S modeling.

Another example of an inherently sequential NLP task is extractive question answering. This involves finding a subsequence of text that contains an answer to a question. An NLP model would receive a "context" sequence (the original text) and a query as inputs. It would then yield a subsequence of the context as an output. Both the inputs and output are sequences. State-of-the-art S2S models, such the LLMs we use in this paper, are naturally suited to such problems.

For our purposes, S2S modeling will mostly involve the transformation of a sequence of words in a paragraph into a sequence of "contextualized" words. The words themselves will be represented using dense vectors called embeddings.[13] The contextualized embeddings will encode information about the meaning of the word in the context in which it was used. For example, the word "run" will be encoded differently in the following two sequences:

Sequence 1: "If depositor concerns are not addressed, there could be a bank *run*."
Sequence 2: "If stock prices increase again tomorrow, it will be the longest bull *run* in history."

Transformer models output sequences of contextualized embeddings, which can be used to perform a wide variety of textual analysis tasks, including sentiment analysis, zero shot classification (classification without training), textual similarity measurement, machine translation, contextual embedding generation, and extractive text summarization. Transformers also can be extended and trained to perform supervised learning tasks, such as sentiment classification. Furthermore, such extensions can be fine-tuned at a low computational cost to yield further improved performance in a specific domain, such as central banking communication.

We use transformer models to extract features from Fed speeches, focusing on elements related to the Fed's mandate. Using transformer models that were fine-tuned for different language tasks enables us to measure subtle features of a speech, such as the speaker's primary concern in a given

[13] See Gentzkow et al. (2019) for an overview of embeddings.

paragraph or the extent to which a statement expresses approval of monetary policy as a means of achieving financial stability. It also allows us to identify whether certain content – such as concern about bank liquidity – is present in a paragraph.

2.2 Large Language Models

We make use of a collection of LLMs called transformer models. Transformer models introduced several innovations that have proven useful for natural language processing, including the attention mechanism, which enabled the model to learn the relationships between words, irrespective of where they appeared in a sequence. Figure 1 illustrates the application of the attention mechanism to sequences taken from two Federal Reserve speeches. See Section A.2 of the Appendix for a discussion of the transformer model's architecture and contributions to text sequence modeling.

Figure 1: Attention Mechanism

Notes: The left panel illustrates the attention mechanism applied to the word "banks" in a sequence from a speech given by Gary Stern in January 2009. The words "liquid(ity)," "financial institutions," and "markets" are not close to "banks" in the sequence, but the attention mechanism determines that they are highly relevant for contextualizing banks. The right panel illustrates self-attention applied to the word "manufacturing" in a sequence taken from a speech given in September 2002 given by Cathy Minehan, then-President of the Federal Reserve Bank of Boston. The attention mechanism identifies the importance of "industrial," "production," and "capacity" for contextualizing "manufacturing" in the sequence. Note that the `[CLS]` "word" indicates the start of a sequence and `##` indicates that a word is outside of the corpus.

In addition to the improvements in model architecture, transformer models also make use of extensive pre-training. This entails training the model on a large, auxiliary dataset that is unrelated to the NLP task of interest. The purpose of pre-training is to create a baseline or "foundation" model that understands a given language using large amounts of text. This model can then be fine-tuned on a small amount of text that is closely related to the language task of interest. In our case, for example, the model can be pre-trained on all English language Wikipedia articles, newspaper articles, and journal articles. It can then be fine-tuned on a much smaller corpus of Federal Reserve speeches.

Because we use a variety of pre-trained and fine-tuned models in this paper – and also extend the pre-training and fine-tuning process – we discuss both topics in the remainder of this subsection.

2.2.1 Pre-training

The NLP literature has shown that pre-training a language model can yield substantial performance benefits on subsequent (downstream) tasks (Dai and Le, 2015; Peters et al., 2018; Salimans and Sutskever, 2018; Howard and Ruder, 2018). Indeed, some of the most substantial gains in the development of language models have come from changes in the pre-training procedure and data, rather than changes in model architecture. The BERT and RoBERTa models, for instance, achieved state-of-the-art performance on NLP benchmarks using the same transformer architecture introduced in Vaswani et al. (2017), but with modified training processes.

BERT. The BERT model, introduced in Devlin et al. (2019), proposed several modifications to the training process used in Vaswani et al. (2017). Most notably, the pre-training process concentrated on two tasks: 1) masked language modeling (MLM); and 2) next sentence prediction (NSP). Both tasks automatically generate labels, allowing for pre-training on arbitrarily large text datasets that have not been labeled by a human.

For concreteness, we demonstrate how the BERT model would use the passage given in the quote below in the training process for each of these two tasks. The text is taken from a speech given by Gary Stern, then-President of the Minneapolis Federal Reserve Bank, in January 2009.

> *"Commercial banks have long been thought of, and indeed have functioned as, the backup source of liquidity for many other financial institutions and markets. Banks continue to play this role, but it has become more challenging today to do so because some lenders find themselves capital-constrained as a result of recent losses and/or sizable, unanticipated additions to their balance sheets of formerly off-balance-sheet instruments."*

The MLM task involves masking randomly-selected words in a sequence and then training the model to predict them. In the quote above, MLM might generate the following sequence and labels.

> **Sequence:** "Commercial $\texttt{[MASK]}_1$ have long been thought of, and indeed have functioned as, the backup source of $\texttt{[MASK]}_2$ for many other financial institutions and markets."
> **Labels:** $\texttt{[MASK]}_1$ = banks, $\texttt{[MASK]}_2$ = liquidity.

Another benefit of pre-training with MLM is that it allows for bidirectionality in the interpretation of sequences. Rather than training BERT to predict the next word in a sequence, conditional on the preceding words, it is instead trained to predict a missing word in a sequence, conditional on all words before and after it. This results in a pre-trained model that has a substantially expanded capacity to understand language.

The other training task, next sentence prediction (NSP), also allows for the automatic generation of labels. This task presents the model with a sequence of two sentences drawn from the corpus. The model must determine whether the second sentence follows the first or whether it is drawn from a different place in the document. Again, returning to the speech from Gary Stern, consider the following three sentences.

> **<u>Sentence A</u>:** "Commercial banks have long been thought of, and indeed have functioned as, the backup source of liquidity for many other financial institutions and markets.
> **<u>Sentence B</u>:** "Banks continue to play this role, but it has become more challenging today to do so because some lenders find themselves capital-constrained as a result of recent losses and/or sizable, unanticipated additions to their balance sheets of formerly off-balance-sheet instruments."
> **<u>Sentence C</u>:** "On the positive side, term funding is more readily available than at the height of the crisis, and risk premia have diminished through much of the financial sector."

In the speech, Sentence B follows Sentence A; whereas Sentence C is selected from a random location in the text. The MLM process might yield sequence (A,B), which could be passed to BERT with the training label `IsNextSentence`. Alternatively, it could yield the sequence (A,C), which could be passed to BERT with the training label `IsNotNextSentence`.

Because neither the MLM nor NSP tasks require labeled data, Devlin et al. (2019) were able to train BERT on two large text corpora: the 800M word Book Corpus from Zhu et al. (2015) and a 2500M-word corpus constructed from English language Wikipedia articles. This increase in the size of the training corpus allowed for a corresponding increases in model size. Making use of the extended training corpus, Devlin et al. (2019) pre-trained two versions of BERT: 1) $\texttt{BERT}_{\texttt{BASE}}$, which has 12 transformer blocks, a hidden dimension size of 768, 12 attention heads, and 110M parameters; and 2) $\texttt{BERT}_{\texttt{LARGE}}$, which has 24 transformer blocks, a hidden dimension size of 1024, 16 attention heads, and 340M parameters.

Devlin et al. (2019) show that the pre-trained BERT model can then be fine-tuned to achieve state-of-the-art performance on downstream tasks, such as question answering and language inference. This entails training a model to take the contextualized embedding output from BERT to use as an input to a supervised learning task. This can be done with a few hours of GPU training and does not require modifications to the model's architecture that are specific to the task. The authors used this approach to achieve state-of-the-art performance on 11 NLP benchmarks.

The fine-tuned versions of BERT introduced in Devlin et al. (2019) remain near the state-of-the-art on the GLUE (Wang et al., 2019) and SQuAD (Rajpurkar et al., 2018) benchmarks, which are used to assess the performance of NLP models on specific language tasks. The models that have since surpassed them also use transfomer model architectures. Fine-tuned BERT models have also demonstrated better performance on domain-specific NLP tasks than models that were trained exclusively using text from that domain. This suggests that pre-trained transformer models are likely to perform as well or better than the alternative model choices for the language tasks we perform in this paper.

RoBERTa. Liu et al. (2019) introduced the RoBERTa model, a "robustly optimized" version of BERT, arguing that the gains from BERT are primarily attributable to the modification of the pre-training process, which allows for an increase in model size by several orders of magnitude. They also argue that this is true of most landmark language models, such as ELMo (Peters et al., 2018), GPT (Salimans and Sutskever, 2018), XLM (Conneau and Lample, 2019), and XLNet (Yang et al., 2019).

The RoBERTa model attempts to provide further improvements over the pre-training process in BERT. Specifically, it employs larger datasets, removes the NSP task, trains on larger sequence lengths, employs a dynamic masking pattern, and uses a new dataset (CC-News).

Liu et al. (2019) find that the resulting modifications to the pre-training process boost the "robustly optimized" version of BERT to match or outperform the models that were introduced after BERT. We use versions of the RoBERTa model for many of the NLP-related tasks in this paper.

2.2.2 Fine-tuning

The success of supervised transfer learning – that is, training on a corpus in one domain and performing prediction in a different domain – was one of the motivations for the construction of BERT.[14] Since BERT does not require human labeling, it opened up the possibility of performing transfer learning with much larger text corpora. This first step is typically described as "pre-training" in the context of LLMs.

Once a deep learning model, such as a transformer model, has been pre-trained, we can fine-tune it to perform a task in our domain of interest. Such models are made up of a sequence of layers, which transform the inputs from the previous layer. We typically perform fine-tuning by *freezing* all of the layers, except the output layer and a few of the layers that directly precede it. We then perform supervised learning on a domain-specific task and with domain-specific data. Note that freezing a layer prevents its parameters from being updated. This means that in $\texttt{BERT}_{\texttt{BASE}}$, for instance, fewer than 100K of the 340M parameters will need to be trained. The rest of the model will serve as a state-of-the-art text feature extractor.

2.3 Text Feature Generation

So far, we have discussed the NLP problem we encounter in extracting information from Federal Reserve speeches, the class of models we use to solve it, and the features that make such models appealing. In this subsection, we discuss three different types of text tasks that this class of models can use for feature extraction. We start with zero shot learning, which we will use to classify text into categories without training a model. We then discuss extractive question answering, which we will use to identify areas of concern for the speeches in our corpus. Finally, we examine semantic textual similarity (STS) measurement, which will allow us to evaluate whether two statements are closely related. In one application, we will use this to determine whether speakers discuss monetary policy as a tool for achieving financial stability.

2.3.1 Zero Shot Classification

For some exercises in this paper, we will need to perform text classification. This normally involves training a supervised learning model with labelled text categories. However, in the absence of an

[14] See Conneau et al. (2017) and McCann et al. (2017) for examples of transfer learning.

appropriate dataset or labels, it is not be possible to perform supervised classification. An alternative approach, introduced in Pushp and Srivastava (2017), trains a general model that can perform classification on arbitrarily chosen categories without first labelling the data.[15]

To perform zero shot classification, we follow the approach taken in Pushp and Srivastava (2017). We start by constructing a corpus using labelled text data that is outside of our domain of interest. We then embed the text sequences and the labels in the same embedding space. Pushp and Srivastava (2017) do this by averaging word embeddings for the sequence; however, we can also do this using a sentence transformer, such as SBERT (sentence BERT), which can generate a contextualized embedding for the sequence. We then concatenate pairs of sequence and label embeddings, and train a supervised model to predict whether or not they match.

We perform zero shot classification by embedding candidate labels and sequences of interest in the same space and then using the model to perform classification. This will yield the probability that the sequence and candidate label match.

Below, we provide an example of zero shot classification with a fine-tuned BERT model. We attempt to classify whether the statement is about one of the following four categories of topics: financial stability, output, inflation, or the labor market. The scores can be interpreted as a probability distribution over the categories.

> **Sequence:** "Banks continue to play this role but it has become more challenging today to do so because some lenders find themselves capital constrained as a result of recent losses and or sizable unanticipated additions to their balance sheets of formerly off balance sheet instruments."
> **Candidate Classes:** ['financial stability', 'output', 'inflation', 'labor market']
> **Scores:** [0.718, 0.203, 0.048, 0.031]

The model identifies "financial stability" as the most probable label, even though it was not trained on a central bank communication corpus. As we will see in later exercises, transformer models perform well on simple zero shot classification tasks on Fed speeches.

2.3.2 Extractive Question Answering

The original BERT model (Devlin et al., 2019) was trained to perform extractive question answering on the Stanford Question Answering Dataset (Rajpurkar et al., 2016), which consists of 100,000 human-labelled questions, answers, and context passages. The context is taken from a Wikpedia article. The model must correctly predict the subsegment of the text that contains the correct answer to the question. BERT exceeded human-level performance on both benchmarks for that task.

We use extractive question answering with pre-trained BERT models in several exercises in this paper. For some exercises, we attempt to determine the speaker's most significant concern in a passage. The two examples given below provide the query, context, and model output for passages in a speech given by then-President of the Federal Reserve Bank of St. Louis, Darryl Francis, in February 1972.

[15] See Yogatama et al. (2017), Zhang et al. (2019), and Yin et al. (2019) for alternative approaches to zero shot text classification.

Query 1: What is the most significant concern in the passage?
Context 1: "The suspension of the convertibility of the dollar into gold and the imposition of a 10 percent import surcharge last summer ran the risk of mass foreign retaliation in the form of destructive trade barriers."
Output 1: mass foreign retaliation

Query 2: What is the most significant concern in the passage?
Context 2: "Another significant aspect of the President's new policies announced August 15 are the measures taken to reverse the deteriorating US balance of payments."
Output 2: deteriorating US balance of payments

Once the most significant concerns have been extracted from a speech, they can then be converted to a set of numerical features, such as contextualized word embeddings via BERT or a sentence embedding via SBERT, that can then be further analyzed via econometric methods.

2.3.3 Semantic Textual Similarity

For several exercises, we will need to measure the similarity between pairs of passages from speeches using a measure called semantic textual similarity (STS). The BERT and RoBERTa models, which we use throughout the paper, achieve state-of-the-art performance on the STS task; however, they require each pair of passages to be input into the model simultaneously to produce an STS score and hence are computationally intensive. For 10,000 sentences, this would require around 50M STS pair computations. To compute semantic textual similarity efficiently, we make use of SBERT models. We also extend the pre-training process and perform fine-tuning to improve performance on our corpus. For a detailed description of this process, see Section A.3 of the Appendix.

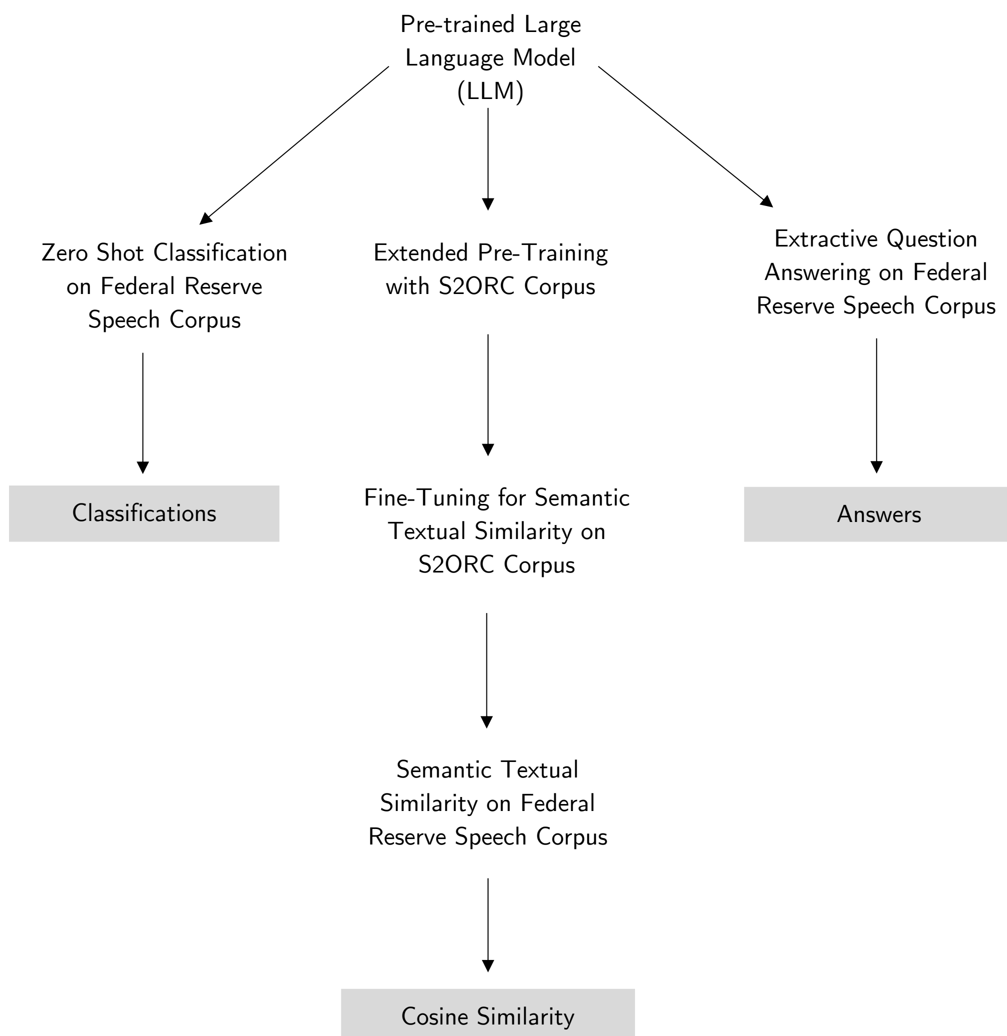
2.3.4 LLM Training Summary

In Figure 2, we summarize the training process for the LLMs used in this paper. All models are initially pre-trained on large corpora, such as the entirety of Wikipedia, by the developers who introduced them. We start with these versions of the models. In exercises that involve zero shot classification or extractive question answering, the models can be directly applied to the speech corpus. In exercises that make use of cosine similarities produced by sentence transformers, we extend the pre-training process and fine-tune the model to the semantic textual similarity task. In some exercises, we combine multiple approaches and models.

3 Interpretation of Text Features

In the previous section, we discussed three methods for extracting text features using transformer models: zero shot classification, extractive question answering, and semantic textual similarity measurement. In this section, we will describe a selection of the features we extracted from Federal Reserve speeches using those methods. Our objective is to examine whether they adequately represent the

Figure 2: LLM Training Summary



Notes : This figure provides an overview of the training process for the LLM models used in this paper. All models are pre-trained on large corpora, such as the entirety of Wikipedia. In exercises that involve zero shot classification or extractive question answering, the models can be directly applied to the Fed speech corpus. In exercises that make use of cosine similarities produced by sentence transformers, we extend the pre-training process and fine-tune the model on the semantic textual similarity task before applying to the Fed speech corpus. In some exercises, we combine multiple approaches and models. For example, in some exercises, we use zero shot classification to identify paragraphs that contain content about financial stability. We then compute the cosine similarity between those paragraphs and other natural language statements (e.g., monetary policy).

concept we intended to measure in the text. We describe how some of the more informative features evolved over time and across district.

Most of the features discussed in this section are constructed from paragraph-length sequences. For relatively simple features, we use zero shot classification with the RoBERTa model to determine whether a paragraph discusses a certain concept, such as financial stability. For more subtle concepts, we extend the pre-training and fine-tuning of the model to improve its capacity to understand central bank texts. We then measure cosine similarity between paragraphs of speech text and the statement we want to evaluate. Finally, we standardize the raw scores by subtracting the sample mean and dividing by the sample standard deviation.

Using S2S models – and, in particular, the RoBERTa model – has at least three advantages. First, in contrast to dictionary-based methods, RoBERTa's classifications are based on the entire paragraph, taking long-run dependencies, negations, and modifiers into account. Second, RoBERTa automatically identifies related terms and, thus, does not rely on the ex-ante identification of all relevant terms. And third, unlike average word embeddings, RoBERTa accounts for the context in which words are used.

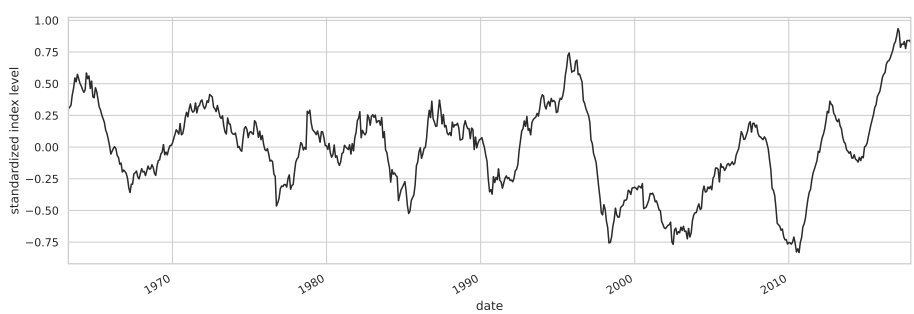
3.1 Time Variation in Text Features

Since we are interested in measuring changes in the Federal Reserve System's interpretation of its mandate over time, we will start by examining the time variation in the text features. In this subsection, we discuss what features we measure, how they perform, and what insights they provide into the Federal Reserve's interpretation of its mandate.

3.1.1 Dual Mandate Content

We first partition the text into content that is related to the Fed's "dual mandate" and content that is not. Because the sample starts prior to the introduction of the dual mandate, we measure whether paragraphs discuss (at least one of) inflation, employment, or output growth, rather than attempting to identify references to the dual mandate itself. The evolution of the series is plotted in Figure 3.

Figure 3: Text Feature: Dual Mandate Content



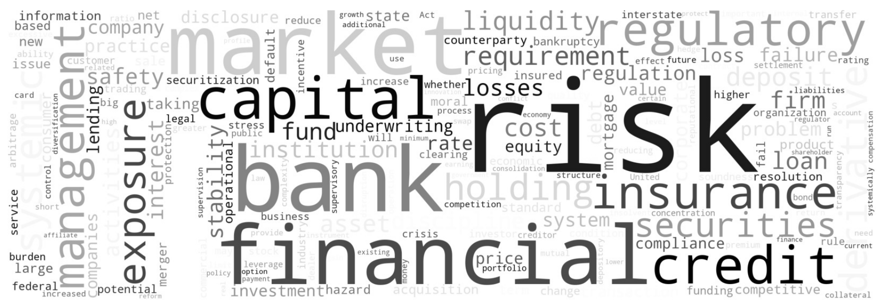
Notes: The figure shows the 2-year rolling mean of the standardized cosine similarity score from BERT for "inflation, employment, and output growth" and each paragraph in the text.

The figure indicates that there is both short- and long-term variation in the dual mandate content of speeches. We examine the content of paragraphs that have low dual mandate content and find that they are largely related to the financial sector, financial crises, and banking regulation. To demonstrate this informally, we use extractive question answering with the RoBERTa model to identify the speaker's concern in each paragraph that has a low dual mandate content score. We then construct word clouds for the pre-Great Moderation period and the Great Moderation period in Figure 4. In both cases, the discussion is dominated by topics related to finance and banking regulation.

Figure 4: Non-Dual Mandate Content Word Cloud

a. 1960-1983

b.1984-2020



Notes: The figures above show word clouds of *concerning* terms that appear in statements with low dual mandate content scores. Such statements are identifed using extractive question answering.

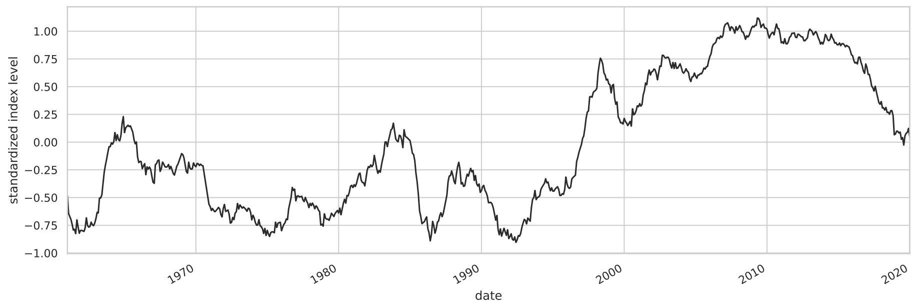
3.1.2 Financial Stability Content

The top panel of Figure 5 shows the 24-month rolling mean of the financial stability index. As expected, the highest levels of the financial stability classification score coincide with the Great Recession. It is preceeded by a spike around 1998, coinciding with the Asian and Russian financial crises and the collapse of hedge fund Long-Term Capital Management (LTCM), which resulted in a bailout of LTCM by a group of private banks that was orchestrated by the Federal Reserve Bank of New York. We can see that financial stability content is declining in the late 1960s and 1970s, followed by another spike around the Latin American Debt Crisis in 1982, when the nine major U.S. banks were heavily exposed to Latin American debt, amounting to 176.5 percent of their capital (Sachs, 1987). Thereafter, the financial stability index declined during the first half of the Great Moderation.

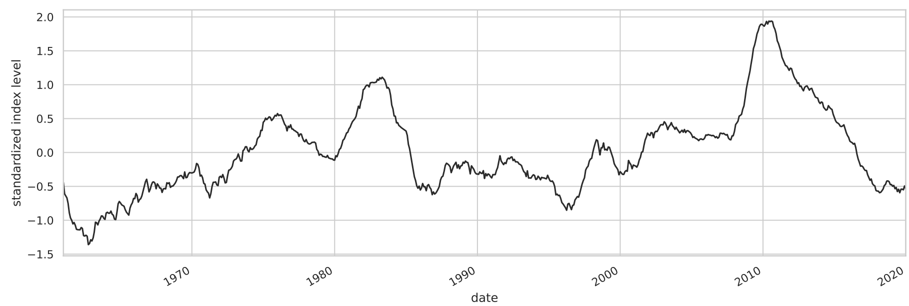
In addition to the financial stability index, we also construct a financial crisis index. This allows us to differentiate between discussions about the Fed's role in achieving financial stability and deliberations about specific financial crises. The bottom panel of Figure 5 plots the 24-month rolling mean of the financial crisis feature. In contrast to the financial stability discussion, there is a general upward trend in financial crisis content prior to the start of the Great Moderation in the mid-1980s. There is also a lull between 2000 and 2007. Interestingly, the Asian and Russian financial crises, as well as

Figure 5: Text Features

a. Financial Stability



b. Financial Crisis



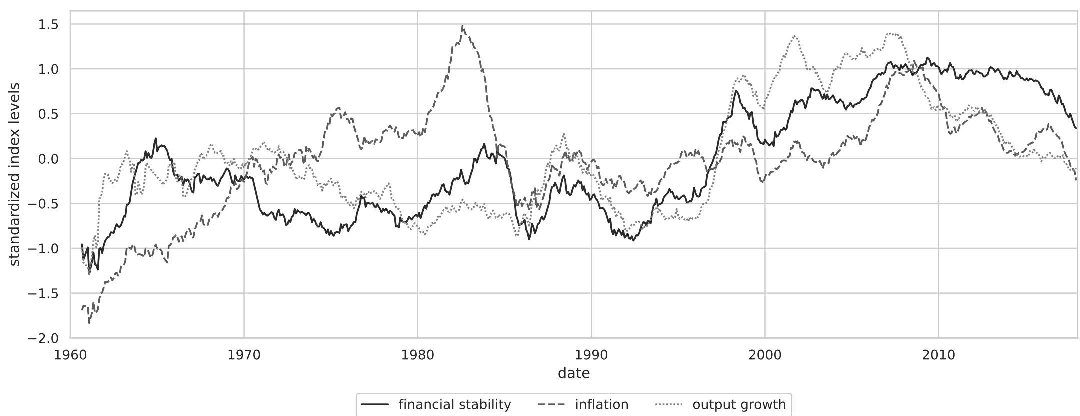
Notes: The figures above show the standardized classification score from BERT for the terms "financial stability" and "financial crisis." This is computed by classifying each paragraph as describing "financial stability" / "financial crisis" or not. We then compute the average score for each month, standardize it, and then plot the 2-year rolling mean.

the collapse of LTCM do not generate a spike in the financial crisis index. This may be explained by the effectiveness of the "Greenspan put" in combination with the much lower exposure of the major U.S. banks to the 1998 external crisis events than the 1982 external crisis events. Also interesting is the peak of the financial crisis index in 2010, the time of the European sovereign debt crisis, and the way in which the index declines rapidly after that peak. In contrast, the financial stability index remains high. Taken together, these contrasting patterns illustrate that while discussion of the global financial crisis itself quickly subsided, discussions of financial stability in its aftermath continued.

3.1.3 Mandate Description

During the period we examine, the Federal Reserve tended to talk about either the components of its dual mandate or financial stability. In Figure 6, we plot the series for inflation and output growth together with the financial stability index to examine how the Fed's focus shifted from dual mandate content to financial stability over time and across cycles. Consistent with inflation data, textual content about inflation in speeches increases during the period in the late 1970s and early 1980s, an era that has been classified as discretionary with large deviations from the Taylor rule (see, e.g., Nikolsko-Rzhevskyy et al., 2014). Another notable feature of the figure is that financial stability and output growth content diverge prior to the Great Moderation, but then positively comove thereafter. This suggests that the Great Moderation could have been a turning point with respect to Fed communication. For this reason, we will include sample splits in our regression exercises in Section 4, so that we can evaluate the pre-Great Moderation and Great Moderation periods separately. It is also interesting that the inflation and output growth features appear to comove in the early 1970s and again during the Volcker years.

Figure 6: Text Features: Inflation, Output Growth, and Financial Stability



Notes : The figure shows the 2-year rolling mean of the standardized classification score from BERT for the terms "financial stability," "output growth," and "inflation."

3.1.4 Financial Stability Position

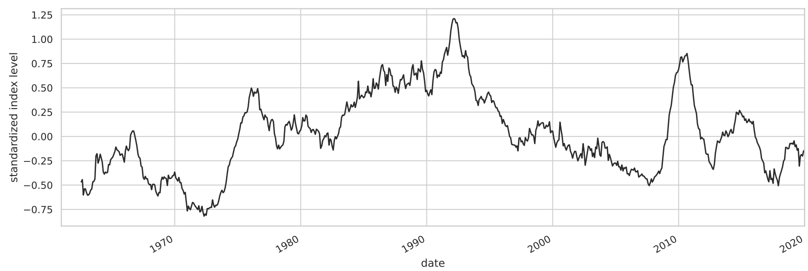
In addition to measuring broad text features related to financial stability, we also construct two features that provide a direct description of the Fed's evolving perception of its mandate. The first feature, shown in the top panel of Figure 7, measures the cosine similarity between each paragraph and the statement "banking regulation should be used to achieve financial stability." Peaks for this series are typically closely related to crisis events and some resemble those seen in the financial crisis text feature in the bottom panel of Figure 5. Yet there are some important differences: the peak in the late 1980s, for instance, seems to follow a similar peak in Figure 5, while the peak in 1992, the year the Basel I capital regulations were implemented, is only apparent in the cosine similarity. In contrast, the bottom panel of Figure 7 measures the cosine similarity between each paragraph and the statement "monetary policy should be used to achieve financial stability," providing us with a measure of how inclined the Fed is to use monetary policy as a means to achieve financial stability.

We can see a clear difference between the evolution of this series and the one for banking regulation, suggesting that the Fed places emphasis on this nuance. In particular, these figures indicate less appetite (or alternatively, less ability, given the zero lower bound) for using monetary policy to achieve financial stability in the past two decades than in the three decades prior. In addition, the more jagged nature of the first plot suggests a more episodic role for banking regulation to achieve financial stability, rather than being a sustained tool.

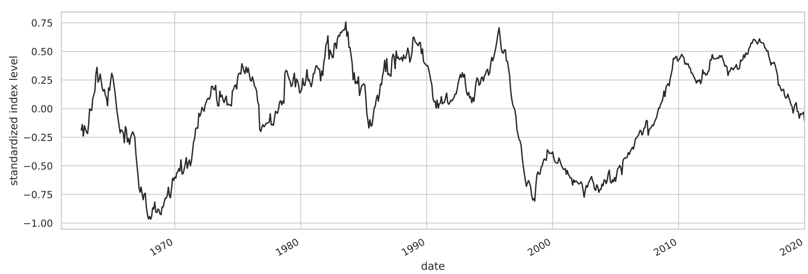
Looking further at the cosine similarity depicted in the bottom panel of Figure 7, it is also apparent that support for the idea that monetary policy should be used to achieve financial stability varies

Figure 7

Cosine Similarity: Banking Regulation and Financial Stability



Cosine Similarity: Monetary Policy and Financial Stability



Notes: Both panels plot standardized cosine similarity scores. The top panel makes use of a cosine similarity score computed between the statement "banking regulation should be used to achieve financial stability" and the contextualized embeddings for paragraphs with high classification scores for "financial stability." Cosine similarity scores are repeated for the same exercise in the bottom panel, but using the statement "monetary policy should be used to achieve financial stability" instead of "banking regulation."

according to the level of accommodation in the Fed's stance, and in particular, the cosine similarity appears to increase as the Fed tightens and decreases (or, in the last two decades remains low or constant) during easing periods. This pattern is consistent with an endorsement of a forward looking "leaning against the wind" view, rather than a "financial instability is caused by monetary tightening" view. We will explore this aspect in our regressions in Section 4.

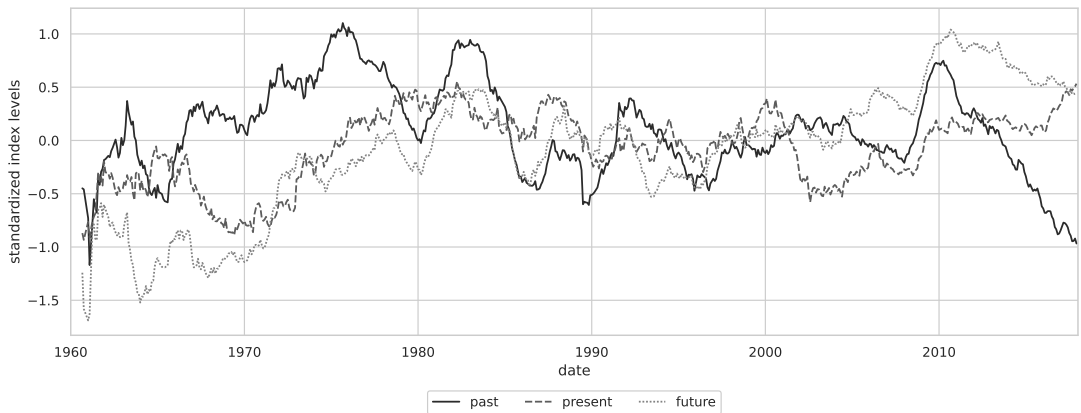
3.1.5 Temporal Focus

In addition to distinguishing between the content of financial stability concerns, it may also be useful to identify the tense or the temporal focus of a statement. This could indicate whether a paragraph is discussing a past crisis, an unfolding crisis, or the prospect of a future crisis and may indicate the extent to which Federal Reserve officials rely on historical precedent, are reactive, or are proactive in making their decisions. We use zero shot learning to classify each tense separately, allowing for the possibility that no clear tense is established in a given statement.

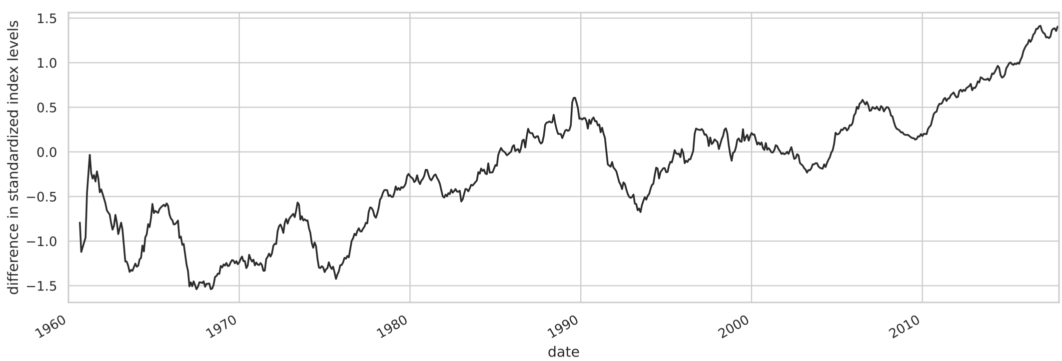
The standardized indices for past, present, and future focus are plotted in the top panel of Figure 8. The bottom panel shows the difference between future and past focus, which has been rising since the start of the Great Moderation. The increased use of future tense coincides with the Fed's emphasis on increased transparency and forward guidance. Prior to the Great Moderation, there is evidence that an increased use of future tense coincides with tightening episodes, whereas the shift towards a

Figure 8: Text Features

Text Features: Past Focus, Present Focus, Future Focus



Text Features: Difference Between Future and Past Focus



Notes: The figure in the top panel shows the standardized classification score for the speech tense for a focus on the "past", "present" and "future." This is computed by classifying each paragraph and then computing the average score for each month, standardizing it, and then plotting the 2-year rolling mean. The figure in the bottom panel shows the 2-year rolling mean of the difference between "future" and "past" focus.

focus on the past coincided with the easing that began in April 1989.

3.1.6 Concern Type and Level

Not all content contained in Fed speeches is equally informative. In order to identify statements and terms that reveal internal concerns, we make use of extractive question answering, as described in Section 2.3.2. Specifically, for each paragraph, we query the model with "What is the most significant concern in the passage?" This yields a concern, which is extracted from the text, along with a score that indicates the model's uncertainty. A low score, for instance, might indicate that no specific concern was stated or that the model was unable to classify the stated concern accurately. Figure 9 visualizes the evolution of "most significant" concerns as they relate to financial stability over time.

Figure 9: Financial Stability Word Clouds

a. 1960-1983

b. 1984-2006

c. 2007-2020

Notes: The figure above shows word clouds of *concerning* terms that appear in statements about "financial stability" during three periods: 1960-1983, 1984-2006, and 2007-2021. We first use zero shot classification with BERT to identify statements about "financial stability." We then use extractive question answering to recover the speaker's stated concern in each paragraph.

3.2 Institutional Variation in Text Features

Another source of variation in text features arises from differences in the institutional arrangements and concerns of different Federal Reserve Banks and the Board of Governors. For instance, the Cleveland and New York Federal Reserve Banks may have different concerns, since there are differences in the compositions of their respective regional economies and the bank holding companies they supervise. Furthermore, given the local appointment processes, some district banks may favor presidents with academic profiles, while others may favor those with private sector experience. In this subsection, we will attempt to document the variation we were able to measure across districts and the Board of

Governors.[16]

Financial Stability Focus. First, we find that the Federal Reserve Banks of New York and Richmond have the highest average financial stability scores, consistent with the fact that most of the systemically important financial institutions (SIFIs) are in those two districts. In addition, many of the speakers with high scores (e.g., Bernanke, Dudley, Evans, Geithner, George, Kocherlakota, Kohn, Kroszner, Lacker, Minehan, Plosser, Stern, Tarullo, and Warsh) were officials in the period surrounding the Great Financial Crisis (2007-2009), a time when the overall financial stability content of speeches was near its highest levels. Many of the other speakers with high scores were members of the Board of Governors; in general, reserve bank presidents' speeches have less financial stability content, although the three highest index scores are associated with reserve bank presidents.[17] The most negative scores are also associated with reserve bank presidents, many of whom served near the start of the Great Moderation (e.g., Eastburn, Kimbrel, Parry, Roos, and Willes). Overall, the results align well with our intuition.

Future Focus. Second, as shown in Table A3 in the Appendix, we find that the Federal Reserve Banks of New York and Richmond also have the highest score on future focus. This is consistent with the previous finding, given the strong connection between future focus and financial stability documented in the cosine similarity results. As noted previously, these districts supervise the largest, most systemically important financial institutions in the US. The NY Fed also has the highest scores for past and present focus; taken together, these results indicate that tense usage is clearer and more distinct in speeches by its presidents than in speeches by presidents of other districts.

Looking across the individual speakers' scores, there is a strong positive correlation between the future focus and financial stability scores (0.67), suggesting that discussions of financial stability involve the use of more forward-looking language.[18] As a result, many of the speakers with high future focus and financial stability scores are the same ones, encompassing many of those that were on the FOMC during the Great Financial Crisis (GFC).

Academic Inclination. There is a long-standing debate in the literature over the extent to which academic discussions influence central bank deliberations and policy-making. Much of the work in this discussion concludes that academic work positively influenced central bank decisions. Some, such as Mankiw (2006), diverge from the majority position and argue that policy-making is largely uninformed by the academic literature. Others, such as Howitt (2012) argue that central bankers often face crises that have not been adequately studied by the academic literature and, thus, typically lead the literature.

On an informational level, Orphanides (2001, 2003) documents the importance of misinformation at the time when the monetary policy decisions are made, while Meltzer (2009a) highlights the role

[16]In the interest of space, we summarize our findings for some of the individual speakers; details of these results are available on request.

[17]In addition to Minehan and Lacker, Loretta Mester, President of the Cleveland Fed has a very high score; the Cleveland Fed co-hosts an annual financial stability conference.

[18]In contrast, the correlation between past or present focus and financial stability is -0.13 and 0.23, respectively.

of misconceptions about economic theory that led the Fed to chase the Phillips curve in an attempt to lower unemployment during the 1970s.

With respect to financial stability, the academic literature has largely argued that macroprudential policy – not monetary policy – should be used to achieve financial stability (Vollmer, 2021). As such, the Fed's interpretation of its mandate and its belief about what may be achievable with monetary policy may also depend on its openness to dialogue with academics. To capture this, we use zero shot learning to identify whether a paragraph refers to an "academic debate or the academic literature."

Just as the NY and Richmond districts have the highest future focus and financial stability scores, so also do they have a high academic focus score. In the case of Richmond, this no doubt reflects the academic backgrounds of the two main presidents in our corpus (Broaddus and Lacker), while for the NY Fed, the presidents' backgrounds are more mixed, but still reflect a high reliance on academic debate and literature. The Federal Reserve Board also has a high academic focus score, unsurprisingly given the more than 200 staff economists that support its work.

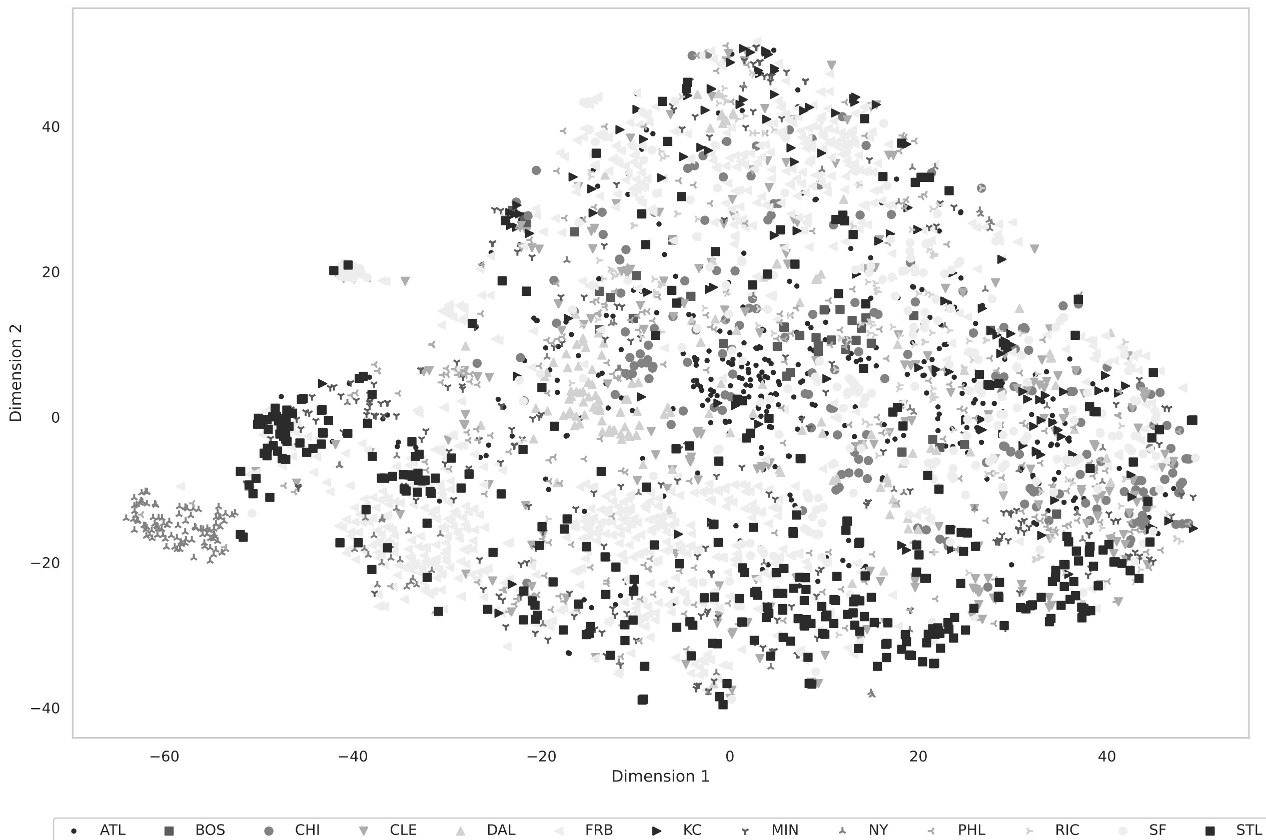
Communication Clusters. Finally, we examine clustering in the financial stability content of Federal Reserve Bank speeches. We cluster on the content of speeches, which we represent using contextualized sequence embeddings produced using the RoBERTa model. For each paragraph of each speech, this yields a 768-dimensional vector, where each dimension corresponds to a text feature. We then compute the average vector for each district-month.

To prepare the data for visualization, we first perform dimensionality reduction using the t-distributed stochastic neighbor embedding (t-SNE) introduced in van der Maaten and Hinton (2008). The t-SNE approach is a nonlinear dimensionality reduction technique that preserves both local and global structure and is intended to produce representations that can be visualized in 2 or 3 dimensions. It is an effective tool for differentiating between and visualizing clusters; however, the distances across clusters are not readily interpretable.

We first apply t-SNE to the full sample, which spans the period between 1960 and 2021, and visualize it in Figure 10. The markers in the visualization represent the different Federal Reserve Banks, as well as the Board of Governors (FRB). The Federal Reserve Bank of New York (FRBNY – denoted NY in the Figure) appears to have a nearly self-contained cluster, suggesting that the content of their financial stability discussions is closer to their own past discussions than it is to contemporaneous communication at other institutions of the Federal Reserve System.

In contrast to the tight and distinct cluster for the FRBNY, the Federal Reserve Bank of St. Louis's communication forms a long cluster that overlaps with the communication of many other Federal Reserve banks. Other regional banks have tight clusters (e.g., ATL, DAL), similar to the FRBNY, but are considerably closer to the clusters of other districts, indicating that their communication is dominated more by contemporaneous views than by past discussions.

Figure 10: t-SNE: Sequence Embeddings for Paragraphs Related to Financial Stability



Notes: The figure visualizes the output of the t-stochastic nearest neighbors (t-SNE) algorithm applied to sequence embeddings for paragraphs that discuss financial stability. We compute embeddings for all documents in the corpus over the period between 1960 and 2021. The sequence embeddings are produced for each paragraph using a RoBERTa sentence transformer model with extended pre-training on abstracts from the S2ORC corpus, as well as fine-tuning on sentence pairs from the S2ORC corpus. Each dot corresponds to a 2-dimensional representation of the average sequence embedding for a given month and institution.

4 Econometric Results

We next move to the empirical tests. We start by exploring structural breaks in the text features in Section 4.1, relating them to macroeconomic and financial events. Sections 4.2 and 4.3 analyze the determinants of financial stability concerns. Section 4.4 explores the association between the text features and financial ratios. Section 4.5 examines whether our text features explain time variation in the historical conduct of monetary policy by conducting extended Taylor rule regressions. Finally, Section 4.6 links our text features to asset valuation-monetary policy regimes.

4.1 Structural Breaks in Text Features

In this subsection, we conduct formal structural break tests using the multiple break test of Bai (1997) and Bai and Perron (1998).[19] The results are presented in Table 1.

[19]The tests are performed separately on each univariate text feature index series using the sequential test option in Eviews with 15% trimming, the HAC option, a degrees of freedom adjustment, and where the error distribution can

Table 1: Structural Breaks for Text Feature Series

Variable	**Break dates**
Dual Mandate	None
Non-Dual Mandate	None
Monetary Policy	1985 Q2
Cosine Similarity Employment and Inflation	1970 Q1 and 2010 Q4
Banking Regulation	1999 Q2
Financial Stability	1996 Q3
Financial Crisis	1969 Q4
Cosine Similarity Monetary Policy and Financial Stability	None
Cosine Similarity Banking Regulation and Financial Stability	1973 Q2, 1983 Q1, 1995 Q1
Cosine Similarity Financial Stability and Inflation	1970 Q1 and 2005 Q4
Academic focus	1997 Q1 and 2010 Q1
Speech Tense: Past Focus	1970 Q1, 1983 Q4, 1997 Q1, 2012 Q4
Speech Tense: Present Focus	1972 Q3, 2000 Q3, 2010 Q3
Speech Tense: Future Focus	1970 Q2, 1996 Q3, 2008 Q3

Notes: All text feature series are quarterly averages of the standardized classification scores from BERT for each paragraph in our text corpus. The tests for structural change are estimated over the full sample (1960-2021) with 15% trimming. The break dates listed represent the first date of the new regime.

Somewhat surprisingly given the more than 60 years of speeches that we analyze, we find no evidence of structural breaks in three of the main features that we extract: the dual mandate feature, the non-dual mandate feature (defined as everything that is not related to the dual mandate), and the cosine similarity between the monetary policy and financial stability features. We expected to find a break in the first two of these features because the dual mandate was only codified via amendment to the Federal Reserve Act in 1977. What this suggests is that well before the official adoption of a dual mandate, communication by Federal Reserve officials seemed to recognize the importance of balancing employment and inflation objectives. Similarly, while the invocation of "unusual and exigent" measures during the 2007-9 global financial crisis may have suggested a new role for the conduct of monetary policy to ensure financial stability, our results indicate that communication surrounding these topics did not experience a significant break during the sample.

For all the other features we consider, the identified breaks coincide with important developments. The earliest identified breaks, which occur in 1969Q4 (financial crisis), 1970 Q1 (cosine similarity between employment and inflation, cosine similarity between financial stability and inflation, and speech tense: past focus) coincide with the recession that began in December 1969 and lasted until November 1970. This period coincided with Fed tightening; hence it is not surprising that the cosine similarities between inflation and both financial stability and employment experience a structural break in 1970Q1, the quarter after the Fed funds rate peaked at 9%. That the textual feature capturing a future focus in speech tense experiences a break in 1970Q2 reinforces the break in the past-focus feature in the previous quarter, coinciding with the end to the recession. There is a second break in the cosine similarity between financial stability and inflation in 2005Q4, the final full quarter

differ across the break subsamples. The maximum number of allowable breaks is set to five.

of the Greenspan chairmanship.[20]

There are two structural breaks that occur during the period leading up to the period of high inflation in the US, one in 1972 Q2 (speech tense: present focus) and one in 1973 Q2 (cosine similarity between banking regulation and financial stability). A few events leading up to the 1973 oil embargo occurred during these dates, with an initial signal of improving relations between Egypt, Saudi Arabia and the US, followed by eventual deterioration, resulting in the embargo.

There are three structural breaks that coincide with the end of the "great inflation" period that lasted from 1965-1982 and the transition into the Great Moderation. The second structural break in the cosine similarity between banking regulation and financial stability is in 1983 Q1, as the US was emerging from the 1980-82 recession. During this period a number of attempts to repeal the 1933 Glass-Steagall Act were introduced, signaling a shift in attitude toward banking regulation, and in 1983 Q4 (the structural break date identified in the speech tense: past focus feature) the International Lending Supervision Act was passed. There were also changes afoot in the conduct of monetary policy, as the Fed moved away from targeting M1 to instead target the quantity of money. This transition is also captured via the 1985 Q2 structural break in the monetary policy feature.

As a result of the Great Moderation and a prolonged period of low inflation, there was little discussion of the price stability side of the dual mandate throughout the 1980s. That changed in the mid-1990s, and the third break in the cosine similarity between banking regulation and financial stability, in 1995Q3, coincides with the first quarter where the "dual mandate" emerges as common parlance. Prior to that quarter, the phrase "dual objective" was used. In addition, two features have breaks in 1996Q3: financial stability has a single break and speech tense: future focus has its second break. During this time (late 1995 and 1996), the Fed was having extensive discussions about changing Fed policy to specify an explicit inflation target, with Vice-Chair Alan Blinder arguing in opposition. The academic focus feature has its first structural break in 1997 Q1, the quarter where the Fed reversed its easing course and implemented a single 25bps hike in the federal funds rate, staying on hold at 5.25% before returning to the easing path in 1998 Q3. The academic focus feature has a second structural break in 2010Q1, the quarter that Ben Bernanke was renewed to a second term as Fed chairman.

The banking regulation feature has a single structural break, in 1999 Q2, coinciding with the passage of the Senate version of what would eventually become the Gramm-Leach-Bliley Act, enacted in 1999 Q4. This act represented the culmination of the various attempts, noted above, to repeal the 1933 Glass-Steagall Act.

The structural breaks identified in the speech tense feature also link to important episodes. The present focus feature has two breaks; the first occurs in 2000 Q3, the quarter after the Fed Funds rate peaked at 6.5%. The second structural break is in 2010Q3. This was the quarter that saw the passage of the Dodd-Frank Act, which codified financial stability as a statutory concern, as well as the European Stability Mechanism and the Fed's second round of quantitative easing (QE2). In addition, according to the Federal Reserve's own history, the September 21, 2010 meeting was the first time a

[20] Although Greenspan served as chairman until January 31, 2006, there are no speeches in the dataset after 2005 Q4.

reference to the employment side of the mandate appeared explicitly in the FOMC statement.[21] It is therefore no surprise that the cosine similarity between inflation and employment exhibits a break in 2010 Q4.

The third break in the future focus feature, 2008Q3, was a critical period as the global financial crisis was unfolding. The last month of this quarter saw the government sponsored enterprises, Fannie Mae and Freddie Mac, go into conservatorship, the collapse of Lehman Brothers and Washington Mutual, and the near-failure of insurer AIG, as well as the September 29th rejection by the House of Representatives of the Emergency Economic Stabilization Act (it passed in October).

Lastly, the fourth break in the past focus, 2012Q4, is the quarter following the FOMC meeting where the Fed's third round of quantitative easing was announced, as well as a prolonging of the expected period that the fed funds rate would remain at the zero lower bound. It might be related to the emergence of non-conventional monetary policy.

Taken together, these structural break results demonstrate that the features appear to capture key episodes in monetary policy, banking regulation and Federal Reserve history, as well as an evolving role for financial stability in those discussions. Yet this descriptive analysis finds no evidence of a break in the cosine similarity between the financial stability and monetary policy features. We further investigate the possibility of this link via a more formal characterization of the determinants of financial stability concerns.

4.2 Empirical Specification

The primary specification is given in Equation (1). We also consider a separate set of specifications that lag each of the macroeconomic and financial variables, but find no substantive differences in the results.[22]

$$y_t = \beta_0 + \beta_1 \tau_{jt}^{fc} + \beta_2 \tau_{jt}^{\pi} + \beta_3 \tau_{jt}^{e} + \beta_4 \tau_{jt}^{bc} + \beta_5 \tau_{jt}^{bl} + \beta_6 \nu_t^{m} + \beta_7 \nu_t^{f} + \zeta_k + \gamma_t + e_{jt} \tag{1}$$

In the first set of regressions, the dependent variable, y_t, is a measure of the financial stability content of a particular paragraph. The explanatory variables include various text features indicated by the variable τ. Such features are computed at the paragraph level and, thus, have both paragraph (j) and time (t) variation. The superscripts on τ indicate what variable or group of controls it encompasses. Additionally, we include district fixed effects (ζ_k) and time fixed effects (γ_t) in some specifications. The following text variables are included in regressions: financial crisis (fc), inflation (π), employment (e), bank capital (bc), and bank liquidity (bl).

The macroeconomic and financial controls are indicated by ν and only have time variation (t). The macro variables (ν^m) include inflation, the output gap, house prices, and the debt-to-gdp ratio. Financial variables (ν^f) include the short term interest rate, a financial crisis indicator, the loan-to-deposit rate, and the natural log of the total loan volume to the non-financial sector. We refer the

[21] See https://www.federalreservehistory.org/essays/humphrey-hawkins-act for documentation and further discussion related to the dates referred to in this paragraph

[22] We omit these results to save space; however, they are available on request.

interested reader to Table A2 and Section 1.3 for details about the macro and financial data definitions and sources.

All regressions use either Newey-West standard errors or cluster at the institution level. We present the results with contemporaneous values of the controls, but do not find any substantive differences if we instead lag them. We also show fixed effect specifications, where non-text controls are dropped and both year-month and institution fixed effects are employed. Finally, the baseline regression spans the full sample (1960-2021), but we also include two sample splits: 1960-1983 (pre-Great Moderation) and 1984-2021 (Great Moderation).

4.3 Regression results

The empirical results are presented in Tables 2-4. In Table 2, the dependent variable is the standardized financial stability index, computed using zero shot learning. In Table 3, it is the measure of cosine similarity between the statement "monetary policy should be used to achieve financial stability" and the content of a paragraph. Table 4 also uses a cosine similarity measure as the dependent variable – between the statement "banking regulation should be used to achieve financial stability" and the content of a paragraph.

Financial Crises and Financial Stability. Our first finding, given in columns (1) to (4) of Table 2, is that a one standard deviation increase in discussion of financial crises is associated with a 0.09 standard deviation increase in discussion of financial stability. Importantly, the result holds even in the presence of macroeconomic and financial controls (column 2), and even when year-month fixed effects are used (column 4). Thus, irrespective of the state of the economy and financial system, an uptick in discussion of financial crises appears to be strongly and statistically significantly associated with discussion of financial stability. This also holds when we lag the macroeconomic controls.[23]

Dual Mandate and Financial Stability. Another important question about financial stability is whether it is typically discussed in the context of the Fed's dual mandate or whether it is treated as a separate concern or objective. In the first four columns of Table 2, the coefficients on the inflation and employment text features are positive and strongly statistically significant in the full sample, indicating that financial stability discussion is positively associated with dual mandate related concerns over the full sample. Furthermore, we can see that the association with financial stability is stronger for employment, where the effect size is roughly three times as large as that for inflation.

Initially, this result may seem at odds with a literature suggesting that aggressive inflation-targeting is conducive to financial stability (Bernanke and Gertler, 2001). However, when the sample is split into two sub-periods (columns (5) and (6)), the relationship between inflation and the discussion of financial stability appears to be negative prior to the Great Moderation, when there were episodes of high inflation and output growth was less stable. In addition, in the first subsample (1960-1983), the employment effect is about 20% stronger than over the full sample. In contrast, during the Great Moderation period, the relationship between inflation and financial stability becomes strongly

[23]The results are available on request.

Table 2: Federal Reserve Speech Focus: Financial Stability

	(1)	(2)	(3)	(4)	(5)	(6)
inflation_{jt} [text]	0.0330***	0.0342***	0.0342***	0.0339***	-0.0321***	0.0639***
	(0.0027)	(0.0027)	(0.0086)	(0.0027)	(0.0043)	(0.0033)
employment_{jt} [text]	0.1256***	0.1253***	0.1253***	0.1261***	0.1504***	0.1139***
	(0.0020)	(0.0020)	(0.0048)	(0.0020)	(0.0035)	(0.0023)
$\text{financial crisis}_{jt}$ [text]	0.0896***	0.0879***	0.0879***	0.0866***	0.0736***	0.0954***
	(0.0027)	(0.0027)	(0.0113)	(0.0027)	(0.0048)	(0.0032)
$\text{bank liquidity}_{jt}$ [text]	0.1506***	0.1466***	0.1466***	0.1444***	0.1683***	0.1351***
	(0.0036)	(0.0036)	(0.0085)	(0.0036)	(0.0063)	(0.0044)
bank capital_{jt} [text]	0.3004***	0.3032***	0.3032***	0.3047***	0.3346***	0.2916***
	(0.0039)	(0.0039)	(0.0126)	(0.0039)	(0.0068)	(0.0047)
past focus_{jt} [text]	-0.1229***	-0.1185***	-0.1185***	-0.1169***	-0.1147***	-0.1194***
	(0.0018)	(0.0018)	(0.0053)	(0.0018)	(0.0031)	(0.0022)
$\text{present focus}_{jt}$ [text]	0.0665***	0.0678***	0.0678***	0.0682***	0.0492***	0.0781***
	(0.0015)	(0.0015)	(0.0055)	(0.0015)	(0.0025)	(0.0019)
future focus_{jt} [text]	0.0890***	0.0882***	0.0882***	0.0881***	0.0746***	0.0917***
	(0.0019)	(0.0019)	(0.0050)	(0.0019)	(0.0033)	(0.0023)
$\text{academic focus}_{jt}$ [text]	0.1673***	0.1639***	0.1639***	0.1632***	0.1726***	0.1596***
	(0.0019)	(0.0019)	(0.0032)	(0.0019)	(0.0037)	(0.0023)
$\text{debt-to-gdp ratio}_{t}$		0.2198***	0.2198**		-0.0638	0.2799***
		(0.0257)	(0.0772)		(0.1695)	(0.0365)
$\text{loan-to-deposit ratio}_{t}$		0.0010**	0.0010		0.0026	0.0032***
		(0.0005)	(0.0018)		(0.0024)	(0.0007)
Sample Period	1960-2021	1960-2021	1960-2021	1960-2021	1960-1983	1984-2021
Macro Controls	NO	YES	YES	NO	YES	YES
Financial Controls	NO	YES	YES	NO	YES	YES
Housing Controls	NO	YES	YES	NO	YES	YES
Month x Year FEs	NO	NO	NO	YES	NO	NO
District FEs	YES	YES	YES	YES	YES	YES
Standard Errors	NW	NW	CL	NW	NW	NW
Adj. R-squared	0.5242	0.5256	0.5256	0.5281	0.5164	0.5310
N	310,425	310,425	310,425	310,425	96,702	213,723

Notes: The dependent variable in all regressions is the zero shot classification score for the term "financial stability." A higher score indicates that the text in a given sequence of words – such as a paragraph or a sentence – describes a topic related to financial stability. All controls that include [*text*] indicate that they are text features measured using zero shot classification. For instance "inflation [text]" is the classification score for whether a sequence describes inflation. The features "past focus," "present focus," and "future focus" classify the tense of a sequence to determine whether the speaker was discussing the past, present, or future. "Academic focus" classifies whether a speaker was discussing an academic debate or the academic literature. Macro controls include the output gap and consumer price inflation. Financial controls include the log of total loans to the non-financial sector, the short term interest rate, and an indicator for whether a financial crisis occurred in a given year. The financial controls, housing controls, debt-to-gdp ratio, and loan-to-deposit ratio are taken from the *Macrohistory Database*, introduced by Jordà et al. (2016). Note that j indexes paragraph and t indexes time. Standard errors are either Newey-West (NW) or clustered (CL). * $p < .1$, ** $p < .05$, *** $p < .01$. Table A.4 in the Appendix presents the results with lagged macroeconomic and financial variables.

positive. In this sub-period inflation targeting also became more prominent at the Federal Reserve, suggesting that, in line with the literature, stabilizing prices and the financial sector are not treated as separate concerns or as being in conflict. In addition, during the Great Moderation, the effect size for employment declines while that for inflation increases.

These findings can also be seen in Figure 11, which shows the coefficient estimates corresponding to the inflation and employment text features from a series of five-year rolling regressions of financial stability discussion on the full set of text, macroeconomic, and financial controls. Consistent with the sample split findings, the Great Moderation appears to have changed how the dual mandate is

Table 3: Federal Reserve Speech Advocacy for the Use of Monetary Policy to Achieve Financial Stability

	(1)	(2)	(3)	(4)	(5)
inflation_{jt} [text]	0.2168***	0.2168***	0.2018***	0.2254***	0.2088***
	(0.0023)	(0.0147)	(0.0023)	(0.0043)	(0.0027)
employment_{jt} [text]	-0.1943***	-0.1943***	-0.1900***	-0.1715***	-0.2067***
	(0.0025)	(0.0129)	(0.0024)	(0.0044)	(0.0030)
$\text{financial crisis}_{jt}$ [text]	0.1041***	0.1041***	0.1035***	0.1236***	0.0986***
	(0.0029)	(0.0111)	(0.0028)	(0.0054)	(0.0035)
$\text{bank liquidity}_{jt}$ [text]	-0.1486***	-0.1486***	-0.1401***	-0.1601***	-0.1376***
	(0.0042)	(0.0129)	(0.0041)	(0.0077)	(0.0049)
bank capital_{jt} [text]	0.0465***	0.0465***	0.0478***	0.0523***	0.0422***
	(0.0041)	(0.0141)	(0.0040)	(0.0075)	(0.0048)
past focus_{jt} [text]	0.0351***	0.0351***	0.0320***	0.0136***	0.0430***
	(0.0024)	(0.0086)	(0.0023)	(0.0042)	(0.0029)
$\text{present focus}_{jt}$ [text]	0.0010	0.0010	0.0024	0.0215***	-0.0114***
	(0.0021)	(0.0065)	(0.0020)	(0.0036)	(0.0026)
future focus_{jt} [text]	0.0617***	0.0617***	0.0585***	0.0413***	0.0696***
	(0.0024)	(0.0072)	(0.0023)	(0.0043)	(0.0028)
$\text{academic focus}_{jt}$ [text]	-0.0691***	-0.0691***	-0.0652***	-0.0923***	-0.0566***
	(0.0025)	(0.0088)	(0.0024)	(0.0050)	(0.0029)
$\text{debt-to-gdp ratio}_t$	-0.0106	-0.0106		1.8618***	-0.2755***
	(0.0509)	(0.5707)		(0.3395)	(0.0748)
$\text{loan-to-deposit ratio}_t$	-0.0091***	-0.0091		0.0124**	-0.0123***
	(0.0010)	(0.0081)		(0.0050)	(0.0014)
Sample Period	1960-2021	1960-2021	1960-2021	1960-1983	1984-2021
Macro Controls	YES	YES	NO	YES	YES
Financial Controls	YES	YES	NO	YES	YES
Housing Controls	YES	YES	NO	YES	YES
Month x Year FEs	NO	NO	YES	NO	NO
District FEs	YES	YES	YES	YES	YES
Standard Errors	NW	CL	NW	NW	NW
Adj. R-squared	0.0893	0.0893	0.1421	0.0969	0.0962
N	310,425	310,425	310,425	96,702	213,723

Notes: The dependent variable in all regressions is the cosine similarity between the contextualized embedding for a given sequence of words and the contextualized embedding that corresponds to the sequence "monetary policy should be used to achieve financial stability." A higher score indicates that a sequence is more likely to be advocating for the use of monetary policy to achieve financial stability. All controls that include [*text*] indicate that they are text features measured using zero shot classification. For instance "inflation [text]" is the classification score for whether a sequence describes inflation. The features "past focus," "present focus," and "future focus" classify the tense of a sequence to determine whether the speaker was discussing the past, present, or future. "Academic focus" classifies whether a speaker was discussing an academic debate or the academic literature. Macro controls include the output gap and consumer price inflation. Financial controls include the log of total loans to the non-financial sector, the short term interest rate, and an indicator for whether a financial crisis occurred in a given year. The financial controls, housing controls, debt-to-gdp ratio, and loan-to-deposit ratio are taken from the *Macrohistory Database*, introduced by Jordà et al. (2016). Note that j indexes paragraph and t indexes time. Standard errors are either Newey-West (NW) or clustered (CL) at the district level. * $p < .1$, ** $p < .05$, *** $p < .01$. Lagging the macroeconomic and financial controls does not substantially change the results.

discussed in conjunction with financial stability. In particular, inflation and financial stability are more likely to be discussed jointly; whereas financial stability and employment are less likely to be discussed together in the same paragraph. This is consistent with the destabilizing role played by inflation and interest rate variability during the 1980s savings and loan crisis, as well as the financial stability concerns in the low inflation-low interest rate environment.

Table 4: Federal Reserve Speech Advocacy for the Use of Banking Regulation to Achieve Financial Stability

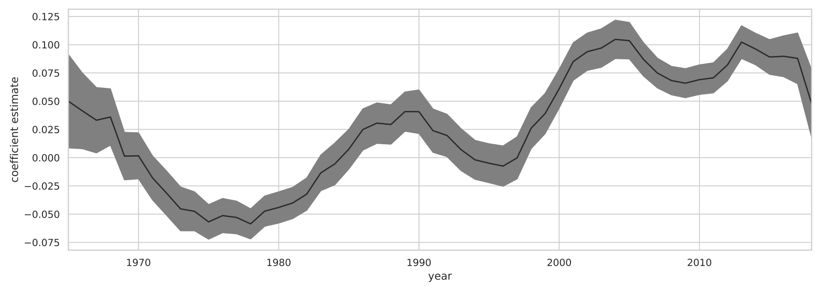
	(1)	(2)	(3)	(4)	(5)
inflation$_{jt}$ [text]	-0.2926***	-0.2926***	-0.2829***	-0.2156***	-0.3182***
	(0.0022)	(0.0201)	(0.0022)	(0.0036)	(0.0027)
employment$_{jt}$ [text]	-0.1952***	-0.1952***	-0.1832***	-0.1703***	-0.2055***
	(0.0022)	(0.0113)	(0.0021)	(0.0036)	(0.0027)
financial crisis$_{jt}$ [text]	0.2049***	0.2049***	0.1990***	0.1266***	0.2372***
	(0.0028)	(0.0124)	(0.0028)	(0.0046)	(0.0034)
bank liquidity$_{jt}$ [text]	-0.1496***	-0.1496***	-0.1392***	-0.1278***	-0.1480***
	(0.0049)	(0.0217)	(0.0048)	(0.0083)	(0.0060)
bank capital$_{jt}$ [text]	0.3697***	0.3697***	0.3488***	0.2920***	0.3876***
	(0.0049)	(0.0186)	(0.0047)	(0.0082)	(0.0059)
past focus$_{jt}$ [text]	0.0002	0.0002	-0.0023	-0.0190***	0.0027
	(0.0022)	(0.0054)	(0.0022)	(0.0036)	(0.0028)
present focus$_{jt}$ [text]	0.0736***	0.0736***	0.0675***	0.0755***	0.0705***
	(0.0019)	(0.0039)	(0.0019)	(0.0031)	(0.0024)
future focus$_{jt}$ [text]	-0.0356***	-0.0356***	-0.0317***	-0.0396***	-0.0321***
	(0.0022)	(0.0044)	(0.0022)	(0.0037)	(0.0027)
academic focus$_{jt}$ [text]	-0.0233***	-0.0233	-0.0210***	-0.0132***	-0.0242***
	(0.0024)	(0.0144)	(0.0024)	(0.0043)	(0.0029)
debt-to-gdp ratio$_t$	0.0716	0.0716		2.3234***	0.5414***
	(0.0453)	(0.2855)		(0.2603)	(0.0657)
loan-to-deposit ratio$_t$	0.0051***	0.0051		0.0043	-0.0026*
	(0.0009)	(0.0073)		(0.0038)	(0.0014)
Sample Period	1960-2021	1960-2021	1960-2021	1960-1983	1984-2021
Macro Controls	YES	YES	NO	YES	YES
Financial Controls	YES	YES	NO	YES	YES
Housing Controls	YES	YES	NO	YES	YES
Month x Year FEs	NO	NO	YES	NO	NO
District FEs	YES	YES	YES	YES	YES
Standard Errors	NW	CL	NW	NW	NW
Adj. R-squared	0.1701	0.1701	0.2043	0.1280	0.1953
N	310,425	310,425	310,425	96,702	213,723

Notes: The dependent variable in all regressions is the cosine similarity between the contextualized embedding for a given sequence of words and the contextualized embedding that corresponds to the sequence "bank regulation should be used to achieve financial stability." A higher score indicates that a sequence is more likely to be advocating for the use of monetary policy to achieve financial stability. All controls that include [*text*] indicate that they are text features measured using zero shot classification. For instance "inflation [text]" is the classification score for whether a sequence describes inflation. The features "past focus," "present focus," and "future focus" classify the tense of a sequence to determine whether the speaker was discussing the past, present, or future. "Academic focus" classifies whether a speaker was discussing an academic debate or the academic literature. Macro controls include the output gap and consumer price inflation. Financial controls include the log of total loans to the non-financial sector, the short term interest rate, and an indicator for whether a financial crisis occurred in a given year. The financial controls, housing controls, debt-to-gdp ratio, and loan-to-deposit ratio are taken from the *Macrohistory Database*, introduced by Jordà et al. (2016). Note that j indexes paragraph and t indexes time. Standard errors are either Newey-West (NW) or clustered (CL) at the district level. * $p < .1$, ** $p < .05$, *** $p < .01$. Table A.4 in the Appendix presents the results with lagged macroeconomic and financial variables.

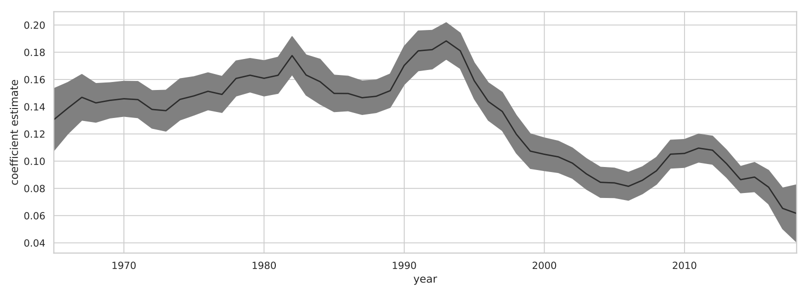
Academic Focus. Next, in Table 2, we find that a focus on academic debates and the academic literature is strongly associated with increased discussion of financial stability. In all specifications, a one standard deviation increase in academic focus is associated with about a 0.16 standard deviation increase in financial stability content, an effect magnitude that exceeds those for either of the components of the Fed's dual mandate. This suggests that, among the topics that the Federal Reserve discusses, the discussion of financial stability appears to have an above average concentration on the

Figure 11: Time Series of Coefficient Estimates from Rolling Regressions

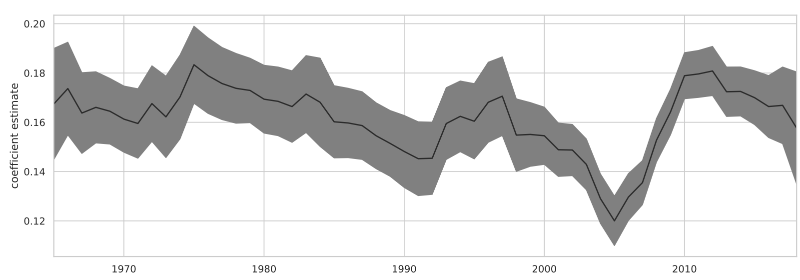
Rolling Regression: Inflation Text Feature Coefficient



Rolling Regression: Employment Text Feature Coefficient



Rolling Regression: Academic Focus Text Feature Coefficient



Notes: The plots above show yearly coefficient estimates from rolling regressions of the financial stability text feature on the inflation, employment, financial crisis, bank liquidity, bank capital, tense focus, and academic focus text features, as well as month-year and district fixed effects. Along with the point estimates, we show 95% confidence intervals. We use a 5-year window for the rolling regressions.

academic discussion, possibly reflecting the need for a conceptual framework.

Table 3 considers the extent to which various text features are discussed in the context of advocating that monetary policy be used to achieve financial stability. The significant negative coefficient on the academic focus variable indicates that Fed speeches that reference the academic literature tend to oppose the use of monetary policy to achieve financial stability (Vollmer, 2021). Table 4 presents similar results in the context of advocating that banking regulation be used to achieve financial stability. Fed speeches that reference the academic literature also tend to oppose the use of banking regulation for this purpose, although the effect is about one-third the magnitude of the effect size for the use of monetary policy.

Comparing the two subsamples across all three tables (columns (5) and (6)) suggests that, although there is little change in the association between academic focus and financial stability overall, the use of academic literature to voice opposition to the use of monetary policy to achieve financial stability appears to have declined during the Great Moderation (relative to the period before the Great Moderation) while the use of it to oppose the use of banking regulation for this purpose has increased. A closer look at how this relationship evolved over time reveals that the influence of academic content over the financial stability discussion increased during the Great Moderation, but declined during Greenspan's tenure, rising again during the Bernanke era. This is illustrated in the third panel of Figure 11, which plots the coefficient estimates corresponding to the academic text feature from a five-year rolling regression.

Speech Tense. Table 2 suggests that statements about financial stability tend to be framed in terms of the present and future, perhaps suggesting that they are in response to ongoing or anticipated events. This tendency increased during the Great Moderation. Advocacy for the use of monetary policy to achieve financial stability typically hinges on the use of past examples and future hypotheticals, as indicated by Table 3; this advocacy increased substantially during the Great Moderation. In contrast, advocacy for the use of bank regulation is typically focused on events in the present, as indicated by Table 4.

Financial Variables and Discussion of Financial Stability. Turning to financial variables in Table 2, we find that an increase in the debt-to-GDP ratio or an increase in total lending appears to be associated with a considerable increase in discussion of financial stability; however, this increased discussion does not necessarily lead to advocacy for a particular approach to policy. Tables 3 and 4 show that a higher debt-to-gdp ratio tends to be associated with increased advocacy of monetary policy and banking regulation as a means of achieving financial stability. The results for the loan-to-deposit ratio, however, are more mixed. Consistent with the view in Meltzer (2009b,a) that the Fed's policy was driven by the forces of political pressure, the association between the advocacy for the use of monetary policy and the public debt-to-gdp ratio was more prominent in Fed communication during the pre-Volcker period when the Fed was subservient to fiscal dominance.

4.4 Financial Markets

We next demonstrate that the concerns articulated in Federal Reserve speeches matter for financial ratios and returns. We do this by estimating a modified version of Equation (1), where the dependent variable is a measure of annual asset returns and the text feature financial stability is introduced as an additional explanatory variable (τ_{jt}^{fs}). Table 5 reports the association between the variables of interest and returns on equity, bonds, risky assets, and safe assets. For sources and return definitions, see Table A2 and Section 1.3.

From a conceptual viewpoint, we expect the contents of Fed speeches to matter not only for asset returns, but also for explaining the difference between risky and safe returns. The impact of Fed communication is likely to occur via the risk-taking channel of monetary policy; through the risk perceptions and risk attitudes of market participants; and through expectations about monetary

Table 5: Federal Reserve Speech Impact on Asset Returns

Return type	(1) Equity	(2) Equity	(3) Bond	(4) Risky	(5) Safe
inflation$_{jt}$[$text$]	-0.0001	-0.0001	0.0019***	-0.0001	0.0010***
	(0.0005)	(0.0005)	(0.0004)	(0.0002)	(0.0002)
employment$_{jt}$[$text$]	0.0025***	0.0025***	0.0012***	0.0018***	0.0006***
	(0.0004)	(0.0004)	(0.0003)	(0.0002)	(0.0002)
financial stability$_{jt}$ [text]	-0.0008*	-0.0007*	-0.0001	-0.0005***	-0.0000
	(0.0004)	(0.0004)	(0.0003)	(0.0002)	(0.0002)
financial crisis$_{jt}$ [text]	-0.0032***	-0.0032***	-0.0052***	-0.0023***	-0.0026***
	(0.0005)	(0.0005)	(0.0004)	(0.0002)	(0.0002)
bank liquidity$_{jt}$ [text]	-0.0032***	-0.0032***	-0.0007	-0.0020***	-0.0004
	(0.0007)	(0.0007)	(0.0005)	(0.0003)	(0.0002)
bank capital$_{jt}$ [text]	0.0031***	0.0031***	-0.0000	0.0015***	-0.0000
	(0.0007)	(0.0007)	(0.0005)	(0.0003)	(0.0003)
past focus$_{jt}$ [text]	0.0003	0.0001	0.0005*	0.0004**	0.0002*
	(0.0004)	(0.0004)	(0.0003)	(0.0002)	(0.0001)
present focus$_{jt}$ [text]	0.0012***	0.0013***	0.0004	0.0006***	0.0002*
	(0.0003)	(0.0004)	(0.0002)	(0.0001)	(0.0001)
future focus$_{jt}$ [text]	0.0017***	0.0018***	-0.0002	0.0008***	-0.0001
	(0.0004)	(0.0004)	(0.0003)	(0.0002)	(0.0001)
past focus$_{jt}$ $*$ financial stability$_{jt}$ [text]		-0.0006*			
		(0.0003)			
present focus$_{jt}$ $*$ financial stability$_{jt}$ [text]		0.0003			
		(0.0004)			
future focus$_{jt}$ $*$ financial stability$_{jt}$ [text]		0.0003			
		(0.0004)			
academic focus$_{jt}$ [text]	-0.0021***	-0.0021***	0.0005*	-0.0006***	0.0003*
	(0.0004)	(0.0004)	(0.0003)	(0.0002)	(0.0002)
debt-to-gdp ratio$_t$	-0.3020***	-0.3020***	0.0172*	-0.1586***	0.0054
	(0.0114)	(0.0113)	(0.0089)	(0.0045)	(0.0044)
loan-to-deposit ratio$_t$	-0.0145***	-0.0145***	0.0056***	-0.0061***	0.0028***
	(0.0002)	(0.0002)	(0.0002)	(0.0001)	(0.0001)
Sample Period	1960-2021	1960-2021	1960-2021	1960-2021	1960-2021
Macro Controls	YES	YES	YES	YES	YES
Financial Controls	YES	YES	YES	YES	YES
Housing Controls	YES	YES	YES	YES	YES
District FEs	YES	YES	YES	YES	YES
Standard Errors	NW	NW	NW	NW	NW
Adj. R-squared	0.1752	0.1752	0.1544	0.1485	0.2196
N	310,425	310,425	310,425	310,425	310,425

Notes: The dependent variable in each regression is a measure of annual asset returns. The returns are taken from Jordà et al. (2016) and include total equity, total bond, risky assets, and safe assets, as specified in the "return type" row of the table. All controls labeled [$text$] indicate that they are text features measured using zero shot classification. For instance "inflation [text]" is the classification score for whether a sequence describes inflation. The features "past focus," "present focus," and "future focus" classify the tense of a sequence to determine whether the speaker was discussing the past, present, or future. "Academic focus" classifies whether a speaker was discussing an academic debate or the academic literature. Macro controls include the output gap and consumer price inflation. Financial controls include the log of total loans to the non-financial sector, the short term interest rate, and an indicator for whether a financial crisis occurred in a given year. The financial controls, housing controls, debt-to-gdp ratio, and loan-to-deposit ratio are taken from the *Macrohistory Database*, introduced by Jordà et al. (2016). Note that j indexes paragraph and t indexes time. Standard errors are either Newey-West (NW) or clustered (CL) at the district level. * $p < .1$, ** $p < .05$, *** $p < .01$. Lagging the macroeconomic and financial controls does not substantially change the results.

conditions, which affect the riskiness of bank lending, valuations, and risk measures (Jiménez et al., 2014; Dell'Ariccia et al., 2017).

We find a positive association between discussion of employment and the returns to equity, bonds, and safe assets. This holds even in the presence of macroeconomic and financial controls, and even when text factors and controls are lagged by one year. Financial stability discussion appears to have a negative association with equity and risky asset returns, but not bond and safe asset returns. In contrast, financial crisis discussion has a negative association for all asset return categories. The same holds for bank liquidity, but the estimated coefficient is insignificant for safe asset returns. Discussion of bank capital has the opposite effect: a positive association with equity and risky asset returns, but insignificant estimated coefficients for bonds and safe asset returns.

4.5 Taylor Rule Regressions

Building on our earlier results, we explore time variation in the historical conduct of monetary policy and whether Fed officials' views as measured through speech features are reflected in their policy decisions. We do this by conducting regressions that extend the Taylor rule (Taylor, 1993, 1999) to include macroeconomic time series data and semantic variables.

Existing work has documented low frequency time variation in the Fed's monetary policy rule (Clarida et al., 2000; Orphanides, 2003; Boivin, 2006; Hamilton et al., 2011; Bianchi et al., 2022; Bauer et al., 2022). Our exercises aim to uncover key drivers of policy rate adjustments by examining the Federal Reserve's speeches. Specifically, we investigate whether non-dual mandate related concerns, as articulated in speeches of Federal Reserve officials, contribute to changes in the policy rule.

4.5.1 Specification of the Monetary Policy Rule

In a recent paper, Carvalho et al. (2021) argue in favor of using ordinary least squares (OLS) to estimate the Taylor rule and document that the endogeneity bias in OLS estimates is small. We also use OLS and first extend the estimation results in Table 2 of Carvalho et al. (2021) to the period ending in 2020 with more recent data. We use the same interest rate rule as in Clarida et al. (2000), which allows for interest rate smoothing. We also estimate versions of an extended Taylor rule, which includes text features from Federal Reserve speeches.[24]

$$r_t = \theta_0 + \theta_{1,1} r_{t-1} + \theta_{1,2} r_{t-2} + \theta_2 E[\pi_{t+1}] + \theta_3 E[x_{t+1}] + \theta_{4,i} \tau_t^i + e_t. \quad (2)$$

The dependent variable r_t is the interest rate. For the subsample prior to 2008, we use the federal funds rate for the interest rate; and for the subsample from 2008 to 2021, when monetary policy was constrained by the zero lower bound, we either use the federal funds rate or the end-of-quarter shadow rate from Wu and Xia (2016).

As explanatory variables, we use lagged values of the interest rate, as well as lagged values of the Greenbook forecast for inflation, $E[\pi_{t+1}]$, and for the output gap, $E[x_{t+1}]$. We use τ_t^i for variables with text features, where the superscripts indicate the specific features used: $i = nd$ refers to non dual mandate, $i = fs$ financial stability, and $i = mf$ refers to advocacy for the use of monetary policy to achieve financial stability.

[24] See Peek et al. (2016) for a study of the ternary mandate using an extended Taylor rule and dictionary-based methods to measure financial stability concerns from FOMC meeting transcripts. In a related paper Istrefi et al. (2021) show that a more negative tone in Federal Reserve speeches on financial stability topics is associated with a more accommodative policy stance.

The choice of including a policy inertia term can be justified by both empirical evidence and theoretical studies (Coibion and Gorodnichenko, 2012). Importantly, the inclusion of interest rate smoothing increases the explanatory power due to an excessive volatility in interest rates predicted by the standard Taylor rule.

Following Carvalho et al. (2021), we define $\rho \equiv \theta_{1,1} + \theta_{1,2}$, $\beta \equiv \theta_2/(1-\rho)$, $\gamma \equiv \theta_3/(1-\rho)$ and $\pi^* \equiv (\theta_0 - (1-\rho)rr^*)/((1-\rho)(1-\beta))$, where ρ is the interest rate smoothing coefficient, β is the coefficient on the inflation gap, γ is the coefficient on the output gap, rr^* is the equilibrium real interest rate and π^* is the inflation target.

4.5.2 Extended Taylor Rule Regressions

Columns (1)-(4) of Table 6 extend the results in Table 2 of Carvalho et al. (2021), which entails the estimation of equation (2) without the inclusion of textual features for the Post-Volcker sample (1986Q1-2020Q4) and for the ZLB subsample (2008Q1-2020Q4). In column (1), we use the federal funds rate throughout the sample. Due to the zero lower bound episode, there is little variation in the dependent variable. As such, we replace it in column (2) of Table 6 with the Wu and Xia (2016) shadow federal funds rate starting in 2008Q1, which also captures the effects of quantitative easing.

We report the estimates for the coefficients β on the inflation gap and γ on the output gap, as well as ρ for the interest rate smoothing and π^* for the inflation target. The coefficient for the inflation gap is highly significant and greater than one for the full sample, indicating that the "Taylor principle" of raising the nominal interest rate more than one-for-one is satisfied, which is an important condition for the existence of a stable inflation rate in macro models. The coefficient for the output gap is also highly significant and close to one. For the ZLB subsamples in columns (3) and (4) the estimates for the coefficient on the inflation gap become insignificant and the estimate for the coefficient on the output gap also becomes insignificant when using the shadow rate. The finding of an insignificant coefficient for the inflation gap for recent years has been established in other work (see, e.g., Bauer et al. (2022)) and may be attributed to the low variation in the Greenbook inflation forecast.

Next, we estimate equation (2), including the text features in the explanatory variable τ_t^i. Column (5) of Table 6 reports the estimated coefficients for the non-dual mandate text feature variable τ_t^{nd} in the post-crisis subsample, which is defined analogously as $\delta^{nd} \equiv \theta^{nd}/(1-\rho)$. We find that more non-dual mandate related talk is negatively associated with the shadow federal funds rate, that is non-dual mandate talk enters the expanded policy rule in an accommodative fashion for the ZLB subsample. In unreported regressions for different pre-ZLB subsamples, the coefficient estimates for δ^{nd} are insignificant.[25] This indicates that the text feature is more important for the ZLB subsample, when the policy discussion was shaped by the experience from the Great Financial Crisis.

To explore the role played by financial stability discussions for the ZLB subsample, we estimate equation (2), including the variables τ_t^{fs} and τ_t^{mf}. The results are reported in columns (6), (7), and (8). We find that the financial stability text feature is associated with more accommodative policy decisions in all three specifications, where the last column includes the lagged text features. The

[25] These regression results are available from the authors on request.

Table 6: Taylor Rule Regressions

	(1) FFR	(2) FFR & Shadow FFR	(3) FFR	(4) Shadow FFR	(5) Shadow FFR	(6) Shadow FFR	(7) Shadow FFR	(8) Shadow FFR
β	1.25**	1.54***	-0.30	-0.63	-1.26	-0.34	-0.56	-0.88
	(0.59)	(0.53)	(0.30)	(0.72)	(0.95)	(0.36)	(0.49)	(0.83)
γ	1.04***	1.30***	0.25**	0.55	0.43	0.37**	0.40**	0.45
	(0.25)	(0.25)	(0.12)	(0.38)	(0.35)	(0.18)	(0.20)	(0.36)
ρ	0.89***	0.91***	0.66***	0.86***	0.88***	0.80***	0.81***	0.85***
	(0.05)	(0.04)	(0.13)	(0.06)	(0.05)	(0.07)	(0.07)	(0.07)
π^*	-0.89	0.30	2.19***	2.41***	1.92***	3.19***	3.06***	2.66***
	(6.10)	(2.55)	(0.56)	(1.09)	(0.88)	(0.64)	(0.74)	(0.95)
non-dual mandate$_t$ [text]					−0.19***			
					(0.07)			
financial stability$_t$ [text]						−0.31**	−0.35***	
						(0.13)	(0.13)	
monetary financial$_t$ [text]							0.12	
							(0.10)	
financial stability$_{t-1}$ [text]								−0.27*
								(0.16)
monetary financial$_{t-1}$ [text]								0.23**
								(0.10)
Sample	1986Q1-2020Q4	1986Q1-2020Q4	2008Q1-2020Q4	2008Q1-2020Q4	2008Q1-2020Q4	2008Q1-2020Q4	2008Q1-2020Q4	2008Q1-2020Q4
Adj. R-squared	0.95	0.96	0.82	0.91	0.92	0.93	0.93	0.92
N	140	140	52	52	52	52	52	52

Notes: We estimate versions of the specification in equation (2). The dependent variable is the end-of-quarter federal funds rate (FFR) and, where indicated, the Wu and Xia (2016) shadow federal funds rate (Shadow FFR) from 2008Q1 onwards. Controls include the lagged values of the interest rate, as well as lagged values of the Greenbook forecast for inflation and for the output gap. All controls labeled [*text*] indicate that they are text features measured using zero shot classification or cosine similarity. The text feature "non dual mandate [text]" is the classification score for whether a sequence is non dual mandate related, the text feature "financial stability [text]" is the classification score for whether a sequence is non dual mandate related, and the text feature "monetary financial [text]" is the classification score for whether a sequence supports the view that financial stability should be achieved using monetary policy. Robust standard errors are reported in parentheses. * $p < .1$, ** $p < .05$, *** $p < .01$.

text feature that measures whether monetary policy should be used to achieve financial stability is associated with less accommodative policy decisions, but the coefficient is only significant when it is included with a lag, which is consistent with a lingering "lean against the wind" rationale.

4.6 Monetary Policy Regimes

Taken together, our text features shed light on the time variation in the historical conduct of monetary policy and help us to identify financial stability-related concerns as an important determinant of that variation. To corroborate our findings, we link our text features with the asset valuation–monetary policy regimes introduced in Bianchi et al. (2022). Specifically, we run a regression with the specification

$$p_t \quad = \quad \theta_0 + \theta_i \tau_t^i + e_t \tag{3}$$

on the sample 1961Q1–2016Q1 of Bianchi et al. (2022), where p_t is the probability of the Fed being in the hawkish monetary policy regime, and τ_t^i contains relevant text features generated from the model. We make use of the *present focus* and *financial crisis* variables to identify whether an ongoing financial crisis is being discussed.

Table 7 presents the empirical results. We find that the hawkish monetary policy regime is significantly associated with an increase in 1) non-dual mandate related text; and 2) discussion about the use of monetary policy to achieve financial stability. On the other hand, a higher fraction of discussions related to either contemporaneous financial crisis developments or discussions about monetary policy and inflation-employment themes are negatively correlated with the hawkish monetary policy stance. Overall, the results are consistent with our extended Taylor rule regression results and the monetary policy regimes in the literature.

Table 7: Monetary Policy Regimes Regressions

	(1)	(2)	(3)	(4)	(5)	(6)
present focus$_t$ $*$ financial crisis$_t$ [text]	-0.085^{**}	-0.091^{***}	-0.092^{***}	-0.085^{**}	-0.091^{***}	-0.092^{***}
	(0.034)	(0.034)	(0.033)	(0.038)	(0.033)	(0.034)
monetary financial$_t$ [text]	0.101^{**}	0.041	0.152^{***}	0.101	0.041	0.152^{**}
	(0.042)	(0.043)	(0.042)	(0.063)	(0.071)	(0.061)
employment inflation$_t$ [text]		-0.001			-0.001	
		(0.046)			(0.077)	
monetary policy$_t$ [text]			-0.198^{***}			-0.198^{**}
			(0.050)			(0.084)
non dual mandate$_t$ [text]	0.097^{**}			0.097		
	(0.044)			(0.064)		
financial stability$_t$ [text]			-0.084^{***}			-0.084
			(0.032)			(0.052)
Robust SE	No	No	No	Yes	Yes	Yes
N	221	221	221	221	221	221

Notes: We estimate versions of the specification in equation (3). The dependent variable is the probability of the Fed being in a hawkish monetary policy regime from Bianchi et al. (2022). All controls are text features measured using either zero shot classification or cosine similarity. The text feature "present focus * financial crisis [text]" is the product of the classification scores for whether a sequence is discussing the financial crisis and whether it is discussing the present; the text feature "monetary financial" is a cosine similarity measure that captures whether a sequence supports the view that financial stability should be achieved using monetary policy; the text feature "non dual mandate" is the classification score for whether a sequence is not related to the dual mandate; and other text features are the classification scores for whether a sequence is related to a particular discussion, including "employment inflation", "monetary policy" and "financial stability". Standard errors are reported in parentheses. * $p < .1$, ** $p < .05$, *** $p < .01$. A value of "Yes" for "Robust SE" indicates that Newey-West standard errors with four lags were used.

5 Conclusion

This paper attempts to characterize how the Federal Reserve system has historically interpreted its institutional mandate and how this interpretation has affected policy decisions and financial markets. We approach this by constructing variables that capture the Fed's latent position about its own mandate. More specifically, we assemble the largest corpus of Fed speeches and apply a collection of large language models to extract a variety of paragraph-level text features. We validate the quality of these features by aggregating them into time series and checking for the presence of structural breaks using Bai and Perron (1998) tests. We find that breaks in the text features we construct align closely with important events in the Fed's history as an institution.

We next partition the speech content into paragraphs that discuss the dual mandate and paragraphs that do not. We find that paragraphs that do not discuss the dual mandate usually discuss financial stability. Other topics, such as climate risk mitigation and economic equality, do not account for a substantial portion of this content, even in the latter half of the sample. This raises the question of whether Federal Reserve officials perceive financial stability to be a third component of its mandate. We find that they do not, even though they sometimes advocate for the use of monetary policy to achieve financial stability. Rather, they typically discuss financial stability as a means of achieving the objectives of the dual mandate or in the context of banking regulation.

We also examine whether the Fed's interpretation of its mandate predicts changes in monetary policy. Estimating an augmented Taylor rule with text features that measure the Fed's latent positions about its mandate, we find that a rise in speech content that is unrelated to the dual mandate is typically associated with more accommodative monetary policy. If we focus specifically on financial stability content, which accounts for much of the speech material that is unrelated to the dual mandate, this association increases in magnitude. In contrast, a rise in advocacy for monetary policy as a means of achieving financial stability is typically associated with tightening, which is consistent with an endorsement of a forward looking "leaning against the wind" view, rather than a "financial instability is caused by monetary tightening" view.

We also show that the Fed's discussion of financial stability affects asset prices. In particular, we find a negative association with asset valuations, even after controlling for macroeconomic and financial variables, including an indicator for financial crises. In a separate exercise, we corroborate this result and its interpretation by demonstrating that financial stability concerns are an important determinant of asset valuation-monetary policy regimes, as defined in Bianchi et al. (2022).

Taken together, these results demonstrate how machine learning can be used in the context of policy evaluation. In the aftermath of the ZLB era, where central banks continue to contemplate the use of communication as an effective policy alternative, natural language processing tools can be employed to shed light on the Federal Reserve's interpretation of its mandate, above and beyond more traditional approaches. By helping economic decision-makers to better understand central banks' objectives, the insights that machine learning tools bring could lead to better-informed decisions, of both policymakers and market participants.

References

Apel, M. and M. Blix Grimaldi, "How Informative are Central Bank Minutes?," *Review of Economics*, 2014, *65.*

Apel, Mikael, Marianna Blix Grimaldi, and Isaiah Hull, "How Much Information Do Monetary Policy Committees Disclose? Evidence from the FOMC's Minutes and Transcripts," *Journal of Money, Credit and Banking*, 2022, *54* (5), 1459–1490.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in Yoshua Bengio and Yann LeCun, eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Bai, Jushan, "Estimating multiple breaks one at a time," *Econometric theory*, 1997, *13* (3), 315–352.

_ **and Pierre Perron**, "Estimating and testing linear models with multiple structural changes," *Econometrica*, 1998, pp. 47–78.

Baker, Scott R., Nicholas Bloom, and Steven J. Davis, "Measuring Economic Policy Uncertainty," *Quarterly Journal of Economics*, 2016, *131*, 1593–1636.

Bauer, Michael D., Carolin E. Pflueger, and Adi Sunderam, "Perceptions about Monetary Policy," *Mimeo*, 2022.

Bernanke, Ben S. and Mark Gertler, "Should Central Banks Respond to Movements in Asset Prices?," *American Economic Review*, 2001, *91* (2), 253–257.

Bianchi, Francesco, "Regime Switches, Agents' Beliefs, and Post-World War II U.S. Macroeconomic Dynamics," *The Review of Economic Studies*, 10 2012, *80* (2), 463–490.

— , **Martin Lettau, and Sydney C Ludvigson**, "Monetary policy and asset valuation," *The Journal of Finance*, 2022, *77* (2), 967–1017.

BIS, *Issues in the Governance of Central Banks*, Bank for International Settlements, 2009.

Boivin, Jean, "Has U.S. Monetary Policy Changed? Evidence from Drifting Coefficients and Real-Time Data," *Journal of Money, Credit and Banking*, 2006, *28* (5), 1149–1173.

Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning, "A large annotated corpus for learning natural language inference," in "Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing" Association for Computational Linguistics Lisbon, Portugal September 2015, pp. 632–642.

Byrne, David, Robert Goodhead, Michael McMahon, and Conor Parle, "The Central Bank Crystal Ball: Temporal information in monetary policy communication," *Mimeo*, 2022.

Carvalho, Carlos, Fernanda Nechio, and Tiago Tristão, "Taylor rule estimation by OLS," *Journal of Monetary Economics*, 2021, *124*, 140–154.

Cer, Daniel et al., "Universal Sentence Encoder for English," in "Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations" Association for Computational Linguistics Brussels, Belgium November 2018, pp. 169–174.

Cho, Kyunghyun et al., "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation," in "Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)" Association for Computational Linguistics Doha, Qatar October 2014, pp. 1724–1734.

Clarida, Richard, Jordi Gali, and Mark Gertler, "Monetary Policy Rules and Macroeconomic Stability: Evidence and Some Theory," *The Quarterly Journal of Economics*, 2000, *115* (1), 147–180.

Coibion, Olivier and Yuriy Gorodnichenko, "Why Are Target Interest Rate Changes so Persistent?," *American Economic Journal: Macroeconomics*, 2012, *4* (4), 126–162.

Conneau, Alexis and Guillaume Lample, "Cross-lingual Language Model Pretraining," in Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2019, pp. 7057–7067.

— , **Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes**, "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data," in "Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing" Association for Computational Linguistics Copenhagen, Denmark September 2017, pp. 670–680.

Correa, R., K. Garud, J.M. Londono, and N. Mislang, "Sentiment in Central Banks' Financial Stability Reports," *Review of Finance*, 2021, *75* (1), 85–120.

Dai, Andrew M. and Quoc V. Le, "Semi-supervised Sequence Learning," in Corinna Cortes, Neil D. Lawrence, Daniel D. Lee, Masashi Sugiyama, and Roman Garnett, eds., *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, 2015, pp. 3079–3087.

Dell'Ariccia, Giovanni, Luc Laeven, and Gustavo A. Suarez, "Bank Leverage and Monetary Policy's Risk-Taking Channel: Evidence from the United States: Bank Leverage and Monetary Policy's Risk-Taking Channel," *The Journal of Finance*, April 2017, *72* (2), 613–654.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Jill Burstein, Christy Doran, and Thamar Solorio, eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics 2019, pp. 4171–4186.

Gentzkow, M., B. Kelly, and M. Taddy, "Text as Data," *Journal of Economic Literature*, 2019, *57.*

Hamilton, James D., Seth Pruitt, and Scott C. Borger, "Estimating the Market-Perceived Monetary Policy Rule," *American Economic Journal: Macroeconomics*, 2011, *3* (3), 1–28.

Hansen, Stephen and Michael McMahon, "Shocking language: Understanding the macroeconomic effects of central bank communication," *Journal of International Economics*, March 2016, *99*, S114–S133.

Hill, Felix, Kyunghyun Cho, and Anna Korhonen, "Learning Distributed Representations of Sentences from Unlabelled Data," in "Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies" Association for Computational Linguistics San Diego, California June 2016, pp. 1367–1377.

Hochreiter, Sepp and Jürgen Schmidhuber, "Long Short-Term Memory," *Neural Computation*, November 1997, *9* (8), 1735–1780.

Howard, Jeremy and Sebastian Ruder, "Universal Language Model Fine-tuning for Text Classification," in "Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)" Association for Computational Linguistics Melbourne, Australia July 2018, pp. 328–339.

Howitt, Peter, "What have central bankers learned from modern macroeconomic theory?," *Journal of Macroeconomics*, 2012, *34* (1), 11–22.

Hull, Isaiah, *Machine Learning for Economics and Finance in TensorFlow 2: Deep Learning Models for Research and Industry*, Berkeley, CA: Apress, 2021.

Istrefi, Klodiana, Florens Odendahl, and Giulia Sestri, "FED Communication on Financial Stability Concerns and Monetary Policy Decisions: Revelations from Speeches," *Banco de España Documentos de Trabajo. N.º 2110*, 2021.

Jiménez, Gabriel, Steven Ongena, José-Luis Peydró, and Jesús Saurina, "Hazardous Times for Monetary Policy: What Do Twenty-Three Million Bank Loans Say About the Effects of Monetary Policy on Credit Risk-Taking?," *Econometrica*, 2014, *82* (2), 463–505.

Kiros, Ryan, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler, "Skip-Thought Vectors," in "Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2" NIPS'15 MIT Press Cambridge, MA, USA 2015, p. 3294–3302.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *CoRR*, 2019, *abs/1907.11692.*

Lo, Kyle, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld, "S2ORC: The Semantic Scholar Open Research Corpus," in "Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics" Association for Computational Linguistics Online July 2020, pp. 4969–4983.

Loughran, T. and B. McDonald, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10Ks," *Journal of Finance*, 2011, *LXVI*, 35–36.

Luong, Thang, Hieu Pham, and Christopher D. Manning, "Effective Approaches to Attention-based Neural Machine Translation," in "Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing" Association for Computational Linguistics Lisbon, Portugal September 2015, pp. 1412–1421.

Mankiw, N. Gregory, "The Macroeconomist as Scientist and Engineer," *Journal of Economic Perspectives*, December 2006, *20* (4), 29–46.

McCann, Bryan, James Bradbury, Caiming Xiong, and Richard Socher, "Learned in Translation: Contextualized Word Vectors," in I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., *Advances in Neural Information Processing Systems*, Vol. 30 Curran Associates, Inc. 2017.

Meltzer, Allan H., *A History of the Federal Reserve*, Vol. 2 of Book 2, 1970–1986, University of Chicago Press, 2009.

— , *A History of the Federal Reserve*, Vol. 2 of Book 1, 1951–1969, University of Chicago Press, 2009.

Nikolsko-Rzhevskyy, Alex, David H. Papell, and Ruxandra Prodan, "Deviations from rules-based policy and their effects," *Journal of Economic Dynamics and Control*, December 2014, *49*, 4–17.

OpenAI, "GPT-4 Technical Report," 2023.

Orphanides, Athanasios, "Monetary Policy Rules Based on Real-Time Data," *American Economic Review*, September 2001, *91* (4), 964–985.

— , "Historical monetary policy analysis and the Taylor rule," *Journal of Monetary Economics*, 2003, *50* (5), 983–1022.

Peek, Joe, Eric S. Rosengren, and Geoffrey M. B. Tootell, "Does Fed policy reveal a ternary mandate?," Working Papers 16-11, Federal Reserve Bank of Boston September 2016.

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer, "Deep Contextualized Word Representations," in Marilyn A. Walker, Heng Ji, and Amanda Stent, eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, Association for Computational Linguistics 2018, pp. 2227–2237.

Pushp, Pushpankar Kumar and Muktabh Mayank Srivastava, "Train Once, Test Anywhere: Zero-Shot Learning for Text Classification," *CoRR*, 2017, *abs/1712.05972.*

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in "Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing" Association for Computational Linguistics Austin, Texas November 2016, pp. 2383–2392.

__ **, Robin Jia, and Percy Liang**, "Know What You Don't Know: Unanswerable Questions for SQuAD," in Iryna Gurevych and Yusuke Miyao, eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, Association for Computational Linguistics 2018, pp. 784–789.

Reimers, Nils and Iryna Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Association for Computational Linguistics 2019, pp. 3980–3990.

Sachs, Jeffrey D., "International Policy Coordination: The Case of the Developing Country Debt Crisis," *Working Paper No. 2287*, June 1987.

Salimans, Alec Radford Karthik Narasimhan Tim and Ilya Sutskever, "Improving Language Understanding with Unsupervised Learning," *Technical Report, Open AI*, 2018.

Schroff, Florian, Dmitry Kalenichenko, and James Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," *CoRR*, 2015, *abs/1503.03832.*

Schularick, Moritz, Lucas ter Steege, and Felix Ward, "Leaning against the Wind and Crisis Risk," *American Economic Review: Insights*, 2021, *3* (2), 199–214.

Sims, Christopher A. and Tao Zha, "Were There Regime Switches in U.S. Monetary Policy?," *American Economic Review*, March 2006, *96* (1), 54–81.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le, "Sequence to Sequence Learning with Neural Networks," in "Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2" NIPS'14 MIT Press Cambridge, MA, USA 2014, p. 3104–3112.

Taylor, John B., "Discretion versus policy rules in practice," *Carnegie-Rochester Conference Series on Public Policy*, 1993, *39*, 195–214.

__ , "A Historical Analysis of Monetary Policy Rules," January 1999.

van der Maaten, Laurens and Geoffrey Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, 2008, *9* (86), 2579–2605.

van Dieijen, Myrthe and Robin L. Lumsdaine, "What Say They About Their Mandate? A Textual Assessment of Federal Reserve Speeches," *Mimeo*, 2019.

Vaswani, Ashish et al., "Attention is All you Need," in Isabelle Guyon et al., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017, pp. 5998–6008.

Vincent, Pascal, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol, "Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion," *J. Mach. Learn. Res.*, December 2010, *11*, 3371–3408.

Vollmer, Uwe, "Monetary policy or macroprudential policies: What can tame the cycles?," *Journal of Economic Surveys (forthcoming)*, 11 2021, pp. 1–29.

Wang, Alex et al., "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," in "7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019" OpenReview.net 2019.

Wang, Kexin, Nils Reimers, and Iryna Gurevych, "TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning," *CoRR*, 2021, *abs/2104.06979.*

Williams, Adina, Nikita Nangia, and Samuel Bowman, "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference," in "Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)" Association for Computational Linguistics New Orleans, Louisiana June 2018, pp. 1112–1122.

Wu, Jing Cynthia and Fan Dora Xia, "Measuring the Macroeconomic Impact of Monetary Policy at the Zero Lower Bound," *Journal of Money, Credit, and Banking*, 2016, *48* (2-3), 253–291.

Wu, Yonghui et al., "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," *CoRR*, 2016, *abs/1609.08144.*

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," in H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, eds., *Advances in Neural Information Processing Systems*, Vol. 32 Curran Associates, Inc. 2019.

Yin, Wenpeng, Jamaal Hay, and Dan Roth, "Benchmarking Zero-shot Text Classification: Datasets, Evaluation and Entailment Approach," in Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Association for Computational Linguistics 2019, pp. 3912–3921.

Yogatama, Dani, Chris Dyer, Wang Ling, and Phil Blunsom, "Generative and Discriminative Text Classification with Recurrent Neural Networks," *CoRR*, 2017, *abs/1703.01898.*

Zhang, Jingqing, Piyawat Lertvittayakumjorn, and Yike Guo, "Integrating Semantic Knowledge to Tackle Zero-shot Text Classification," in Jill Burstein, Christy Doran, and Thamar Solorio, eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, Association for Computational Linguistics 2019, pp. 1031–1040.

Zhu, Yukun et al., "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books," in "2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015" IEEE Computer Society 2015, pp. 19–27.

Òscar Jordà, Moritz Schularick, and Alan M. Taylor, *Macrofinancial History and the New Business Cycle Facts*, University of Chicago Press, May 2016.

A Appendix

Table A1: Most Common Journals in Sentence Pair Corpus

Journal	Article Count
SSRN Electronic Journal	4,164
Journal of Banking and Finance	1,388
Journal of Money, Credit and Banking	618
IMF Working Papers	590
Journal of Finance	585
National Bureau of Economic Research	427
Applied Economics	369
Econometric Reviews	357
Economic Modelling	356
Journal of International Money and Finance	348
Other journals	20,579
Total number of articles	29,781

Notes: This table provides article counts for the 10 most common journals and working paper series in the corpus we construct to train our NLP models. In total, the final corpus includes 283 journals and 29,781 articles. It is a subset of the 2.3M economics articles in the S2ORC corpus (Lo et al., 2020).

Table A2: Macroeconomic and Financial Series

Variable	Description	Source
bond return	total return on government bonds	Jordà et al.
cpi inflation	consumer price index (1990 = 100)	Jordà et al.
crisis	dummy for financial crisis	Jordà et al.
debt-to-gdp ratio	public debt-to-GDP ratio	Jordà et al.
equity return	total return to equity	Jordà et al.
house prices	nominal house prices	Jordà et al.
interest rate	short term interest rate	Jordà et al.
ltd ratio	loan-to-deposit rate	Jordà et al.
output gap	percentage difference between actual and potential (real) GDP	CBO, BEA, FRED
risky return	total return on risky assets	Jordà et al.
safe return	total return on safe assets	Jordà et al.
total loans	total loans originated to non-financial sector	Jordà et al. Jordà et al.

Notes: The output gap variable is taken from the St. Louis Fed's FRED database. It is computed as the percentage difference between actual and potential (real) GDP. The underlying measures of actual and potential GDP are computed by the Bureau of Economic Analysis (BEA) and Congressional Budget Office (CBO), respectively. The remaining macroeconomic and financial variables are taken from the Macrohistory Database (Jordà et al., 2016).

A.1 Sequence-to-Sequence Modeling

Sequence-to-sequence (S2S) modeling poses challenges that are not present in other NLP tasks. Among other problems, dense neural networks require fixed-length inputs and outputs and, thus, cannot be used for S2S tasks, such as machine translation, which requires variation in input and output length. This is true even for a given translated sentence, which may be best expressed using a different number of words in two languages. Consequently, models that are suitable for text classification purposes, such as dense neural networks, are not usable for S2S tasks.

Early breakthroughs in S2S modeling centered around the use of a variant of recurrent neural networks (RNN) called a long-short term memory (LSTM) model (Hochreiter and Schmidhuber, 1997).[26] LSTMs explicitly treat input data, such as words in a sentence, as a sequence, rather than as a set of features.[27] This allows for a more parsimonious parameterization than a dense neural network would permit. LSTMs are also able to handle variable length input sequences, which provides an advantage over dense neural networks for S2S modeling.

The initial innovation in S2S modeling with LSTMs involved the use of an encoder-decoder architecture (Sutskever et al., 2014; Cho et al., 2014). The encoder maps a sequence of symbols to a latent vector, which can be viewed as a compressed representation of the input text. The decoder then maps the latent vector to a sequence of symbol predictions. Since the output sequence must also be permitted to have a variable length in many applications, the model is trained to output an end-of-sentence (`[EOS]`) token, which terminates the sequence of predictions.

LSTM-based models with an encoder-decoder architecture provided the initial means of performing high-quality machine translation and also generated many spillover benefits for related NLP tasks. However, the introduction of the attention mechanism (Bahdanau et al., 2015) and later transformer models (Vaswani et al., 2017) fundamentally changed how sequences are modeled in S2S contexts and in NLP more broadly. This provided a foundation for the set of NLP tools we use in this paper to measure the content of central bank communications.

A.2 The Transformer Model

In this subsection, we provide a detailed overview of transformer models and their advantages over earlier S2S modeling techniques.

Time Complexity. One advantage of the transformers model is that they reduce the computational complexity of certain components of the training process relative to LSTM-based S2S models. In particular, the attention mechanism has a time complexity of $\mathcal{O}(T^2D)$; whereas recurrent operations, such as LSTM cells, have a time complexity of $\mathcal{O}(TD^2)$. This suggests that transformer models will tend to have a training time advantage when embeddings are large.

Another important advantage of transformer models is that they require $\mathcal{O}(1)$ sequential operations; whereas LSTM-based S2S models require $\mathcal{O}(T)$. This implies a substantial training time

[26] See Apel et al. (2022) for an application of LSTM models to natural language processing tasks in economics.

[27] LSTMs also allow for long-term dependence between words in a sequence and correct a version of the vanishing gradient problem. See Hull (2021) for an overview of LSTMs in the context of economics.

advantage for transformer models, since they can parallelize operations that must be performed in sequence for LSTM-based S2S models.

Attention in Transformer Models. Transformer models apply three distinct forms of scaled dot product attention: 1) encoder-decoder attention; 2) encoder self-attention; and 3) decoder self-attention. Encoder-decoder attention uses the query vectors, $\mathcal{Q}$, from the previous decoder layer and key and value vectors, $\mathcal{K}$ and $\mathcal{V}$, from the current encoder layer.

Self-attention, in contrast, is applied to words in the same sequence and in the same encoder or decoder layer. We make a distinction between encoder and decoder self-attention because decoder self-attention only uses the sequence of words up to and including the word being predicted; whereas encoder self-attention uses the entire sequence. Figure 1 in Section 2 illustrates the self-attention mechanism from the RoBERTa model applied to word sequences in the Federal Reserve speeches in our corpus.[28]

The Attention Mechanism Bahdanau et al. (2015) argued that the latent vector in LSTM-based encoder-decoder models created a bottleneck that made it difficult to improve model architecture and training. As a solution to this problem, they proposed using the attention mechanism, which eliminated the need to encode the entire input sequence in a single latent vector. For a given symbol, such as a word, the attention mechanism determines which symbols are related to it without explicitly considering the temporal ordering of the sequence. This allows for symbols that are not close together to be closely related. Luong et al. (2015) demonstrated how this could be used effectively on machine translation tasks.

In practice, attention proved to be an invaluable construct for NLP. We will briefly describe the attention mechanism below in an LSTM encoder-decoder context, focusing on the scaled dot product variant, which was later used in transformer models. For concreteness, consider the passage given in the quote below, taken from a speech given by Gary Stern, then-President of the Minneapolis Federal Reserve Bank, in January 2009. We will first convert the sequence of words to a sequence of embedding vectors, as shown in Equation (4).

> *"Commercial banks have long been thought of, and indeed have functioned as, the backup source of liquidity for many other financial institutions and markets. Banks continue to play this role, but it has become more challenging today to do so because some lenders find themselves capital-constrained as a result of recent losses and/or sizable, unanticipated additions to their balance sheets of formerly off-balance-sheet instruments."*

In Equation (4), M is the embedding dimension and T is the sequence length. The mapping between words and embeddings is constructed outside of the model, typically through a separate training process.[29] In a minimal LSTM model with a single cell, the sequence of embeddings is

[28] (The RoBERTa model is a "robustly optimized" version of the BERT model, which is described in detail below.)

[29] For some models, such as BERT – which we use in most NLP exercises – we will use embeddings based on sub-word units, such as the WordPiece embeddings (Wu et al., 2016). This allows for the use of full words, individual characters, and multi-character strings. The word "growing," for example, can be split into the word "grow" and the multi-character string "ing."

processed in order with each step yielding a hidden state, which is then combined with the next input embedding in the sequence. Equation (5) provides the sequence of hidden states, which have the same dimension as the embedding vector in this architecture.

$$\left\{e_1, e_2, ..., e_t, ..., e_T\right\} = \left\{\begin{bmatrix} e_{11} \\ e_{12} \\ \vdots \\ e_{1M} \end{bmatrix} \begin{bmatrix} e_{21} \\ e_{22} \\ \vdots \\ e_{2M} \end{bmatrix} \cdots \begin{bmatrix} e_{t1} \\ e_{t2} \\ \vdots \\ e_{tM} \end{bmatrix} \cdots \begin{bmatrix} e_{T1} \\ e_{T2} \\ \vdots \\ e_{TM} \end{bmatrix}\right\} \tag{4}$$

$$\left\{h_1, h_2, ..., h_t, ..., h_T\right\} = \left\{\begin{bmatrix} h_{11} \\ h_{12} \\ \vdots \\ h_{1M} \end{bmatrix} \begin{bmatrix} h_{21} \\ h_{22} \\ \vdots \\ h_{2M} \end{bmatrix} \cdots \begin{bmatrix} h_{t1} \\ h_{t2} \\ \vdots \\ h_{tM} \end{bmatrix} \cdots \begin{bmatrix} h_{T1} \\ h_{T2} \\ \vdots \\ h_{TM} \end{bmatrix}\right\} \tag{5}$$

The relationship between the contemporaneous hidden state, h_t, the previous hidden state, h_{t-1}, and the input embedding, e_t, is given by Equation (6). Note that $\mathcal{W}^E$ and $\mathcal{W}^H$ are shape-preserving linear transformations of e_t and h_{t-1} and $\mathcal{G}$ is an elementwise nonlinear activation function.

$$h_t = \mathcal{G}\left(\mathcal{W}^E e_t + \mathcal{W}^H h_{t-1}\right) \tag{6}$$

In a standard encoder-decoder architecture, the terminal state, which we denote h_T, is a fixed-length vector that encodes a summary of the entire sequence. Attention modifies the standard LSTM construction by retaining the hidden states and scoring them. It then applies the softmax function to the hidden states, and then multiplies the softmaxed states by the original (untransformed) states. The procedure introduces three additional sets of trainable weights: $\mathcal{W}^Q$, $\mathcal{W}^K$, and $\mathcal{W}^V$.

For each hidden state h_t, there are associated query (q_t), key (k_t), and value (v_t) vectors, which are computed as $\mathcal{W}^i h_t$, where $i \in \{Q, K, V\}$. Stacking those row vectors into matrices $\mathcal{Q}$, $\mathcal{K}$, and $\mathcal{V}$, we can define scaled dot product attention for h_t in Equation (7), where $D = dim(k_t)$.

$$\mathcal{Z} = \sigma\left(\frac{\mathcal{Q}\mathcal{K}^T}{\sqrt{D}}\right)\mathcal{V} \tag{7}$$

Note that σ is a rowwise softmax function, defined in Equation (8), where X_t is row t in $\mathcal{Q}\mathcal{K}^T/\sqrt{D}$.

$$\sigma(X_{t\hat{d}}) = exp(X_{t\hat{d}}) / \sum_{d \in D} exp(X_{td}) \tag{8}$$

The elementwise product of the row vectors $\mathcal{Q}_t$ and $\mathcal{K}_s$ measures the extent to which they are related to or *attend to* each other. Dividing by $\sqrt{D}$ improves computational performance, but is not strictly necessary. Applying the rowwise softmax function σ amplifies the strength of strong associations and also normalizes the sum of the attention weights to be equal to one.

The weights are then multiplied by $\mathcal{V}$, which is a matrix of linear transformations of the hidden vectors. Each row of the vector $\mathcal{Z}_t$ is a weighted sum of the embedding vectors, where each weight depends on the extent to which a given hidden vector attends to another. Since each hidden vector is most closely associated with a specific embedding in the sequence (i.e., h_t is closest to e_t), attention provides us with a contextualized embedding for each input word. That is, rather than using a fixed embedding, we incorporate the context of other words that are most closely related to it in a sentence.

To be consistent with the subsection that follows on transformer models, we have provided a description of how to compute attention for all hidden vectors. However, for most LSTM-based encoder-decoder models, we will exclusively use $\mathcal{Z}_T$, which is the row vector associated with the final hidden state, h_T. $\mathcal{Z}_T$ is sometimes called the *context* vector and is concatenated with h_T and passed to the decoder. Similar to the earlier variant of S2S models, such as Bahdanau et al. (2015) and Luong et al. (2015), the decoder takes the output of the encoder as an input and then sequentially predicts symbols until it terminates with an (`[END]`) token.

Multi-headed Attention. One innovation of transformer models is to make use of multi-headed attention, which is enabled by the removal of sequential elements from the model. This amounts to an h-way partition of the query, key, and value matrices, such that $\mathcal{W}^j = \{\mathcal{W}_1^j, ..., \mathcal{W}_h^j\}$ for $j \in \{Q, K, V\}$, where $\mathcal{W}_i^j \in \mathbb{R}^{DxD/h}$. For the original transformer model, $M = 512$, $D = 64$, which yields a model with $h = 8$ attention heads. The model then attends to each subspace separately and in parallel, yielding a computational time that is similar to that of a single-headed model.

Positional Encodings. Rather than using sequential elements like recurrence or convolution, transformers modify input embeddings by encoding positional information. For each input embedding in a sequence, positional encodings are generated using Equation (9).

$$p(t,m) \quad = \quad \begin{cases} sin\left(\frac{t}{10000^{2m/M}}\right) & \text{if } m \text{ even} \\ cos\left(\frac{t}{10000^{2m/M}}\right) & \text{if } m \text{ odd} \end{cases} \tag{9}$$

The positional encodings are then added to the input embeddings to create positional embeddings, which contain both information about word features and position within the embedding and sequence. The positional embedding is given in Equation (10) and is the input to the model.

$$\left\{\tilde{e}_1, ..., \tilde{e}_T\right\} \quad = \quad \left\{ \begin{bmatrix} e_{11} + p(1,1) \\ e_{12} + p(1,2) \\ \vdots \\ e_{1M} + p(1,M) \end{bmatrix} \; ... \; \begin{bmatrix} e_{T1} + p(T,1) \\ e_{T2} + p(T,2) \\ \vdots \\ e_{TM} + p(T,M) \end{bmatrix} \right\} \tag{10}$$

A.3 Semantic Textual Similarity

Sentence BERT An alternative to using BERT directly to compute semantic textual similarity is to construct sentence embeddings, which we can compute individually for each passage, and then measure the cosine similarity between pairs of embeddings. This provides a more computationally efficient means of performing this comparison. See Equation (11) for the construction of cosine similarity and note that S is a sentence embedding.

$$sim(S_i, S_j) = \frac{S_i \cdot S_j}{\|S_i\| \|S_j\|} \tag{11}$$

Sentence embeddings have been explored in Kiros et al. (2015), Conneau et al. (2017), and Cer et al. (2018). We will make use of the approach in Reimers and Gurevych (2019), which modifies the pre-trained BERT and RoBERTa models to produce contextualized sentence embeddings, rather than contextualized word embeddings. This approach uses Siamese and triplet networks (Schroff et al., 2015) to train the model, which have objective functions that are comparable to cosine similarity. It is trained using the SNLI (Bowman et al., 2015) and NLI (Williams et al., 2018) datasets.

SBERT Pre-training. The sentence BERT (SBERT) models employed in the paper were pre-trained first by Devlin et al. (2019) and then Reimers and Gurevych (2019). We refine the pre-training using an unsupervised learning process called a Transformer-based Sequential Denoising Auto-Encoder (TSDAE). Specifically, we use the training process in Wang et al. (2021), along with the data described in Section 1, which is compiled from the S2ORC corpus (Lo et al., 2020).

The TSDAE approach to training sentence embeddings was based on earlier work by Vincent et al. (2010) and Hill et al. (2016). The training task entails injecting noise into the input embeddings and then training the model to recover the denoised embeddings. Much like the MLM and TSP tasks for BERT, TSDAE does not require labels, making it an attractive choice for refining the pre-trained model on domain-specific text. It also achieves state-of-the-art results, which approach the performance of supervised methods on domain-specific texts.

SBERT Fine-tuning. In addition to pre-training the SBERT model, we also fine-tune it on the semantic textual similarity (STS) task using pairs of sentences drawn from paper abstracts in the S2ORC corpus (Lo et al., 2020). See Section 1 for an overview of the construction of the dataset.

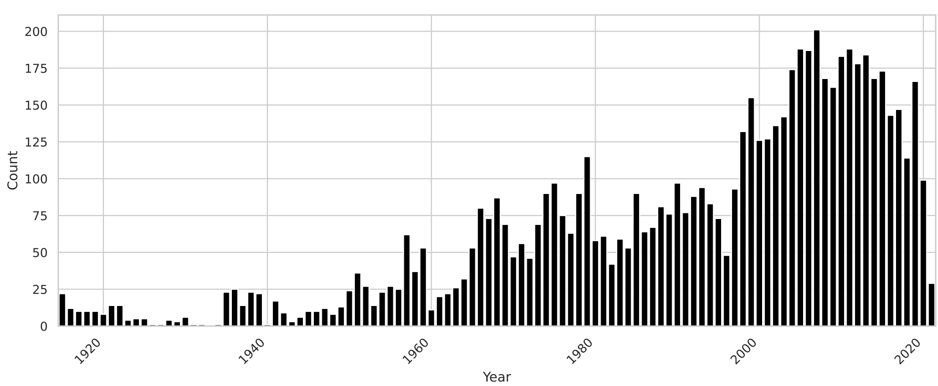
A.4 Text Features

Table A3: Mean Text Feature Values by Federal Reserve District

District	Past Focus	Present Focus	Future Focus	Financial Stability	Academic Focus
ATL	-0.05	0.00	0.05	-0.04	-0.06
BOS	0.15	-0.14	0.29	0.39	0.19
CHI	-0.05	0.00	-0.01	-0.08	-0.09
CLE	0.00	0.00	-0.06	-0.04	-0.05
DAL	0.10	-0.14	-0.03	-0.10	0.01
FRB	0.02	0.01	-0.02	0.02	0.03
KC	0.10	0.03	0.07	0.07	0.03
MIN	0.02	-0.09	-0.05	-0.03	-0.04
NY	0.19	0.16	0.19	0.20	0.23
PHL	0.00	0.00	0.03	0.03	-0.03
RIC	0.09	-0.07	0.22	0.38	0.19
SF	0.08	-0.04	-0.03	-0.13	-0.01
STL	0.06	-0.10	-0.14	-0.04	-0.04

Notes: The table above provides the mean values of selected text features for each Federal Reserve district bank over the entire sample period. All text features are measured at the paragraph level. The first column lists the Federal Reserve district. The next three columns provide measures of tense. Note that tense usage is not mutually exclusive and not all passages have a clear focus on a single tense. As such, it is possible for a given district bank to have positive or negative scores for all three tenses. The remaining two columns provide mean values for features that 1) indicate whether a paragraph is about financial stability; and 2) indicate whether a paragraph references academic work or an academic discussion.

Figure A.1: Federal Reserve Speech Count



Notes: The figure above plots the annual count of Federal Reserve speeches in the corpus. While the corpus spans the period between 1914 and 2020, both coverage and speech frequency increase considerably in the 1960s and again in the 1990s.