Using Natural Language Processing to Identify Monetary Policy Shocks

Alexandra Piller, Marc Schranz and Larissa Schwaller

Working Paper 25.05

Using Natural Language Processing to Identify Monetary Policy Shocks *

Alexandra Piller† Marc Schranz‡ Larissa Schwaller§

October 17, 2025

Abstract

Identifying the causal effects of monetary policy is challenging due to the endogeneity of policy decisions. In recent years, high-frequency monetary policy surprises have become a popular identification strategy. To serve as a valid instrument, monetary policy surprises must be correlated with the true policy shock (relevant) while remaining uncorrelated with other shocks (exogenous). However, market-based monetary policy surprises around Federal Open Market Committee (FOMC) announcements often suffer from weak relevance and endogeneity concerns. This paper explores whether text analysis methods applied to central bank communication can help mitigate these concerns. We adopt two complementary approaches. First, to improve instrument relevance, we extend the dataset of monetary policy surprises from FOMC announcements to policy-relevant speeches by the Federal Reserve Board chair and vice chair. Second, using natural language processing techniques, we predict changes in market expectations from central bank communication, isolating the component of monetary policy surprises driven solely by communication. The resulting language-driven monetary policy surprises exhibit stronger instrument relevance, mitigate endogeneity concerns and produce impulse responses that align with standard macroeconomic theory.

Keywords: FOMC Statements, Central Bank Communication, Monetary Policy Shocks, Proxy SVAR, Machine Learning, Neural Network, Natural Language Processing.

JEL Codes: C45, E52, E58.

*We are grateful to the University of Bern for their financial support. We thank Luca Benati, Martin Brown, Refet S. Gürkaynak, Peter Karadi, Daniel Kaufmann, Costanza Naguib, Galo Nuño, Oliver Pfäuti, Ulf Söderström, Dalibor Stevanovic, Xin Zhang and the Data Science Lab of the University of Bern for their comments. We also greatly benefited from comments from seminar participants at the SNB Research Conference 2025; EFiC 2025; JME-SNB-Gerzensee Conference 2024; Research Seminar at the Riksbank, ESIF Economics and AI+ML Meeting 2024; SSES Annual Congress 2024; Norges Bank Research Workshop: Women in Central Banking 2023; Brown Bag Seminar and Macro PhD Seminar in the Department of Economics at the University of Bern.

†Study Center Gerzensee; alexandra.piller@szgerzensee.ch

‡Department of Economics, University of Bern; marc.schranz@unibe.ch

§Department of Economics, University of Bern; larissa.schwaller@unibe.ch

1 Introduction

Central bankers set policy based on economic projections, while their anticipated decisions in turn shape future outcomes, complicating efforts to identify the causal effects of monetary policy. To address this challenge, economists increasingly rely on data-driven identification methods, to identify monetary policy shocks. High-frequency financial market data has emerged as a valuable tool for this purpose, measuring unanticipated monetary policy changes through price movements in futures contracts within narrow windows around policy announcements (Kuttner, 2001; Gürkaynak et al., 2005b). These price changes are called monetary policy surprises.[1] Gertler and Karadi (2015) extend this approach by combining high-frequency surprises with the Proxy-SVAR framework[2], using such market-based monetary policy surprises as instruments to identify monetary policy shocks.

Despite their widespread use, market-based monetary policy surprises face significant validity concerns. To serve as valid instruments, they must satisfy two conditions: relevance (correlation with the underlying monetary policy shock) and exogeneity (no correlation with other shocks affecting the economy). Recent research has identified weaknesses on both fronts. Ramey (2016) highlights weak instrument problems, while Cieslak and Schrimpf (2019), Miranda-Agrippino and Ricco (2021), Bauer and Swanson (2023a), and Bauer and Swanson (2023b) document substantial endogeneity issues, showing that these market-based surprises are correlated with pre-announcement information on macroeconomic and financial variables.

These validity concerns have emerged alongside a fundamental shift in how monetary policy is conducted. Central bank communication has become an increasingly important aspect of monetary policy, especially since the Great Financial Crisis (Woodford, 2005; Blinder et al., 2008; Gardner et al., 2022; Kerssenfischer and Schmeling, 2024). During the 2008 to 2015 zero lower bound period, traditional policy tools were constrained and policy communication became crucial. FOMC

[1] A positive surprise indicates that the policy announcement shifted the expected path of short-term interest rates upward, serving as a proxy for a contractionary monetary policy shock.

[2] The Proxy-SVAR framework was developed by Stock and Watson (2012) and Mertens and Ravn (2013).

statements became more detailed and Federal Reserve Board chair and vice chair speeches more frequent. Given the Federal Reserve's meticulous language choices, these communications significantly shape market expectations. Consequently, markets respond not only to policy actions but also to the specific language used. This suggests that analyzing the textual content of Federal Reserve communications may offer a way to improve identification of monetary policy shocks.

In this paper, we explore whether text analysis applied to central bank communication can help mitigate the weak instrument and endogeneity problems of market-based monetary policy surprises. In our analysis, we pursue two complementary strategies to address the validity concerns. First, to enhance instrument strength, we extend the dataset of surprises from FOMC announcements to policy-relevant Federal Reserve Board chair and vice chair speeches. To determine policy relevance of speeches, we analyze the language in each speech and retain those that mention both inflation and labor, aligning with the Federal Reserve's dual mandate. Second, we fine-tune a neural network to isolate the component of market-based surprises predictable from central bank communication text, thereby filtering out confounding factors. Our language-driven surprises improve instrument strength, reduce correlation with pre-announcement economic data, and generate impulse responses consistent with standard economic theory.

The first part of our analysis extends the market-based monetary policy surprise dataset by incorporating both FOMC announcements and speeches by the Federal Reserve Board chair and vice chair, following Bauer and Swanson (2023a). Over our sample period, 1996 to 2019, we examine 200 FOMC announcements and 636 speeches. To determine relevance, we use the dictionary from Gardner et al. (2022) to analyze speech content and retain only those that reference both inflation and labor, consistent with the Federal Reserve's dual mandate. This yields 441 policy-relevant speeches by the Federal Reserve Board chair and vice chair.[3] Including these policy-relevant speeches strengthens the instrument and mitigates weak instrument concerns. However, the expanded dataset also produces macroeconomic responses

[3]In comparison, Bauer and Swanson (2023a) use 295 Federal Reserve Board Chair speeches spanning from 1988 to 2019.

that diverge from theoretical expectations. Analogous to other studies, our estimates display the price puzzle[4].

The second part of our analysis uses Natural Language Processing (NLP) to build language-driven monetary policy surprises directly from communication content. We employ the XLNet-Base language model by Yang et al. (2020), a transformer-based architecture (Vaswani et al., 2017), and fine-tune it to predict changes in market expectations from monetary policy communications. We train the model on a random subsample of 178 FOMC statements[5] and 441 policy-relevant speech transcripts, enabling it to extract policy signals directly from text. This approach filters out confounding factors such as market momentum or trader sentiment. Because the texts themselves do not contain raw economic or financial data, the resulting monetary policy surprises are less vulnerable to endogeneity. Language-driven surprises, therefore, provide a cleaner and more reliable tool for identifying monetary policy shocks.

Our findings demonstrate that text-based methods can partly address key shortcomings of market-based monetary policy surprises. We show that endogeneity concerns are more severe for FOMC announcements but less pronounced for Federal Reserve Board chair and vice chair speeches, though both benefit from text-based cleansing. The language-driven monetary policy surprises display markedly lower correlations with pre-announcement information, confirming their improved exogeneity. Monetary policy shocks identified through these measures generate impulse responses that are both economically significant and consistent with conventional economic theories. In particular, when using our language-driven surprises as instruments, we are able to eliminate the price puzzle. Overall, our results show that Federal Reserve communication contains valuable information and is an essential dimension for evaluating causal effects of monetary policy shocks.

[4] The price puzzle refers to the empirical finding that a contractionary monetary policy shock leads to an increase rather than the expected decrease in the price level.

[5] As detailed later, the Federal Reserve began issuing press statements in 1996 when the federal funds rate target changed, extending this practice in May 1999 to all meetings.

Related Literature. Our paper relates to two strands of the literature. First, it contributes to the vast line of research on the identification of monetary policy shocks using high-frequency data (Gürkaynak et al., 2005b; Nakamura and Steinsson, 2018; Cieslak and Schrimpf, 2019; Miranda-Agrippino and Ricco, 2021; Bauer and Swanson, 2023a,b). Second, it contributes to the rapidly expanding research on text analysis in the context of monetary policy.

Bauer and Swanson (2023a) address the weak instrument problem by expanding the set of monetary policy surprises from FOMC announcements to include policy-relevant speeches by the Federal Reserve Board chair or vice chair. They add monetary policy surprises related to post-FOMC press conferences, testimonies to Congress, and speeches by the Federal Reserve Board chair at the Jackson hole symposium. Out of these speeches, they label those as policy relevant that led to a substantial (three basis points or more) reaction in the two-quarter-ahead Eurodollar futures contract and that had moved markets according to their reading of the market commentary in the *The Wall Street Journal* or *New York Times* that afternoon or the following morning. Other studies, such as Jayawickrema and Swanson (2023) and Kerssenfischer and Schmeling (2024), also emphasize the importance of speeches by the Federal Reserve Board chair. Jayawickrema and Swanson (2023) find that speeches by the chair are more important than FOMC announcements for Treasury yields, stock prices, and some interest rate futures. They conclude that including these speeches is key to capturing the primary source of variation in monetary policy. Kerssenfischer and Schmeling (2024) analyze which types of news mainly drive asset prices, finding that chair speeches rank among the most important scheduled releases. Similar to Bauer and Swanson (2023a), they also filter out policy-relevant speeches. However, they employ an automatic approach to identify relevant speeches by counting the number of news reports mentioning each speech.

Our paper builds on previous work by implementing an alternative method for categorizing policy-relevant speeches. Using the dictionary from Gardner et al. (2022), we analyze terminology in Federal Reserve Board chair and vice chair speech transcripts. We classify a speech as policy-relevant if it mentions at least one word related to inflation and one word related to labor – directly reflecting the Federal

Reserve's dual mandate of price stability and maximum employment.

Nakamura and Steinsson (2018) were the first to demonstrate the potential contamination of market-based monetary policy surprises with information effects. They show that Federal Reserve announcements convey information about the central bank's assessment of the economic outlook, leading markets to revise their expectations not only about monetary policy but also about future macroeconomic conditions. This "Fed information effect" suggests that standard monetary policy surprise measures may not be purely exogenous but instead reflect both policy actions and signals about the economy. Miranda-Agrippino and Ricco (2021) further analyze the Fed information effect and propose a way of addressing it. They project market-based monetary policy surprises onto Greenbook forecasts and forecast revisions for real output growth, inflation, and unemployment and back out the residuals to obtain orthogonalized surprises.

Bauer and Swanson (2023a,b) challenge the Fed information effect hypothesis, proposing instead a "Fed response to news" channel as an alternative explanation. This effect still stems from information frictions, but regarding the Federal Reserve's responsiveness rather than economic conditions. Specifically, the public does not know the true response intensity of the Federal Reserve and updates its estimate with every policy communication. Bauer and Swanson (2023a) create an alternative monetary policy surprise series by removing components correlated with pre-announcement economic and financial data. While their series improves upon previous measures, concerns remain that other factors, such as market momentum or trader attitudes, influence federal funds futures (FFF) prices even within narrow windows (Lucca and Moench, 2015; Neuhierl and Weber, 2018). Although their approach includes controls for several economic and financial factors, the set remains limited, leaving open the possibility that other pre-announcement data might still be correlated with the monetary policy surprises.

We contribute to this literature by proposing an alternative approach to refining market-based monetary policy surprises. Recognizing the significant role of language in Federal Reserve communications, we extract the component of existing surprises that is predictable from FOMC statement text or Federal Reserve speech transcripts. By leveraging the relationship between Federal Reserve communication texts and changes in interest rate expectations, we develop a new surprise series that substantially reduces endogeneity issues.

Our paper also connects to the growing literature employing text analysis to construct monetary policy surprise series. For example, Ochs (2021) generates sentiment measures for the FOMC minutes. He uses pre-specified word combinations to which a sentiment class is assigned. In a similar paper, Aruoba and Drechsel (2024) create sentiment measures for FOMC statements using the dictionary by Loughran and McDonald (2011). In contrast to these papers, we employ a transformer-based model to capture relationships and nuances in communication that might otherwise be overlooked and that have become integral to policy implementation.

Doh et al. (2020) construct a new monetary policy surprise measure based on the Universal Sentence Encoder algorithm, designed to capture contextual nuances in FOMC statements. They exploit cross-sectional variations across statements to identify tone and novelty. Handlan (2022) uses the XLNet model to predict intraday changes in FFF contracts from FOMC statement text. To account for the Fed information effect, she additionally cleans her text shocks using alternative statements.

Our approach is closely related to Handlan (2022) but differs by not explicitly correcting for the information effect.[6] Furthermore, we expand our analysis to include speech transcripts by Federal Reserve Board members, broadening our study to encompass more comprehensive central bank communications.

The remainder of this paper is structured as follows. Section 2 describes the data. Section 3 explains the model used to assess the impact of monetary policy shocks on macroeconomic variables. Section 4 extends the dataset of market-based monetary policy surprises by incorporating Federal Reserve Board chair and vice chair speeches, assessing their impact on instrument strength. Section 5 details the text analysis methodology, including the NLP model and its training, and evaluates the language-driven monetary policy surprises. Finally, Section 6 concludes.

[6] As discussed earlier, Bauer and Swanson (2023a) challenge the notion that the Fed information effect is responsible for the endogeneity issues of monetary policy surprises.

2 Data

In our analysis, we utilize three types of data: high-frequency financial data, text data, and monthly macroeconomic data. With the high-frequency financial data, we construct a dataset of market-based monetary policy surprises around FOMC announcements and Federal Reserve Board chair and vice chair speeches. To apply our text analysis approach and derive our language-driven monetary policy surprise series, we match these market-based surprises with the corresponding FOMC statements or speech transcripts. The different monetary policy surprise series are evaluated by using each series as an instrument to identify the monetary policy shock and then assess the shock's impact on a selection of key monthly macroeconomic variables.

It is important to distinguish between the terms we use: "FOMC announcements" or "Federal Reserve Board speeches" refer to the policy communication events, while "FOMC statements" and "speech transcripts" refer to the corresponding text documents published at these events.

2.1 FOMC Announcements

We consider FOMC announcements from January 1996 to December 2019, encompassing eight regularly scheduled meetings per year, typically spaced six to eight weeks apart. Occasionally, the FOMC also holds unscheduled meetings, which occur when unexpected action is required before the next scheduled meeting. We include both types of meetings in our analysis, resulting in a sample of 200 announcements. For the text analysis part of our study, we have to exclude 22 of these announcements as the corresponding statements are not available. The Federal Reserve began consistently publishing a press statement after each meeting starting in May 1999. Prior to that, from 1996 to 1998, the Federal Reserve only issued an explicit statement when there was a change in the federal funds rate target. Thus, our final sample consists of 178 FOMC announcements with press statements. These statements not only communicate the interest rate decision but also provide information on the future economic outlook, forward guidance, and other unconventional policy measures. Over the years, the length of these statements has significantly increased, ranging from approximately 75 to 780 words during our sample period. All statements, including the announcement dates, are from the website of the Federal Reserve Board.

2.2 Federal Reserve Board Chair and Vice Chair Speeches

Building on Jayawickrema and Swanson (2023) and Bauer and Swanson (2023a), we expand the set of policy events beyond FOMC announcements to include speeches by the Federal Reserve Board chair and vice chair. Our sample period aligns with that of the FOMC announcements, spanning from 1996 to 2019. The dataset covers a range of events, such as remarks at the Jackson Hole Economic Symposium, testimonies to Congress, and other chair and vice chair speeches. The number of speeches held each month varies greatly over time, ranging from none to as many as nine. During the Great Financial Crisis, the frequency of speeches was particularly high.

At the annual Jackson Hole Economic Policy Symposium, the Federal Reserve Board chair typically delivers an opening speech to an audience including central bankers, economists, financial market participants, academics, U.S. government representatives, and the media. This speech provides a comprehensive overview of the Federal Reserve's perspectives on the current state of the U.S. and global economies, highlighting key trends and important policy directions. The chair's address often outlines future policy trajectories and the challenges associated with the conduct of monetary policy. During our sample period, the chair delivered 22 speeches at Jackson Hole. However, because precise time stamps are unavailable for eight of these speeches, our dataset includes only 15. These symposium speeches range in length from approximately 1,850 to 7,750 words, reflecting the depth and breadth of the topics covered.

The Federal Reserve Board chair also gives semiannual testimonies to Congress. During these testimonies, the chair provides an overview of the current economic conditions and the rationale behind recent monetary policy decisions. He or she discusses issues such as inflation, employment, and economic growth, and addresses concerns related to financial stability and regulation. The testimony includes an introductory statement followed by a question-and-answer session, allowing for further clarification and discussion. These testimonies aim to ensure accountability and transparency of the Federal Reserve's actions and policies. The testimonies are held twice a year. Each time, the chair presents the testimony once to the Senate and once to the House of Representatives within a few days. Since the introductory statement remains unchanged, we only include the earlier date in our dataset. We assume that the question-and-answer session does not significantly impact interest rate expectations. Moreover, including the question-and-answer part would widen the event window considerably, increasing the risk of capturing effects unrelated to Federal Reserve communications. Given the sample considered, the Federal Reserve Board chair gave 48 testimonies to Congress between 1996 to 2019. However, not all of the transcripts are available, reducing the number to 39. The testimonies contain between 1,200 to 5,700 words.

Additionally, we consider 582 other speeches, out of which 465 were given by the Federal Reserve Board chair and 117 by the vice chair. The length of these speeches varies from around 150 to 20,900 words.

Some policy communication events, such as FOMC announcements or testimonies to Congress, occur in well-defined settings, making it easy to determine when their information reaches financial markets. However, for some speeches, pinpointing this moment is less straightforward. In these cases, we used the timestamps provided on the documents to establish when the speech became publicly available. If no such information was available, the speech was excluded. Additionally, speeches by the Federal Reserve Board chair and vice chair are delivered across various locations in the U.S. and internationally. To ensure consistency, we converted all speech times to U.S. Central Time, aligning with the time zone of the financial market where Eurodollar futures contracts are traded. Similarly, timestamps for FOMC announcements were converted from U.S. Eastern Time to U.S. Central Time. Speech dates and transcripts are from the websites of the Federal Reserve Board and the Federal Reserve Bank of St. Louis.

2.3 High-Frequency Monetary Policy Surprises

To measure shifts in market expectations caused by central bank communication, we extract the high-frequency changes in the price of futures contracts around each announcement or speech. These price changes are often referred to as monetary policy surprises. The rationale for using changes in futures prices is based on the forward-looking nature of financial markets. The FFF market allows participants to hedge against fluctuations in the federal funds rate. On any given day, the FFF market continuously reflects the market's expectations of the average federal funds rate over the remainder of the month. Thus, upward or downward revisions in FFF rates following an FOMC announcement or a Federal Reserve Board chair or vice chair speech indicate that market participants were surprised by the policy announcement and had to adjust their expectations.

As highlighted by Nakamura and Steinsson (2018), financial markets are forward-looking and react only to unexpected components of policy decisions, not to anticipated changes. The construction of monetary policy surprises builds on this idea, measuring intraday price changes in FFF contracts within a narrow time window around Federal Reserve communication events. This approach aims to eliminate reverse causality, ensuring that any observed changes in the FFF rate are attributable solely to the policy announcement rather than any other economic event.

We use Eurodollar futures contracts instead of FFF contracts due to data availability.[7] Nonetheless, Eurodollar futures rates are a reasonable choice. According to Gertler and Karadi (2015), they are the best predictors of future federal funds rate values at horizons beyond six months and are as good as FFF at horizons of less than six months.

We access historical intraday financial market data from Tick Data, LLC, covering Eurodollar futures contracts from December 1981 to June 2023. Eurodollar futures settle based on the spot 90-day Eurodollar deposit rate at expiration, and we focus on contracts that expire approximately one quarter ahead.[8] We convert the

[7]FFF are not available in Tick Data until 2010, while Eurodollar futures are.

[8]Eurodollar futures expire on the International Monetary Market dates: the third Wednesday of March, June, September, and December. We specify the 15th of the expiration month as the

raw data, which reports individual trades, into minute-by-minute data, recording the high and low prices for each minute.[9]

For the FOMC announcements, we follow Gürkaynak et al. (2005b) and measure the change in the Eurodollar futures rate using a 30-minute window, starting 10 minutes before the announcement and ending 20 minutes after.[10] To account for multiple trades within one minute, we use the midpoint between the high and low prices for the minutes marking the beginning and the end of the window. To calculate the monetary policy surprises, we take the difference between the average price at the end of the window and the average price at the beginning of the window, and then multiply this difference by minus one. This scaling is necessary because we want the surprises to reflect changes in interest rate expectations: a decrease in the futures price indicates an increase in interest rate expectations.

For speeches by the Federal Reserve Board chair and vice chair, we consider a time window of 50 minutes, starting 10 minutes before the speech and ending 40 minutes after the speech starts. These communications tend to be more extensive than FOMC statements and contain broader information, which may require investors more time to process. Although some speeches or testimonies can last over an hour, we avoid extending the window too much to minimize the risk of capturing fluctuations in futures rates unrelated to the monetary policy communication. Additionally, the transcript is typically uploaded to the Federal Reserve's website at the start of the speech, providing market participants immediate access to the entire document without the need to listen to the speech in real-time. Thus, we believe the 50-minute window is a reasonable choice. As with the FOMC announcements, we calculate the midpoint between the high and low prices for the minutes marking

roll date, transitioning from the current contract to the next. For example, the first quarter always begins on December 15th. This timing follows the convention in the literature, where the 15th is chosen because the contract expiration date typically falls on a Monday near the middle of the month.

[9] If there is only one trade in a given minute, or if all trades occur at the same price, then the high and low prices for that minute will be identical.

[10] Although the 30-minute window around FOMC announcements has become standard in the literature, Tran (2025) suggests that a longer event window may be more appropriate. For consistency and comparability, we retain the 30-minute window.

the beginning and end of the window and scale the change within the 50-minute time window by minus one.

Some of the speeches partially occur when markets are closed.[11] To address closed markets, we consider three scenarios. First, if the entire speech window falls outside trading hours, we exclude the speech from our dataset as we cannot measure the corresponding change in market expectations. Second, if the speech begins outside trading hours but markets open while the speech is ongoing, we retain the data point if 70 percent[12] of the speech window falls within trading hours. Similarly, if the speech starts during trading hours but ends after markets have closed, we apply the same 70 percent rule, retaining the speech if at least 70 percent of the speech window occurs within trading hours.

Table 1 presents summary statistics for the monetary policy surprises associated with the different types of U.S. monetary policy announcements: FOMC announcements, chair speeches at the Jackson Hole Symposium, chair testimonies to Congress, other chair speeches, and vice chair speeches. The table includes data for monetary policy surprises based on the one-quarter-ahead Eurodollar futures rate (ED2) – the primary focus of our analysis – as well as surprises constructed from current-quarter, two-quarter-ahead, and three-quarter-ahead Eurodollar futures rates (ED1, ED3, and ED4, respectively). First, we observe that the statistics are relatively similar across all four horizons of the futures. The biggest differences are seen for the surprises associated with the current-quarter Eurodollar futures rate. For this very short-term horizon, changes in the futures rate predominantly reflect surprises related to the effective change in the policy rate. For the other horizons, the surprises capture additional elements such as forward guidance. Given our interest in capturing not only the effect of policy rate changes but also other effects transmitted through language, we focus on a different horizon. We have chosen the ED2 for our

[11] Starting from July 2003, Tick Data includes almost around-the-clock electronic trading data, meaning these instances mainly occur in earlier years.

[12] The 70 percent threshold was selected as it provided a balance between ensuring that a large part of the time window fell within trading hours and keeping a majority of the speeches. There are 28 speeches where the time window only partially overlaps with trading hours. Employing the threshold, 12 of these speeches are dropped.

analysis, as it balances the immediate impact of policy decisions with anticipatory elements. Second, the standard deviations and the range of changes (minimum and maximum) indicate that chair speeches and, to a slightly lesser extent, testimonies to Congress are as impactful as FOMC announcements. The other two announcement types, Jackson Hole speeches, and vice chair speeches are considerably less important. Lastly, the mean changes for all five announcement types are close to zero, as expected. FOMC announcements show a slight easing bias of about 1 basis point, but this is relatively small compared to the standard deviations of these changes.

2.4 Macroeconomic Data

When evaluating the effects of FOMC announcements or Federal Reserve Board speeches on macroeconomic variables, we use monthly data on industrial production, the consumer price index, the excess bond premium[13], and the two-year Treasury yield. Industrial production and the consumer price index are taken from the FRED database. The two-year Treasury yield is from Bauer and Swanson (2023a), who took it from the Gürkaynak et al. (2007) database on the Federal Reserve Board's website. The excess bond premium from Gilchrist and Zakrajšek (2012) is available on the Federal Reserve's website. The sample goes from January 1973 to December 2019. The start is determined by the earliest availability of the excess bond premium, while the end is chosen such as to exclude the dramatic swings of the COVID-19 pandemic and its aftermath.

[13] Gilchrist and Zakrajšek (2012) construct a corporate bond credit spread index – the so-called GZ credit spread, which is based on a large micro-level dataset. They then decompose the GZ credit spread into two parts: one part capturing the systematic movements in default risk of individual firms and a residual component – the excess bond premium. The excess bond premium can be interpreted as the variation in the pricing of default risk, meaning it is a measure of the tightness of financial conditions.

Table 1: Summary Statistics for U.S. Monetary Policy Surprises

	FOMC announcements	Jackson Hole speeches	Testimonies to Congress	Chair speeches	Vice chair speeches
Number	200	15	39	465	117
Standard dev. (bp)					
ED1	4.9	0.6	2.3	1.4	0.8
ED2	5.2	1.3	4.2	2.3	1.3
ED3	5.8	2.1	5.8	2.7	1.5
ED4	5.8	2.7	6.4	3.0	1.6
Min. change (bp)					
ED1	-32.5	-2.0	-7.0	-13.0	-3.5
ED2	-27.3	-2.3	-8.3	-17.0	-4.0
ED3	-29.0	-2.5	-9.5	-19.0	-4.0
ED4	-24.0	-3.0	-12.5	-20.5	-4.5
Max. change (bp)					
ED1	18.3	0.8	4.3	6.5	4.3
ED2	12.0	2.5	9.0	18.0	8.0
ED3	17.8	5.0	15.5	22.3	7.8
ED4	24.3	7.3	15.0	26.3	9.3
Mean change (bp)					
ED1	-0.8	-0.1	-0.1	0.0	0.0
ED2	-1.0	0.1	0.5	0.0	0.2
ED3	-1.0	0.2	0.7	0.0	0.2
ED4	-1.0	0.4	0.7	0.0	0.1

Note: Changes for current-quarter, one-, two- and three-quarter-ahead Eurodollar futures rate (ED1, ED2, ED3 and ED4, respectively) are in basis points. Sample period is 1996 to 2019.

3 Proxy-SVAR Methodology

In this paper, we propose enhancements to existing market-based monetary policy surprises and evaluate their effectiveness within the Proxy-SVAR framework established by Gertler and Karadi (2015). As this framework builds the foundation of our analysis, we outline it in the following.

We start by considering the following structural VAR:

$$\mathbf{A}\mathbf{Y}_t = \sum_{j=1}^{p} \mathbf{C}_j \mathbf{Y}_{t-j} + \boldsymbol{\varepsilon}_t, \tag{1}$$

where $\mathbf{Y}_t$ is a vector of observables, $\mathbf{A}$ and $\mathbf{C}_j$ $\forall j \geq 1$ are conformable coefficient matrices, and $\boldsymbol{\varepsilon}_t$ is an $n \times 1$ vector of white noise structural shocks. When multiplying both sides with $\mathbf{A}^{-1}$, the reduced-form VAR representation follows:

$$\mathbf{Y}_t = \sum_{j=1}^{p} \mathbf{B}_j \mathbf{Y}_{t-j} + \mathbf{u}_t, \tag{2}$$

with $\mathbf{u}_t$ being the reduced-form VAR residuals, $\mathbf{B}_j = \mathbf{A}^{-1}\mathbf{C}_j$, and $\mathbb{E}[\mathbf{u}_t\mathbf{u}_t'] = \boldsymbol{\Sigma}$ for some positive definite matrix $\boldsymbol{\Sigma}$. The VAR residuals are modeled as linear combinations of the underlying structural shocks, namely

$$\mathbf{u}_t = \mathbf{S}\boldsymbol{\varepsilon}_t. \tag{3}$$

It follows that $\mathbf{S} = \mathbf{A}^{-1}$ and $\mathbb{E}[\mathbf{u}_t\mathbf{u}_t'] = \mathbb{E}[\mathbf{S}\mathbf{S}'] = \boldsymbol{\Sigma}$.

Let us then define $Y_t^p \in \mathbf{Y}_t$ to be the monetary policy indicator, i.e., the variable for which the exogenous variation is due to the monetary policy shock ε_t^p. To estimate the impulse responses to a monetary policy shock, we need to estimate the equation

$$\mathbf{Y}_t = \sum_{j=1}^{p} \mathbf{B}_j \mathbf{Y}_{t-j} + \mathbf{s}\varepsilon_t^p, \tag{4}$$

where $\mathbf{s}$ is the column of $\mathbf{S}$ associated with the effects of ε_t^p. Because we are only interested in the impulse responses to a monetary policy shock, it is sufficient to identify $\mathbf{s}$ and not the entire matrix $\mathbf{S}$. We use an external instruments strategy to obtain $\mathbf{s}$.

We define $\mathbf{m}_t$ as the $k \times 1$ vector of instruments. $\boldsymbol{\varepsilon}_t^q$ is a vector of structural shocks other than the monetary policy shock. For $\mathbf{m}_t$ to be a valid set of instruments, the exogeneity and relevance conditions must be satisfied:

$$\begin{aligned} &\mathbb{E}[\mathbf{m}_t\varepsilon_t^p] \neq 0 \\ &\mathbb{E}[\mathbf{m}_t(\boldsymbol{\varepsilon}_t^q)'] = \mathbf{0}, \end{aligned} \tag{5}$$

meaning that the instruments are correlated with the monetary policy shock ε_t^p but orthogonal to any other structural shock $\boldsymbol{\varepsilon}_t^q$, where $q \neq p$. In the following application, we will always focus on a single instrument: some version of the monetary policy surprises. Notice that the market-based surprises or the language-based surprises in the next section are all intradaily changes in Eurodollar futures prices. To use them as an instrument in the Proxy-SVAR, we convert them to a monthly series by summing over all the high-frequency surprises within each month.[14]

The identification of $\mathbf{s}$ works as follows: First, we estimate the VAR using least squares estimation and get the reduced-form residuals $\mathbf{u}_t$. These residuals can then be split up into u_t^p, the residual associated with the equation of the policy indicator, and $\mathbf{u}_t^q$, the residuals of all other variables. Moreover, we define $s^p \in \mathbf{s}$ to be the response of u_t^p to a unit increase in ε_t^p. Similarly, $\mathbf{s}^q \in \mathbf{s}$ is the response of $\mathbf{u}_t^q$ to an increase of ε_t^p by one unit. Second, we perform a two-stage least squares regression. In the first stage, we regress u_t^p on the instrument $\mathbf{m}_t$. Consequently, the variation in the fitted value $\hat{u}_t^p$ is only due to the monetary policy shock ε_t^p. In the second stage, we regress $\mathbf{u}_t^q$ on $\hat{u}_t^p$:

$$\mathbf{u}_t^q = \frac{\mathbf{s}^q}{s^p}\hat{u}_t^p + \boldsymbol{\xi}_t. \tag{6}$$

This regression yields a consistent estimate of $\frac{\mathbf{s}^q}{s^p}$ because $\hat{u}_t^p$ is uncorrelated with the error term $\boldsymbol{\xi}_t$. An estimate for s^p can be obtained from the estimated variance-covariance matrix $\boldsymbol{\Sigma}$. In the next step, $\mathbf{s}^q$ can be computed. Based on the estimates of s^p, $\mathbf{s}^q$, and the VAR coefficients ($\mathbf{B}_j$s), we can calculate the impulse responses of all variables in $\mathbf{y}_t$ to a monetary policy shock ε_t^p.

We estimate the VAR using frequentist methods. To obtain confidence bands around the point estimates, we employ bootstrapping methods, with 10,000 bootstrap replications.[15] Moreover, we choose a lag order of $p = 12$. Based on the Ljung-Box Q-test, this lag order is the smallest for which the VAR residuals are no longer serially correlated. Additionally, this lag order is in line with Gertler and

[14] In months for which no surprise occurs, i.e., without FOMC announcements or Federal Reserve Board chair and vice chair speeches, the monthly monetary policy surprise is equal to zero.

[15] We are using the wild bootstrap procedure of Mertens and Ravn (2013) and Gertler and Karadi (2015).

Karadi (2015), Ramey (2016) and Bauer and Swanson (2023a).[16]

4 Market-Based Monetary Policy Surprises

Most existing series of monetary policy surprises focus exclusively on the reactions of market participants to FOMC announcements (Gürkaynak et al., 2005a; Gertler and Karadi, 2015; Miranda-Agrippino and Ricco, 2021; Handlan, 2022). However, several studies have demonstrated that speeches by the Federal Reserve Board chair and vice chair also contain significant policy information that influences interest rate expectations (Bauer and Swanson, 2023a,b; Kerssenfischer and Schmeling, 2024). Additionally, Bauer and Swanson (2023a) show that including monetary policy surprises around Federal Reserve Board chair or vice chair speeches enhances the relevance of these surprises as instruments for identifying monetary policy shocks. Following Bauer and Swanson (2023a), we expand the set of existing market-based surprises to include relevant speeches. To identify the relevant speeches, we utilize a dictionary approach to analyze the content of the speech transcripts and include only those that contain policy-relevant topics.

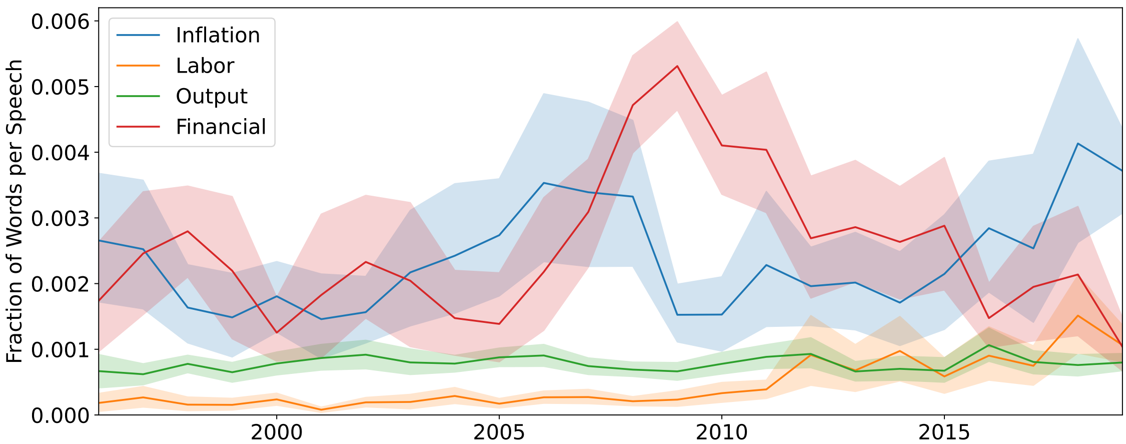
4.1 Identification of Policy-Relevant Speeches

The Federal Reserve Board chair and vice chair deliver speeches on a wide range of topics, many of which extend beyond monetary policy. These can include ceremonial addresses or discussions on subjects such as bank regulation, securities market regulation, fiscal policy, and various other economic and financial issues. As shown by Bertsch et al. (2025), Federal Reserve communication often addresses topics such as financial stability, establishing it as a prominent and recurring theme for these speeches. For our analysis, we focus exclusively on central bank communications that have potential implications for U.S. monetary policy. To identify the speeches relevant to our study, we employ a dictionary-based approach. Specifically, we utilize the dictionary developed by Gardner et al. (2022), which includes lists of words

[16] The results remain qualitatively and quantitatively very similar if we choose a lag order of $p = 6$.

related to inflation, labor, output, and financial topics. For instance, words in the inflation category include “inflation”, “price”, and “cost”, while words in the labor category include “employment”, “job losses”, and “hiring”. We count the occurrences of words related to these topics in each speech.

Figure 1: Chair and Vice Chair Speeches: Categories Sampled by Year



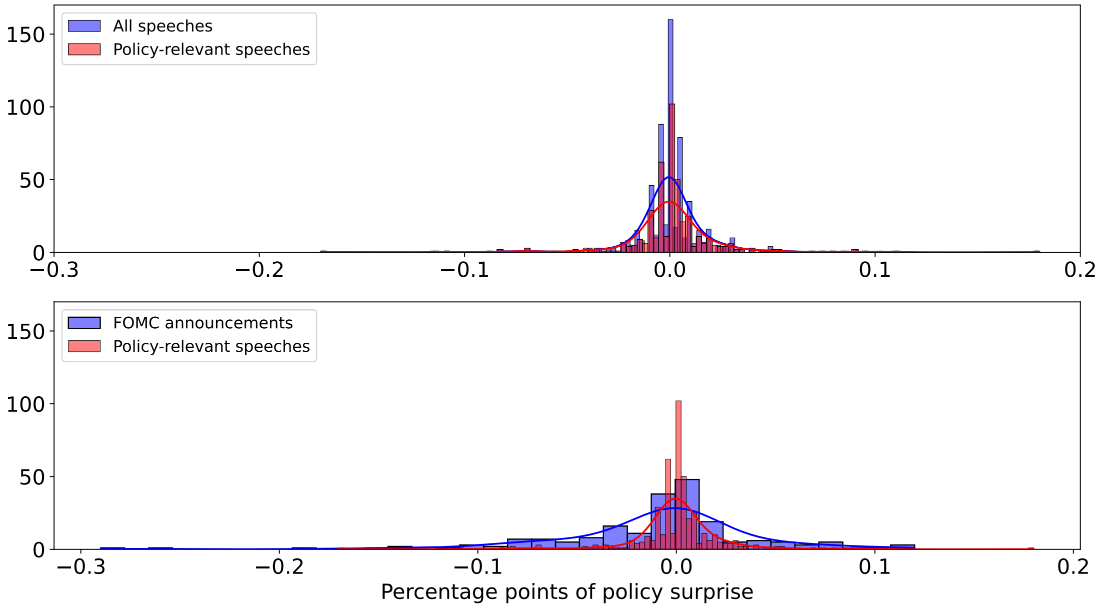
Note: The categories are assigned using the dictionary by Gardner et al. (2022). Observations are sampled by year. The solid lines are the medians per year and the shaded areas show the entire distribution (from minium to maximum fractions).

Figure 1 illustrates the evolution of the average frequency with which words from each category are mentioned per speech transcript. We see that around the Great Financial Crisis, financial topics were discussed more frequently than in other years. Additionally, inflation topics peaked right before the Great Financial Crisis and in 2018, while labor topics appeared more often starting from around 2011.

To determine the relevance of a speech for our application, we focus on the inflation and labor categories. If a speech transcript contains at least one word related to inflation and one word related to labor, we classify it as policy-relevant. Otherwise, it is labeled as non-relevant. Given the Federal Reserve’s dual mandate to promote maximum employment and stable prices, we assume that these two topics are always addressed when the speech pertains to monetary policy. Using these classification criteria, we identify a subset of 441 policy-relevant speeches.

Figure 2 presents a histogram comparing the distribution of monetary policy surprises for policy-relevant Federal Reserve Board chair and vice chair speeches

Figure 2: Distributions of Policy Relevant Federal Reserve Board Chair and Vice Chair Speeches versus All Speeches or FOMC Announcements



Note: Distributions of market-based monetary policy surprises from policy-relevant Federal Reserve Board chair and vice chair speeches (red) against all speeches (top blue) and FOMC announcements (bottom blue).

against all speeches (top) and FOMC announcements (bottom). By excluding non-policy-relevant speeches, we observe a significant decrease in the number of surprises centered around zero. This indicates that our classification method, which relies solely on input text, effectively filters out speeches lacking substantial monetary policy information, thereby refining our dataset to include those speeches that impact market expectations predominantly. The distribution of the policy-relevant speeches is closer to the distribution obtained from the FOMC announcements.

4.2 Monetary Policy Effects on Macroeconomic Variables

Following Bauer and Swanson (2023a), our VAR specification includes the log of industrial production, the log of the consumer price index, the Gilchrist and Zakrajšek (2012) excess bond premium, and the two-year Treasury yield. We include the excess bond premium because Caldara and Herbst (2019) finds it to be necessary to identify monetary policy shocks correctly. Furthermore, as discussed in Gertler and

Karadi (2015) and Bauer and Swanson (2023a), we use the two-year Treasury yield instead of the federal funds rate as the policy rate variable.[17] Unlike the federal funds rate, the two-year Treasury yield was largely unconstrained during the U.S. zero lower bound period from 2009 to 2015, making it a better measure of the stance of monetary policy. Moreover, an important advantage of using a government bond rate as the policy indicator is that its innovations do not only capture traditional monetary policy shocks, i.e., monetary policy surprises related to the current federal funds rate, but also shocks to forward guidance. Swanson and Williams (2014) and Hanson and Stein (2015) argue that the Federal Reserve's forward guidance strategy operates with a roughly two-year horizon, which makes the two-year Treasury yield the preferred government bond rate.[18] Additionally, the speeches do not convey information about changes in the current federal funds rate; rather, they influence market participants' expectations regarding future policy rate changes. Hence, employing the two-year Treasury yield in conjunction with our expanded market-based surprises as an instrument is a natural choice.

We identify the monetary policy shock using three different surprise series as instruments. The first series includes only FOMC announcements. The second series expands to include monetary policy surprises from both FOMC announcements and all speeches by the Federal Reserve Board chair and vice chair. The third series refines upon the second by incorporating only those speeches labeled as policy relevant.

Table 2 reports the robust F-statistics from the first-stage regression for each of our three instrument specifications. The specification using only FOMC announcements yields an F-statistic of 2.90, well below conventional thresholds for strong instruments. Expanding our instrument set to include all speeches by the Federal Reserve Board chair and vice chair substantially improves identification strength, raising the F-statistic to 6.21. Our preferred specification, which refines this ap-

[17] Although Gertler and Karadi (2015) advocate for the two-year Treasury yield, they use the one-year Treasury yield in their VAR due to an insufficiently large F-statistic for their first-stage instrumental variables regression with the two-year yield as the policy indicator.

[18] The Federal Reserve's forward guidance strategy, focusing on managing expectations of the path of the short rate two years into the future, supports the use of the two-year Treasury yield.

Table 2: Robust F-statistics for market-based monetary policy surprises

	F-statistic
FOMC announcements	2.90
FOMC announcements and *all* speeches	6.21
FOMC announcements and *policy-relevant* speeches	7.73

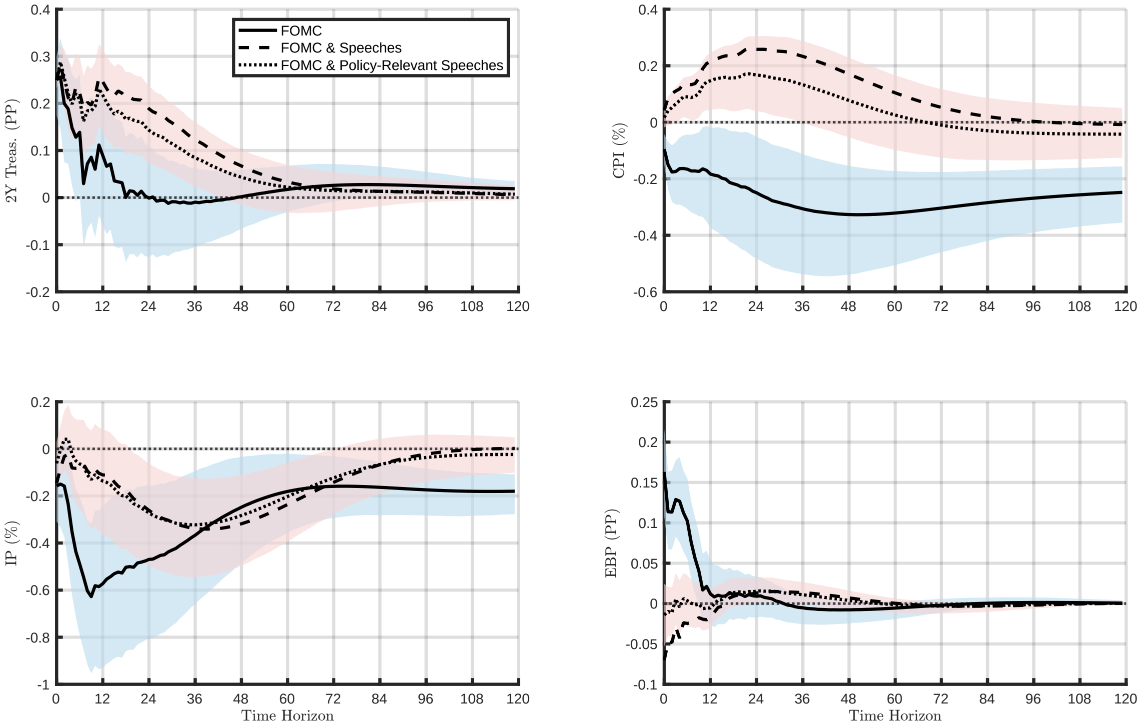
proach by incorporating only policy-relevant speeches alongside FOMC announcements, further strengthens identification with an F-statistic of 7.73.[19] These results demonstrate that carefully broadening the set of market-based surprises significantly enhances instrument relevance, though we acknowledge that our preferred specification still falls somewhat short of the conventional threshold for strong instruments.

Figure 3 depicts the corresponding impulse responses, normalized to reflect a 25 basis point increase in the two-year Treasury yield. The responses are shown for the same three sets of monetary policy surprises discussed above. The impulse responses obtained with the surprise series that include speeches differ markedly from those based solely on FOMC announcement. The latter responses align with conventional wisdom: a monetary policy shock leads to a decline in the consumer price index and a contraction in industrial production. Additionally, the excess bond premium rises, indicating tighter financial conditions. In contrast, the impulse responses derived from the other two surprise series, which contain speeches, reveal notable deviations. The consumer price index responds positively, exhibiting the so-called price puzzle. While industrial production still decreases, as expected, the excess bond premium shows little to no significant reaction, a finding that contradicts traditional economic theories.

To sum up, we find that adding Federal Reserve Board chair and vice chair speeches to the surprise dataset improves the F-statistic substantially. However, the dynamic responses to a monetary policy shock also change. When considering speeches in the instrument series, the identified monetary policy shock gives rise

[19] Stock and Watson (2012) propose a rule of thumb that instruments are considered weak when the first-stage F-statistic in two-stage least squares regression falls below ten.

Figure 3: Impulse Responses to a Monetary Policy Shock (Market-Based Surprises)



Note: The blue shaded areas represent the 5-95 percentiles of the impulse responses identified with only FOMC announcements and the red shaded areas the 5-95 percentiles of the impulse responses identified with *policy-relevant* speeches. The impulse responses are normalized to a 25 basis point increase in the two-year Treasury yield (2Y Treas.) and show the reaction of the consumer price index (CPI), industrial production (IP) and the excess bond premium (EBP). Horizontal axis: time horizon in months.

to a price puzzle. This finding is counterintuitive according to standard economic theory. Hence, we need further refinement.

5 Language-Based Monetary Policy Surprises

In this section, we focus even further on central bank communication. Advances in NLP enable us to directly link central bank statements and speech transcripts to financial market reactions. In particular, we map policy communication texts to market-based monetary policy surprises, isolating the portion of the surprises driven exclusively by FOMC statements and speech transcripts[20]. This approach

[20] Appendix D presents the results only using FOMC statements for training.

allows us to construct language-based surprises that capture solely the impact of central bank communication while abstracting from other influencing factors, such as market momentum or trader sentiment. To validate their exogeneity, we test whether these surprises are predictable from past economic information.

5.1 Natural Language Processing

To construct the language-based monetary policy surprises, we follow a four-step process. First, we select a pre-trained neural network capable of understanding English. We opt for the XLNet-Base model. Second, we pre-process the central bank communication texts to create structured inputs for our NLP model. Third, we fine-tune the pre-trained model on our specific task. That is, we train the model to understand the relationship between central bank communication texts and market-based high-frequency surprises. Finally, using the trained model, we predict the changes in market expectations associated with each FOMC statement or speech transcript. These predictions are what we call the language-driven monetary policy surprises.

5.1.1 Select Pre-Trained Neural Network for Text Processing

To capture the semantics of central bank communications, we use XLNet-Base, a Natural Language Understanding algorithm developed by Yang et al. (2020).[21] XLNet combines two techniques – autoregressive language modeling and auto-encoding – to learn textual content. Both methods involve predicting missing words, but while autoregressive modeling predicts words at the end of a sequence, auto-encoding predicts missing words from anywhere within a sentence. Consequently, XLNet not only understands individual words but also captures sentence structure and longer textual contexts. Because XLNet learns representations through different approaches, it is highly versatile, making it suitable for tasks beyond simple word prediction. The model can be fine-tuned to map text to other text, categories, or continuous numbers. In our case, we want to teach the model to link central bank communica-

[21] The base model consists of 12 attention heads/layers with 768 dimensions and two feed-forward layers with 768 and 3072 dimensions, resulting in approximately 117 million parameters.

tion texts to changes in market expectations, which is a continuous financial market variable. We start from the base version's pre-trained network architecture and word representations, which have been trained on a vast corpus of text to acquire general English language skills.[22] Such off-the-shelf pre-trained models are widely used in NLP because training them from scratch requires substantial computational power and vast amounts of text data.

A legitimate concern when employing off-the-shelf models for text analysis is lookahead bias. As described by Sarkar and Vafa (2024), lookahead bias arises when a model unintentionally uses information from the future to predict outcomes in the past. However, we argue that this issue is mitigated in our setting for two key reasons. First, Yang et al. (2020) trained their model on a restricted corpus of text data, yet demonstrated that this limitation did not compromise the model's generalizability relative to comparable architectures. Second, the monetary policy surprises used to fine-tune our neural network are proprietary and not publicly accessible, making it highly unlikely that the model has been previously exposed to this information.

5.1.2 Pre-Processing Text Data

In the pre-processing step, we format the text of FOMC statements and speech transcripts in such a way that the language model can process it. For the FOMC statements, we make only minimal modifications to preserve the original wording. We replace long word combinations with abbreviations[23], standardize the formatting of numbers, and remove repetitive words, dates, and committee member names to prevent the model from drawing incorrect conclusions during training. A full list of modifications is provided in Appendix A.

The speech transcripts require additional adjustments. As described in Section 2, they tend to be significantly longer than FOMC statements, often spanning multiple pages. While neural networks can process long text inputs, their performance

[22]The model was trained on a text corpus from five sources: The Book Corpus and English Wikipedia (13GB), Giga5 text (16GB), Clue Web 2012-B (19GB), and a Common Crawl (110GB).

[23]We only use abbreviations that the Federal Reserve itself uses in at least one of the other statements.

can deteriorate when exceeding a certain length – especially for smaller models like XLNet, which were trained on short sequences. To address this, we need to reduce the length of the speech transcripts. One possibility is to truncate each speech transcript at the model's recommended token[24] length, which is 512 tokens for the XLNet model. However, this approach risks omitting important information mentioned toward the end of the speech. Instead, we opt for summarization, ensuring that the most relevant content is preserved while keeping the input length manageable.

Summarization helps discipline the model by directing its focus toward the policy-relevant parts of a speech transcript. Unlike FOMC statements, which are short, well-structured, and carefully worded, speeches exhibit greater variability in both structure and phrasing. As a result, market participants are likely to focus on the key takeaways rather than the precise wording. By summarizing speech transcripts, we replicate this process, ensuring that our dataset better reflects how financial markets extract and interpret central bank communication. Furthermore, given the varying length and structure of speeches, summarization enhances consistency, reduces noise, and improves the model's ability to learn from these text documents.

We use the Mistral Large 2 model, a state-of-the-art language model with 123 billion parameters, for the summarization task. This model can process inputs of up to 128,000 tokens and outperformed alternative approaches,[25] making it the most suitable choice for our application. To ensure that the most relevant content is preserved, we instruct the model to: generate fluent, first-person summaries that maintain the speaker's voice rather than third-person bullet points; and (ii) focus on monetary policy topics, using a predefined dictionary from Gardner et al. (2022) along with additional key terms such as accommodative, contractionary, stance, and federal funds rate. Each summary is limited to a maximum of 15 sentences. In the end, we pre-process every summary in the same way as we pre-processed the FOMC statements.

[24]Tokens are the input feed to the language model. They capture the text and its words, where, as a rule of thumb, one token corresponds to 4 characters on average.

[25]We also tested Falcon-7B-Instruct, GPT-4, BART fine-tuned on CNN/Daily Mail, and a smaller variant of T5 fine-tuned for summarization.

5.1.3 Fine-Tune the Neural Network

As described earlier, every NLP model is associated with a specific vocabulary, a collection of distinct words, it was initially trained on. Through this initial training, the XLNet model has already gained an understanding of the words and their relationships among each other. In more technical terms, as part of the training, the words are converted into tokens so a computer can easily process them. The XLNet model has already mapped the tokens in an N-dimensional space where similar tokens lie closer together. For example, the tokens for apples and oranges lie closer together than the tokens for apples and inflation. To use the model by Yang et al. (2020), we convert the FOMC statements and the summaries of the speech transcripts into tokens used in the XLNet vocabulary.

Moreover, we want our model to learn the mapping from FOMC statements and speech transcripts to marked-based surprises. The objective is for the model to predict the monetary policy surprises based on an unknown statement or speech. The model provided by Yang et al. (2020) can not yet execute this task. Thus, we add layers to their neural network structure that are suitable for obtaining continuous predictions. Specifically, we break down the text using convolutional layers so that the model can extract the relevant information and appropriately predict the monetary policy surprises.[26] Appendix B presents our model architecture, and Appendix C explains the training algorithm in further detail.

As it is standard in the machine learning literature, we apply k-fold cross-validation. We split our dataset, consisting of the FOMC statements and speech transcripts, together with the respective marked-based surprise, into different parts. Notably, we always have a training set that the model adapts its parameters on and a test set that it runs the model on but does not adapt its parameters to. Such train and test splits are important in machine learning because the models are prone to overfitting, i.e., to learn too much from the training data, thereby being unable to predict data it has not yet seen. To train our model, we apply five-fold cross-

[26]This procedure adds roughly 4 million parameters to the model. Thus, the final model counts around 221 million parameters. Training these additional parameters typically requires around three days, though the exact duration may vary depending on the specifications.

validation. Thus, we always have 80 percent of our data in the training set and 20 percent in the test set. This procedure also implies that for the same hyperparameter described in Appendix C, we have five different parameter values, depending on the data the model was trained on.

To find the optimal hyperparameters for our model, we experiment with different combinations of the number of epochs[27] and the learning rate.[28] During these experiments, we monitor the mean squared error (MSE) on the training (in-sample) and test (out-of-sample) data. The MSE is the mean squared difference between the prediction and the corresponding true marked-based surprise. Based on these test runs, we fix the learning rate of our model to 1e-5 and the number of epochs to 10. The MSE for this set of hyperparameters for the five-fold cross-validation is displayed in Table 3.

Table 3: MSE of 5-Fold Cross-Validation after 10 Epochs

Number of Split	In-Sample MSE	Out-of-Sample MSE
1	5.040045e-05	0.0023990888
2	0.00016885748	0.0015901675
3	0.00068839284	0.0011396597
4	0.0014133948	0.0010364184
5	0.00012260459	0.0022924726

5.1.4 Predict Surprises using Text Data

The predicted monetary policy surprises for each FOMC statement or speech transcript are generated using the model parameters from the first split of the cross-validation after ten epochs, as described above.

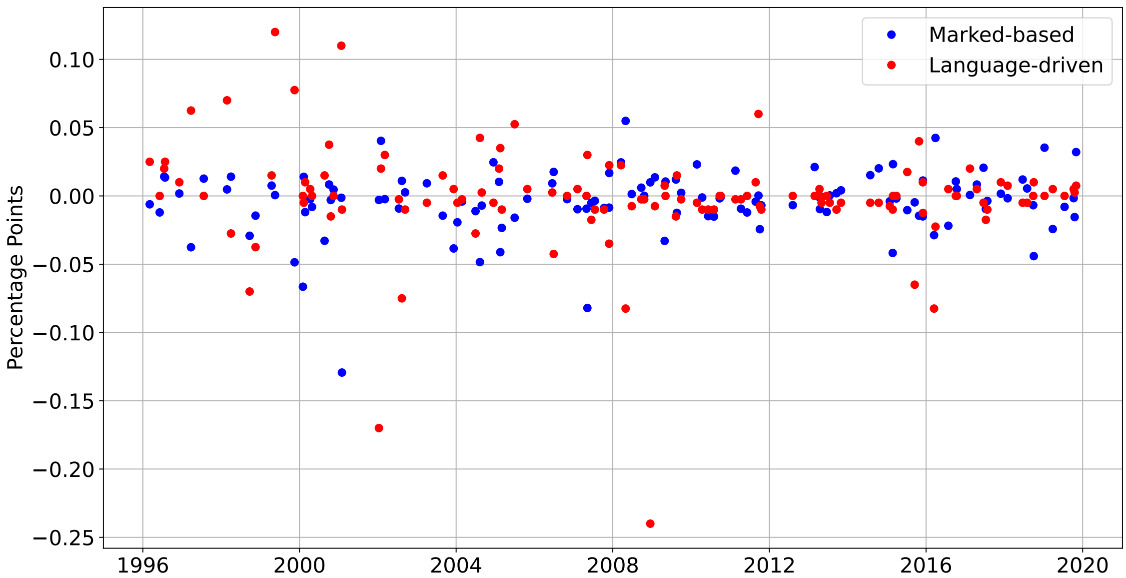
Figure 4 presents the out-of-sample MSE for this split, focusing on the test set of FOMC statements and speech transcripts. While the model captures changes in

[27] The number of epochs defines how often the model sees the same data set to adapt its parameters.

[28] The learning rate defines by how much the neural network adapts its parameters after each iteration.

market-based surprises to some extent, it does not achieve a perfect fit. This result is unsurprising since the number of data points is small for a machine-learning task. However, considering the limited number of available texts, along with their length and complexity, the model performs remarkably well on the test set. Moreover, market-based surprises are expected to be influenced not only by the content of central bank communications but also by other factors such as market momentum and trader sentiment. As a result, some degree of deviation between predicted and actual monetary policy surprises is both expected and desirable. Overall, the results indicate that the model successfully learns patterns from the training data that allow it to predict market reactions. In Appendix G, we dive deeper into the model and check some of the predictions. We want to understand how much each text passage contributes to a particular prediction. This check is necessary to confirm that our model learned the patterns correctly and can interpret the statements accurately.

Figure 4: Predictions and Market-Based Surprises



Note: Market-based surprises and out-of-sample model predictions of the language-driven surprises are displayed in blue and in red, respectively.

5.2 Predictability with Economic and Financial Data

The primary objective of our NLP task is to isolate the component of market-based monetary policy surprises that stems solely from central bank communication. By doing so, we aim to remove the correlation between market-based surprises and past economic and financial data. To verify whether our language-based surprises are indeed uncorrelated to economic and financial information available before an announcement or speech, we conduct a regression analysis following Bauer and Swanson (2023a).

To capture past economic and financial conditions, we construct the following variables: (1) the most recent nonfarm payroll surprises (NFP_SURP), (2) the 12-month employment growth in total nonfarm payrolls (NFP_12M), (3) the three-month growth in the S&P 500 stock market index (SP500_3M), (4) the three-month change in the slope of the yield curve (SLOPE_3M), (5) the three-month growth in the Bloomberg Commodity Spot Price index (BCOM_3M), and (6) the average skewness of the ten-year Treasury yield over the past month (TR_SKEW).[29] Except for nonfarm payroll surprises, constructing these variables is straightforward; further details can be found in Bauer and Swanson (2023a). For nonfarm payroll surprises, we take the difference between the actual nonfarm payroll release and the median forecast from a survey of financial market participants conducted before the release.[30]

Table 4 presents the regression results. The second column reports estimates for market-based surprises associated with FOMC announcements and Federal Re-

[29] The S&P 500 stock market index and the Bloomberg Commodity Spot Price index are from Datastream, provided by LSEG Data & Analytics and accessible via a University of Bern license. The implied skewness is based on the paper by Bauer and Chernov (2024) and obtained from the website of the Federal Reserve Bank of San Francisco. Total nonfarm payrolls and the yield curve slope are from the St. Louis FRED database.

[30] Since we lack direct access to this survey (owned by Haver Analytics), we approximate the series as follows: first, we construct a time series of median expectations for months with FOMC meetings, where complete data is available. Then, for months without meetings, we estimate missing values using linear interpolation. This approach ensures that our series matches Bauer and Swanson (2023a) for FOMC statement dates while providing an approximation for speech dates.

serve Board chair and vice chair speeches. Consistent with concerns raised in the literature, we find evidence that these monetary policy surprises may capture factors beyond monetary policy shocks. Specifically, half of the coefficients on past economic or financial variables are statistically significant at the 5 percent level, suggesting that market-based surprises are correlated with economic and financial information available prior to the announcements or speeches. In contrast, the first column displays results for our language-based surprise series. Here, with the exception of the ten-year Treasury yield skewness, we find no statistically significant relationships with past economic and financial information, which indicates a substantial improvement over the market-based series. A similar pattern emerges in columns three and four, which repeat the analysis but restrict the sample to monetary policy surprises associated with only FOMC statements. The last two columns, which focus solely on speech transcripts, show no big differences between market-based and language-driven surprises. Since none of the coefficients in these columns are statistically significant, we conclude that predictability concerns arise only in the context of FOMC statements, not speeches.

These results suggest that our language-driven approach provides a cleaner measure of monetary policy surprises by filtering out influences from prior economic information. Unlike the market-based series, the language-driven surprises exhibit much weaker systematic relationship with past economic conditions. Moreover, the few remaining significant coefficients likely stem from extreme events or outliers, to which OLS is particularly sensitive.

To control for such potential outliers, the same regression exercise is conducted using median regression.[31] The results again show an improvement for the language-driven surprise series. During crises, such as the dot-com bubble burst or the 2008 financial crisis, markets tend to underestimate the Federal Reserve's actions, leading to larger negative surprises in periods of heightened uncertainty and volatility. This increases the likelihood of influential outliers.[32] The findings confirm that the language-driven series exhibit a weaker correlation with past economic and financial

[31]The results are presented in Appendix F

[32]Scatterplots for the two coefficients with the lowest p-values are shown in Appendix F.

Table 4: Regression on monetary policy surprises

	All		Statements		Speeches	
	LD	MB	LD	MB	LD	MB
NFP_SURP	0.0304	0.0280	0.0986	0.0875	0.0019	-0.0040
	(0.0914)	(0.1493)	(0.0393)	(0.0821)	(0.8957)	(0.7699)
NFP_12M	0.0010	0.0020	0.0024	0.0055	0.0007	0.0009
	(0.1686)	(0.0156)	(0.2893)	(0.0296)	(0.3065)	(0.1987)
SP500_3M	0.0309	0.0528	0.0334	0.1214	0.0206	0.0221
	(0.1570)	(0.0362)	(0.6211)	(0.0871)	(0.2198)	(0.2073)
SLOPE_3M	-0.0037	-0.0042	-0.0115	-0.0115	-0.0005	-0.0013
	(0.2847)	(0.2462)	(0.2475)	(0.2358)	(0.8195)	(0.5829)
BCOM_3M	-0.0139	0.0094	-0.0105	0.0623	-0.0147	-0.0082
	(0.4492)	(0.6633)	(0.8504)	(0.3445)	(0.3446)	(0.5867)
TR_SKEW	0.0118	0.0123	0.0302	0.0305	0.0033	0.0029
	(0.0162)	(0.0138)	(0.0171)	(0.0272)	(0.3987)	(0.3814)
N	619	619	178	178	441	441
R2	0.03	0.05	0.10	0.19	0.01	0.01

Note: p-values in parenthesis. The abbreviations MB and LD stand for market-based and language-driven monetary policy surprise series, respectively. In the former case, the raw changes in market prices within the tight time window around the communication is used as surprise series. In the latter case, our predicted market reactions from the language model are uses as surprise series. We use nonfarm payroll surprises (NFP SURP), the 12-month employment growth in total nonfarm payrolls (NFP 12M), the three-month growth in the S&P 500 stock market index (SP500 3M), the three-month change in the slope of the yield curve (SLOPE 3M), the three-month growth in the Bloomberg Commodity Spot Price index (BCOM 3M), and the average skewness of the ten-year Treasury yield over the past month (TR SKEW).

data, enhancing their properties regarding exogeneity.

5.3 Comparison with Other Monetary Policy Surprises

We compare our language-driven surprises to other monetary policy surprises from the literature. Table 5 summarizes the names, notation, and description of each monetary policy surprise series considered in our analysis. All of these series derive at least partially from high-frequency data.

To assess the similarity between our newly constructed language-driven surprise

Table 5: Description of other monetary policy surprise series

Series name	Description
Nakamura and Steinsson (2018) surprises	First principal component of the change in the current-month and next-month fed funds futures and in Eurodollar futures in 1,2,3,4 quarters, from January 1995 to March 2014
Miranda-Agrippino and Ricco (2021) surprises	Surprise change in the three-month-ahead fed funds futures, orthogonal to the central bank's economic projections and past market surprises, from January 1990 to December 2009
Swanson (2021) federal funds rate factor	Surprise change in the federal funds rate, from July 1991 to June 2019
Swanson (2021) forward guidance factor	Surprise change in forward guidance, from July 1991 to June 2019
Bauer and Swanson (2023a) surprises	First principal component of the change in Eurodollar futures in 1,2,3,4 quarters, from January 1988 to December 2019
Bauer and Swanson (2023a) orthogonalized surprises	Residual of the regression of BS onto macroeconomic and financial variables

Note: All of the monetary policy surprises mentioned in this table are available only for FOMC announcements, not for Federal Reserve Board speeches.

series and existing measures, we compute correlation coefficients. As the comparative monetary policy surprises are only available for FOMC announcements, our correlation analysis is limited to these events and does not include Federal Reserve Board chair and vice chair speeches.

As shown in Table 6, our language-driven monetary policy surprises exhibit substantial correlations with most existing monetary policy surprise measures, except for the surprises of Miranda-Agrippino and Ricco (2021). This suggests that our text-based approach successfully captures significant aspects of monetary policy shocks identified through more traditional marked-based methods. Two findings from this comparative analysis merit particular attention. First, our language-driven surprise series correlates much more strongly with the federal funds rate factor of

Swanson (2021) than with the forward guidance factor. This result challenges our initial expectations and suggests that, even when analyzing Federal Reserve communications, the current policy decision remains a dominant component captured by our measure. However, we note that forward guidance aspects likely play a more substantial role in our surprise series derived from Federal Reserve Board speeches, which are not included in this particular comparison. Second, our language-driven surprises demonstrate stronger correlation with the "raw" surprises of Bauer and Swanson (2023a) than with their orthogonalized counterparts. While one might initially expect the opposite, this finding likely reflects the different approaches employed to enhance exogeneity.

Table 6: Correlation between language-driven surprises and other surprise series

Series	Correlation
Nakamura and Steinsson (2018) surprises	0.5954
Miranda-Agrippino and Ricco (2021) surprises	-0.0504
Swanson (2021) federal funds rate factor	0.6275
Swanson (2021) forward guidance factor	0.2927
Bauer and Swanson (2023a) surprises	0.7376
Bauer and Swanson (2023a) orthogonalized surprises	0.6941

5.4 Central Bank Communication and Macroeconomic Variables

We now evaluate the macroeconomic effects of monetary policy shocks identified using language-driven surprises as instruments, comparing them directly with those identified using market-based surprises. Table 7 reports the robust F-statistics from first-stage regressions for both types of surprise series. Note that for market-based surprises, we include only dates with corresponding FOMC statements to ensure both series utilize identical observation sets. The F-statistics for language-driven surprises are comparable to those for market-based surprises. This demonstrates that our text analysis methodology preserves instrument relevance, while simultaneously improving exogeneity properties.

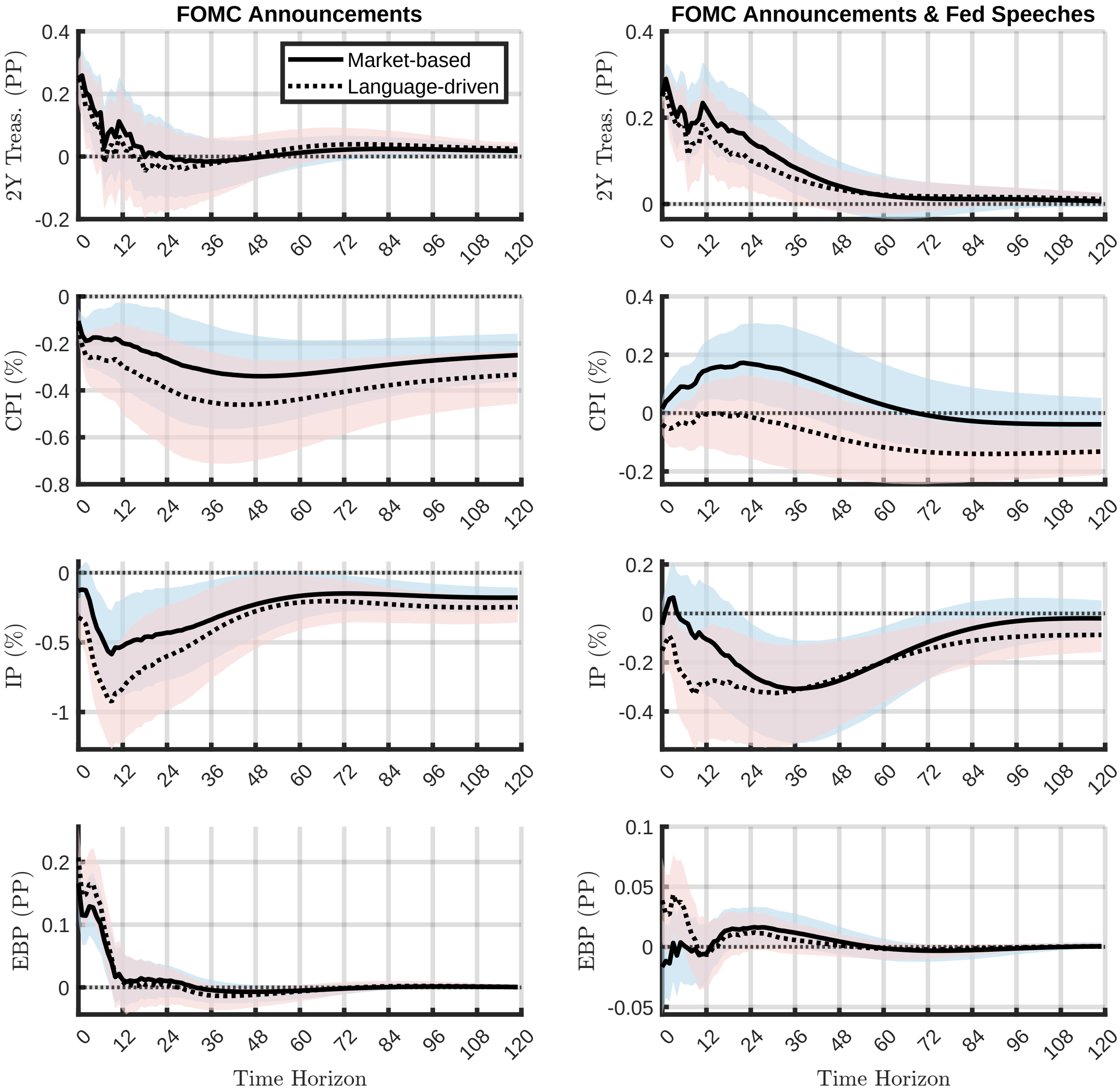
Table 7: Robust F-statistics for Language-Driven vs. Market-Based Surprise Series

	Language-Driven	Market-Based
FOMC announcements	2.09	2.64
FOMC announcements and *policy-relevant* speeches	6.43	7.39

Figure 5 displays the impulse responses to a monetary policy shock when using either market-based surprises or language-driven surprises as instruments for the identification. First, if we consider surprise series containing only FOMC announcements for identification, the dynamic responses of the macroeconomic variables differ only slightly. Second, if we use the surprises containing both FOMC announcements and Federal Reserve Board chair and vice chair speeches, the differences are bigger. Most importantly, with the language-based surprises, we no longer observe a price puzzle. The consumer price index does not react significantly on impact, but decreases below zero in the medium to long run. Moreover, industrial production decreases faster and stays negative for a prolonged period. Lastly, the excess bond premium increases (instead of decreasing), which is economically more intuitive.

While the F-statistics for both our language-driven and market-based surprise series fall below the conventional threshold of ten suggested by Stock and Watson (2012), our methodology represents a deliberate trade-off between instrument strength and improved exogeneity, a balance that appears justified given the more theoretically consistent results and the comparable performance relative to standard market-based approaches. Our language-driven approach yields F-statistics comparable to those of established market-based measures when using identical observation sets. This relative performance suggests our text-based methodology is at least as effective as conventional approaches for identifying monetary policy shocks. Despite the substantial difference in F-statistics between our two language-driven surprise specifications (2.09 for FOMC announcements versus 6.43 for FOMC announcements plus Federal Reserve Board speeches), both generate qualitatively similar impulse responses, suggesting some robustness in our identification approach. In particular, the impulse responses obtained using language-driven surprises align more closely

Figure 5: Impulse Responses to a Monetary Policy Shock (Language-Driven Surprises)



Note: The blue shaded areas represent the 5-95 percentiles of the impulse responses identified with market-based surprises and the red shaded areas the 5-95 percentiles of the impulse responses identified with language-driven surprises. The impulse responses are normalized to a 25 basis point increase in the two-year Treasury yield (2Y Treas.) and show the reaction of the consumer price index (CPI), industrial production (IP) and the excess bond premium (EBP). Horizontal axis: time horizon in months.

with theory, particularly by eliminating the price puzzle and generating more economically intuitive responses in the excess bond premium.

6 Conclusion

This paper improves the identification of monetary policy shocks by combining NLP techniques with an expanded set of central bank communications. We extend the traditional market-based surprise measures—typically derived solely from FOMC announcements—to also include policy-relevant speeches by the Federal Reserve Board chair and vice chair. This expansion significantly improves the relevance of these monetary policy surprises as instruments for identifying monetary policy shocks.

By leveraging a neural network trained on FOMC statements and speech transcripts, we construct a language-driven surprise series that isolates the component of market reactions driven purely by central bank communication. This approach mitigates endogeneity concerns inherent in traditional market-based surprises by filtering out confounding factors such as trader sentiment and market momentum. Our empirical findings confirm that language-based surprises produce impulse responses to monetary policy shocks that align more closely with economic theory.

Our results underscore the increasing importance of central bank communication as a monetary policy tool and demonstrate the potential of NLP in macroeconomic research. Future work could further refine our approach by incorporating additional forms of policy communication or testing alternative machine-learning architectures.

7 References

Aruoba, S. Boragan and Thomas Drechsel, "Identifying Monetary Policy Shocks: A Natural Language Approach," *NBER Working Paper 32417*, 2024.

Bauer, Michael and Mikhail Chernov, "Interest Rate Skewness and Biased Beliefs," *Journal of Finance*, 2 2024, *79* (1), 173–217.

Bauer, Michael D and Eric T Swanson, "A Reassessment of Monetary Policy Surprises and High-Frequency Identification," *NBER Macroeconomics Annual*, 2023, *37.*

_ **and** _ , "An Alternative Explanation for the "FED Information Effect"," *American Economic Review*, 2023, *113* (3), 664–700.

Bertsch, Christoph, Isaiah Hull, Robin L. Lumsdaine, and Xin Zhang, "Central bank mandates and monetary policy stances: Through the lens of Federal Reserve speeches," *Journal of Econometrics*, 5 2025, *249*, 105948.

Blinder, Alan S., Michael Ehrmann, Marcel Fratzscher, Jakob De Haan, and David Jan Jansen, "Central Bank Communication and Monetary Policy: A Survey of Theory and Evidence," *Journal of Economic Literature*, 12 2008, *46* (4), 910–945.

Caldara, Dario and Edward Herbst, "Monetary Policy, Real Activity, and Credit Spreads: Evidence from Bayesian Proxy SVARs," *American Economic Journal: Macroeconomics*, 1 2019, *11* (1), 157–192.

Cieslak, Anna and Andreas Schrimpf, "Non-Monetary News in Central Bank Communication," *Journal of International Economics*, 5 2019, *118*, 293–315.

Doh, Taeyoung, Dongho Song, and Shu-Kuei Yang, "Deciphering Federal Reserve Communication via Text Analysis of Alternative FOMC Statements," *Federal Reserve Bank of Kansas City, Research Working Paper no. 20-14*, 10 2020.

Fukushima, Kunihiko, "Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position," *Biological Cybernetics*, 4 1980, *36* (4), 193–202.

Gardner, Ben, Chiara Scotti, and Clara Vega, "Words Speak as Loudly as Actions: Central Bank Communication and the Response of Equity Prices to Macroeconomic Announcements," *Journal of Econometrics*, 12 2022, *231* (2), 387–409.

Gertler, Mark and Peter Karadi, "Monetary Policy Surprises, Credit Costs, and Economic Activity," *American Economic Journal: Macroeconomics*, 2015, *7* (1), 44–76.

Gilchrist, Simon and Egon Zakrajšek, "Credit Spreads and Business Cycle Fluctuations," *American Economic Review*, 6 2012, *102* (4), 1692–1720.

Gürkaynak, Refet S, Brian Sack, and Eric Swanson, "The Sensitivity of Long-Term Interest Rates to Economic News: Evidence and Implications for Macroeconomic Models," *American Economic Review*, 2 2005, *95* (1), 425–436.

Gürkaynak, Refet S., Brian Sack, and Eric T. Swanson, "Do Actions Speak Louder Than Words? The Response of Asset Prices to Monetary Policy Actions and Statements," *International Journal of Central Banking*, 2005, *1* (1).

_ , _ , and Jonathan H. Wright, "The U.S. Treasury Yield Curve: 1961 to the Present," *Journal of Monetary Economics*, 2007, *54* (8), 2291–2304.

Handlan, Amy, "Text Shocks and Monetary Surprises: Text Analysis of FOMC Statements with Machine Learning," *Job Market Paper*, 2022.

Hanson, Samuel G. and Jeremy C. Stein, "Monetary Policy and Long-Term Real Rates," *Journal of Financial Economics*, 3 2015, *115* (3), 429–448.

Jarociński, Marek and Peter Karadi, "Deconstructing monetary policy surprises-The role of information shocks," *American Economic Journal: Macroeconomics*, 4 2020, *12* (2), 1–43.

Jayawickrema, Vishuddhi and Eric Swanson, "Speeches by the Fed Chair Are More Important Than FOMC Announcements: An Improved High-Frequency Measure of U.S. Monetary Policy Shocks," *unpublished manuscript, University of California, Irvine*, 7 2023.

Kerssenfischer, Mark and Maik Schmeling, "What moves markets?," *Journal of Monetary Economics*, 7 2024, *145.*

Kuttner, Kenneth N, "Monetary Policy Surprises and Interest Rates: Evidence from the Fed Funds Futures Market," *Journal of Monetary Economics*, 2001, *47* (3), 523–544.

Lecun, Y., L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition," *Proceedings of the IEEE*, 1998, *86* (11), 2278–2324.

Loughran, Tim and Bill McDonald, "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks," *Journal of Finance*, 2 2011, *66* (1), 35–65.

Lucca, David O. and Emanuel Moench, "The Pre-FOMC Announcement Drift," *The Journal of Finance*, 2 2015, *LXX* (1), 329–371.

Mertens, Karel and Morten O. Ravn, "The Dynamic Effects of Personal and Corporate Income Tax Changes in the United States," *American Economic Review*, 6 2013, *103* (4), 1212–1247.

Miranda-Agrippino, Silvia and Giovanni Ricco, "The Transmission of Monetary Policy Shocks," *American Economic Journal: Macroeconomics*, 7 2021, *13* (3), 74–107.

Nakamura, Emi and Jón Steinsson, "High-Frequency Identification of Monetary Non-Neutrality: The Information Effect," *The Quarterly Journal of Economics*, 8 2018, *133* (3), 1283–1330.

Neuhierl, Andreas and Michael Weber, "Monetary Momentum," *NBER Working Paper 24748*, 6 2018.

Ochs, Adrian C R, "A New Monetary Policy Shock with Text Analysis," *Cambridge Working Papers 2148, Faculty of Economics, University of Cambridge*, 6 2021.

Ramey, V. A., "Macroeconomic Shocks and Their Propagation," in "Handbook of Macroeconomics," Vol. 2, Elsevier B.V., 2016, pp. 71–162.

Sarkar, Suproteem K and Keyon Vafa, "Lookahead Bias in Pretrained Language Models," *Working Paper*, 10 2024.

Stock, James H. and Mark W. Watson, "Disentangling the Channels of the 2007-09 Recession," *Brookings Papers on Economic Activity*, 2012, *1*, 81–135.

Swanson, Eric T., "Measuring the effects of federal reserve forward guidance and asset purchases on financial markets," *Journal of Monetary Economics*, 3 2021, *118*, 32–53.

_ **and John C. Williams**, "Measuring the Effect of the Zero Lower Bound on Medium- and Longer-Term Interest Rates," *American Economic Review*, 10 2014, *104* (10), 3154–3185.

Tran, Paul L., "How Long Do Markets Need to Fully React to Monetary Policy Announcements?," *Job Market Paper*, 2025.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention Is All You Need," *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

Woodford, Michael, "Central Bank Communication and Policy Effectiveness," *NBER Working Paper 11898*, 2005.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le, "XLNet: Generalized Autoregressive Pretraining for Language Understanding," *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, 1 2020.

A Text Cleaning

We perform basic text cleaning by replacing repetitive and technical words with an abbreviation. Table 8 provides an overview of the abbreviations used. We use one word for the abbreviation except the *Federal Open Market Committee* is replaced with *Committee (FOMC)* because the FOMC statements usually refer to the *Committee* in their statements, whereas the Federal Reserve Board Chair and Vice Chair Speeches usually refer to the *FOMC*. However, both refer to the Federal Open Market Committee. Additionally, we restructure percentage numbers to match the following format: X.XX percent. Especially in the FOMC announcements, rate changes are sometimes marked in fractions, e.g. 1/4, making it hard to interpret for an NLP model. Thus, we similarly restructure all percentage numbers to facilitate comprehension. Finally, in the FOMC statements, we replaced the introductory sentence *Information received since the Committee (FOMC) met in January* with *Information received* for every month to prevent the model from learning from the timeline.

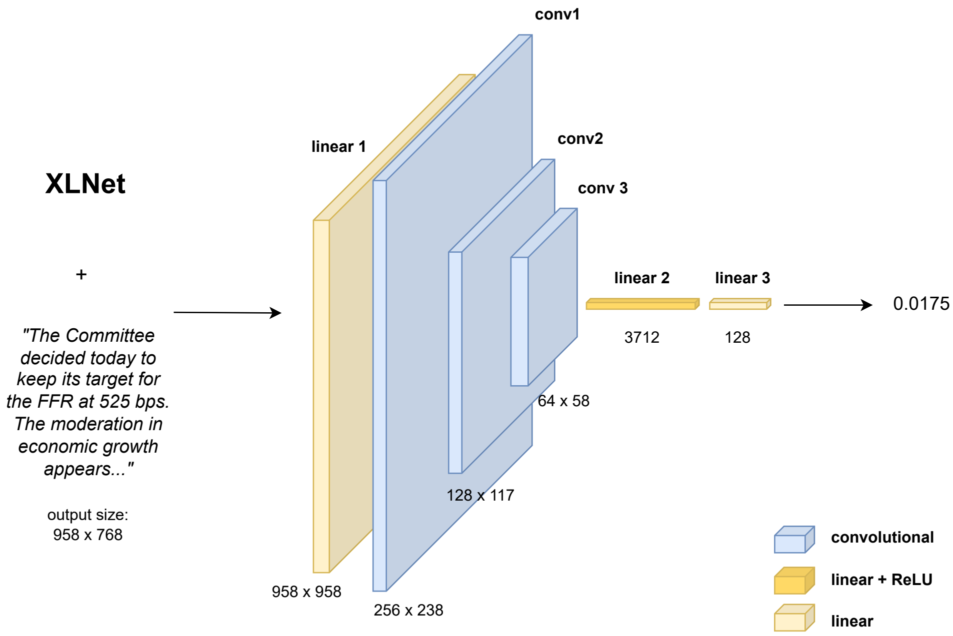
Table 8: Text Cleaning

Words	Abbreviation
Federal Open Market Committee	Committe (FOMC)
federal funds rate	FFR
Board of Governors	BOG
Federal Reserve	FR
basis points	bps
basis point	bps
-basis-point	bps
mortgage-backed securities	mbs
Term Asset-Backed Securities	TABS

B Neural Network Architecture

As mentioned, we use a pre-trained language model, XLNet-Base, developed and trained by Yang et al. (2020) and provided by the platform *Hugging Face*. The backbone neural network consists of 12 layers and 768 hidden states. On top of this, to train the model on our specific task, we add another six layers. A graphical representation of our additional structure is shown in figure 6. First, we increase the number of hidden states to match the length of our texts (number of tokens). Every token has a hidden state and associated weights in our first layer. Second, we add three convolutional layers, simultaneously breaking down the number of notes and tokens. Convolutional neural networks (CNNs) were first proposed by Fukushima (1980) but gained greater attention in machine learning when Lecun et al. (1998) presented LeNet, an algorithm that detected handwritten numbers. CNNs learn via filter optimization and thus symmetrically reduce the number of notes and tokens. The notes remain fully connected using this procedure, making it prone to overfitting. Nonetheless, since we work with a small, heterogeneous text data set, we profit from the connectivity but must check that our model is not overfitting. Third, we add two linear layers, including activation functions, to decrease the number of notes to a single prediction. As an activation function, we use the Rectified Linear Unit (ReLU). In a neural network, the activation function transforms the summed weighted input from the node into the node's activation or output. ReLU is a piecewise linear function that will output the input directly if it is positive. Otherwise, it will output zero. It has become the default activation function for many types of neural networks because a model that uses it is easier to train and often performs better.

Figure 6: Neural Network Architecture



Note: The graph presents the architecture of our neural network, taking as input the pre-trained XLNet and the FOMC statement. The input runs through different linear and convolutional layers, adapting its size to condense the information to a single number.

C Overview of the Training Algorithm

1. Scale the continuous labels, i.e., the changes in the federal funds futures, removing the median and scaling the data according to the quantile range.

2. Define the hyperparameter:

   (a) Learning rate: Defines how much the neuronal network should adapt its parameter after each iteration. We chose a rather low learning rate.

   (b) Number of epochs: Defines how often the model sees the same data set to adapt its parameters.

   (c) Loss function: Defines how the model should penalise its results compared to the true label. Since we work with linear prediction, we take the mean squared error.

(d) Batch size: Defines the number of statements we show to the model simultaneously.

3. Split the statements into training and test data. We use a 5-fold cross-validation, splitting the data set into five different subsets, always taking one of the splits as a test split and letting the model train with the other four splits.

4. Set the model to training mode to adapt its parameters.

5. Train the model by adapting its parameters such that the loss becomes decreases.

6. Stop updating the parameters

7. Evaluate the model using the test data.

8. Repeat from step four until the number of epochs (defined before) is reached. If the model is already overfitting, but the number of epochs is not yet reached, we should stop before.

9. Repeat from step three until all splits are tested.

10. Unscale the results.

D Fine-Tune the Neural Network using FOMC statements

In this section, we map policy communication text to market-based monetary policy surprises, isolating the portion of the surprises driven exclusively by the FOMC statements. In contrast to Section 5, we abstain from using speech transcripts as part of the training set and use these only for predicting changes in market expectations. Truly, the statements are relatively short but comprise vocabulary that is highly policy-relevant.

D.1 Natural Language Processing

To obtain the prediction, we follow the same four steps explained in Section 5.1 with the only difference that we use a different dataset for fine-turning the neural network. We want our model to learn the mapping from FOMC statements to market-based surprises and apply this mapping to the speech transcripts. To train our model, we apply again a five-fold cross-validation using solely the FOMC statements. Thus, we have 80% of our FOMC statements in the training set and 20% in the test set. We use the same hyperparameters and algorithm as described in Appendix B and C, respectively. Based on our test runs, we fixed the learning rate of our model to 1e-5 and the number of epochs to 8. The MSE of this set of hyperparameter for the five-fold cross-validation is displayed in Table 9.

Table 9: MSE of 5-Fold Cross-Validation after 8 Epochs

Number of Split	In-Sample MSE	Out-of-Sample MSE
1	0.0012598837	0.0012116632
2	0.0004515733	0.003678869
3	0.002779334	0.0034142113
4	0.0028541489	0.0032873761
5	0.0012911019	0.0020300713

After training our neural network, we obtain predictions from the FOMC statements and speech transcripts. In other words, we obtain a monetary policy surprise series that contains largely out-of-sample predictions.

D.2 Relation to Past Economic Information

After obtaining our predictions, we verify again whether our language-based surprises are uncorrelated to economic and financial information available before the announcement or speech. We conduct the same regression as in Section 5. The results are presented in Table 10 and are similar to the ones presented earlier in the paper. The only exception is the coefficient of nonfarm payroll surprises (NFP SURP) that is here significant on the 5% level, for our combined results.

Similar to before, these results suggest that our language-driven approach exhibits no systematic relationship with past economic data. Hence, also using a limited training sample help to improve the surprise series to capture only the reaction to the central bank communication and abstract for other past information.

Table 10: Regression on monetary policy surprises

	All		Statements		Speeches	
	LD	MB	LD	MB	LD	MB
NFP_SURP	0.0091	0.0280	0.0214	0.0875	0.0026	-0.0040
	(0.0259)	(0.1493)	(0.0434)	(0.0821)	(0.2913)	(0.7699)
NFP_12M	0.0002	0.0020	0.0010	0.0055	-0.0001	0.0009
	(0.3470)	(0.0156)	(0.1353)	(0.0296)	(0.4183)	(0.1987)
SP500_3M	0.0083	0.0528	0.0255	0.1214	0.0026	0.0221
	(0.1438)	(0.0362)	(0.1419)	(0.0871)	(0.4490)	(0.2073)
SLOPE_3M	-0.0003	-0.0042	-0.0036	-0.0115	0.0010	-0.0013
	(0.6716)	(0.2462)	(0.0707)	(0.2358)	(0.0248)	(0.5829)
BCOM_3M	0.0001	0.0094	0.0139	0.0623	-0.0048	-0.0082
	(0.9829)	(0.6633)	(0.3641)	(0.3445)	(0.1001)	(0.5867)
TR_SKEW	0.0023	0.0123	0.0052	0.0305	0.0007	0.0029
	(0.0099)	(0.0138)	(0.0327)	(0.0272)	(0.3539)	(0.3814)
N	619	619	178	178	441	441
R2	0.04	0.05	0.20	0.19	0.03	0.01

Note: p-values in parenthesis. The abbreviations MB and LD stand for Market Based and Language Driven monetary policy surprise series, respectively. In the former case, the raw changes in market prices within the tight time window around the communication is used as surprise series. In the latter case, our predicted market reactions from the language model are uses as surprise series.

D.3 Monetary Policy Effects on Macroeconomic Variables

Analogous to the Section 4 and 5, we assess the dynamics of the monetary policy shock identified when using language-driven surprises as instruments. First, we report the F-statistics of the first-stage regression for two instrument series: the language-driven surprises related to the FOMC statements and the language-driven surprises related to the FOMC statements and policy-relevant Federal Reserve Board chair and vice chair speech transcripts.

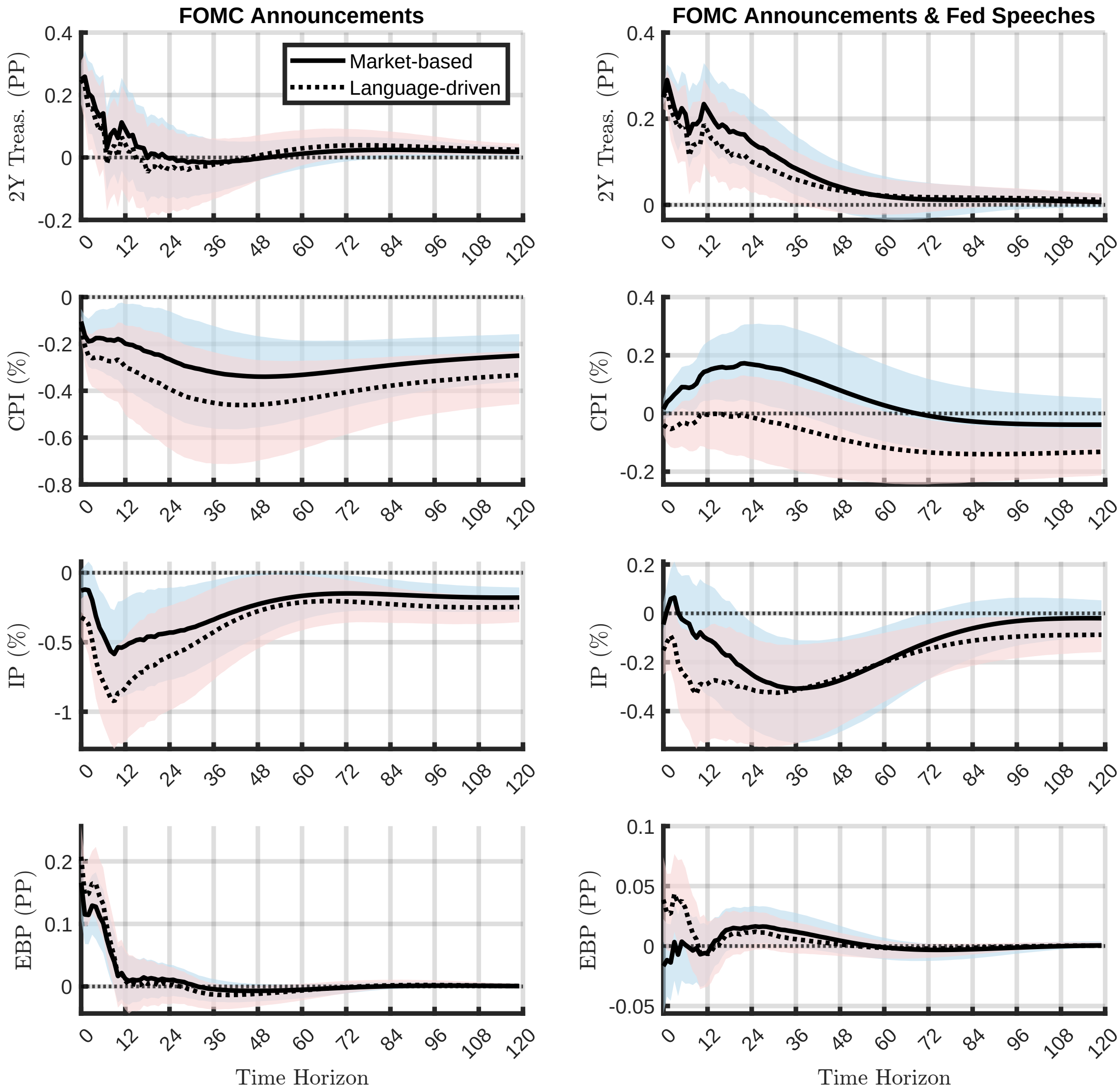
Table 11 reports the robust F-statistics for both surprise series. These values, especially for the FOMC statements and policy-relevant speeches, are drastically lower compared to the results obtained in Section 4 and 5. This indicates that the model, trained only on FOMC statements, retrieves less information from the policy-relevant speeches.

Table 11: F-statistics for language-driven surprises, alternative fine-tuning

	F-statistic
FOMC statements	3.31
FOMC statements and *policy-relevant* speech transcripts	2.97

Figure 7 displays the impulse responses to a monetary policy shock when using either market-based surprises or language-driven surprises as instruments for identification. First, for the results using only FOMC announcements for identification, we observe marginally different results. With the language-based surprises, we obtain a negative impact on the CPI only after some years. However, the reaction of industrial production is much stronger using our series. Second, for the results using both FOMC announcements and Federal Reserve Board chair and vice chair speeches, the deviation is even bigger. Using the language-driven surprises, the two-year treasury yield mean-reverts much quicker. Moreover, the CPI reacts negatively on impact and remains negative for all periods. In contrast, industrial production decreases some months after impact but reverts shortly after. Similar to the results obtained in Section 5, the excess bond premium increases in reaction to the monetary policy shock. In essence, the results are similar to the ones obtained in Section 5 with the big difference that our instrument remains weak even when adding the speech transcripts.

Figure 7: Impulse Responses to a Monetary Policy Shock (Language-Driven Surprises)



Note: The blue shaded areas represent the 5-95 percentiles of the impulse responses identified with market-based surprises and the red shaded areas the 5-95 percentiles of the impulse responses identified with language-driven surprises. The impulse responses are normalized to a 25 basis point increase in the two-year Treasury yield (2Y Treas.) and show the reaction of the consumer price index (CPI), industrial production (IP) and the excess bond premium (EBP). Horizontal axis: time horizon in months.

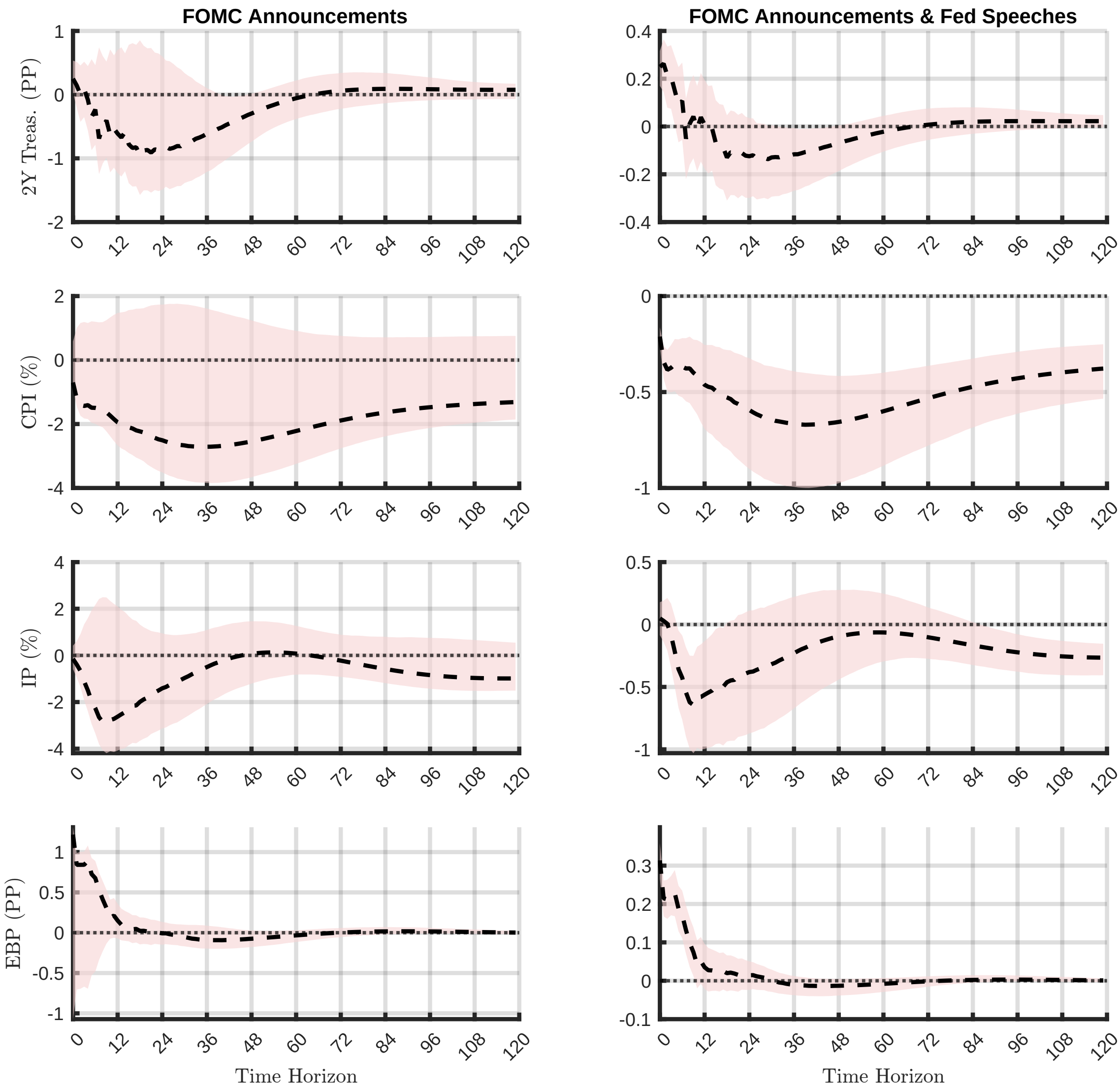
E Assessment of the Information Effect's Influence

To assess whether information effects continue to influence our language-driven surprises, we conduct an exercise inspired by the methodology in Jarociński and Karadi (2020). According to their findings, an unexpected monetary policy tightening raises interest rates and lowers stock prices, whereas a positive central bank information shock leads to increases in both. Following a "poor man's sign restriction" approach, we exclude all positive monetary policy surprises that coincide with rising stock prices, and vice versa.[33]

Figure 8 presents the impulse responses obtained using this restricted subset of surprises as instruments. The responses are qualitatively in line with the patterns observed in Section 5.4, suggesting that, if information effects are present in our language-driven surprises, they do not play a dominant role. Note that the confidence bands are wide when restricting the analysis to surprises around FOMC announcements, reflecting the limited number of observations available when applying the sign restriction filter. Despite this, the point estimates remain largely consistent with those from the unrestricted analysis.

[33] We only have access to daily stock price data. Therefore, we use daily changes in stock prices and match them to the high-frequency monetary policy surprises.

Figure 8: Impulse responses to a monetary policy shock, restricted subset of language-driven surprises based on poor man's sign restrictions



Note: The impulse responses are identified with a restricted subset of language-driven surprises. The restricted subset is determined by excluding the surprises, which move in the same direction as stock prices. The impulse responses are normalized to a 25 basis point increase in the two-year Treasury yield. Horizontal axis: time horizon in months.

F Robustness to Outliers

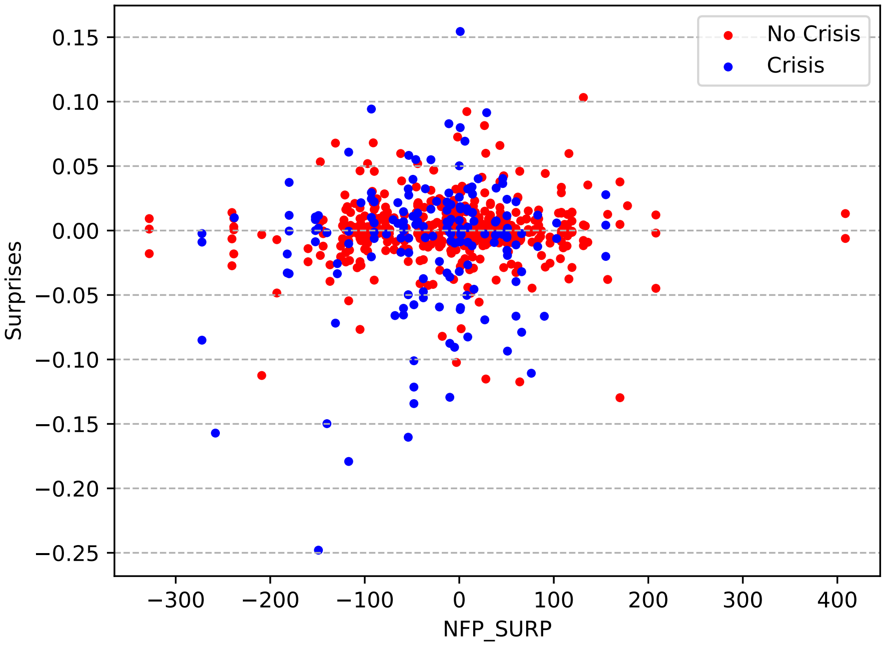
This section comprises the additional results described in section 5. Table 12 displays the results from the median regression, using the same specifications as for the OLS regression in section 5. Figures 9 and 10 plot the outcomes, the language-driven surprises, on the two covariates with the lowest p-values, the nonfarm payroll surprises and the average skewness of the ten-year treasury yields. Dates during times of crisis are plotted in blue, other dates are plotted in red. Times of crisis include the dot-com bubble burst between 1.1.2020 and 31.12.2002 and the financial crisis between 1.7.2007 and 1.1.2010.

Table 12: Median Regression on monetary policy surprises

	All		Statements		Speeches	
	LD	MB	LD	MB	LD	MB
NFP_SURP	0.0049	-0.0013	0.0120	-0.0049	0.0010	-0.0000
	(0.5508)	(0.8398)	(0.7818)	(0.9002)	(0.9038)	(0.9998)
NFP_12M	0.0003	0.0004	0.0035	0.0035	0.0001	0.0000
	(0.5536)	(0.2395)	(0.1281)	(0.0913)	(0.7897)	(0.9994)
SP500_3M	0.0037	0.0016	-0.0140	0.0538	0.0071	0.0000
	(0.7184)	(0.8426)	(0.7820)	(0.2429)	(0.4955)	(0.9999)
SLOPE_3M	-0.0002	-0.0007	-0.0030	-0.0119	-0.0006	-0.0000
	(0.9058)	(0.5335)	(0.6497)	(0.0519)	(0.6782)	(0.9995)
BCOM_3M	-0.0069	0.0007	0.0042	0.0368	-0.0070	-0.0000
	(0.4230)	(0.9133)	(0.9174)	(0.3164)	(0.4239)	(1.0000)
TR_SKEW	0.0037	0.0005	0.0129	0.0148	0.0025	0.0000
	(0.1297)	(0.8011)	(0.2375)	(0.1370)	(0.3234)	(0.9999)
N	619	619	178	178	441	441

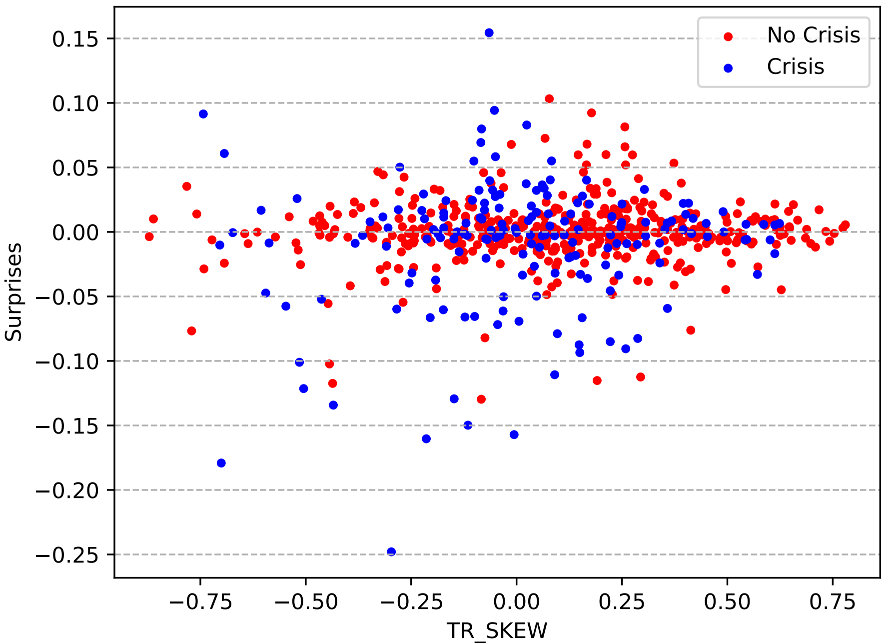
Note: p-values in parenthesis. The abbreviations MB and LD stand for Market Based and Language Driven monetary policy surprise series, respectively. In the former case, the raw changes in market prices within the tight time window around the communication is used as surprise series. In the latter case, our predicted market reactions from the language model are uses as surprise series. We use nonfarm payroll surprises (NFP SURP), the 12-month employment growth in total nonfarm payrolls (NFP 12M), the three-month growth in the S&P 500 stock market index (SP500 3M), the three-month change in the slope of the yield curve (SLOPE 3M), the three-month growth in the Bloomberg Commodity Spot Price index (BCOM 3M), and the average skewness of the ten-year Treasury yield over the past month (TR SKEW).

Figure 9: Language driven surprises on nonfarm payroll surprises



Note: The graph presents the surprises used in the analysis on the y-axis and the most recent nonfarm payroll surprises on the x-axis.

Figure 10: Language driven surprises on average skewness of the treasury yield

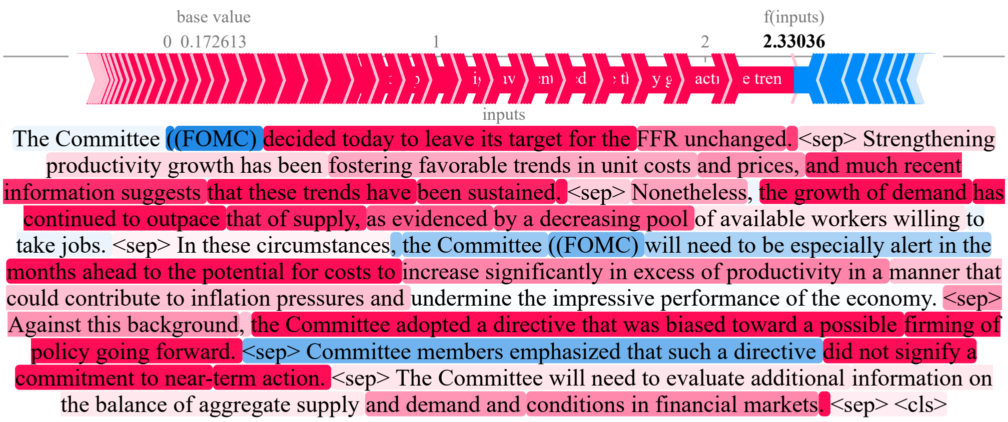


Note: The graph presents the surprises used in the analysis on the y-axis and the average skewness of the ten-year treasury yield in the last month on the x-axis.

G Understand the Model Output

We want to understand what our model learned from FOMC statements and policy-relevant speech transcripts. To explain the output of our model, we use SHAP (SHapley Additive exPlanations) values. They help to understand how much each input word contributes to a particular prediction. Essentially, they break down the model's prediction for a specific text by attributing portions of the prediction to each feature. We provide different examples. In Figure 11, we can see that the overall prediction of the model is an increase of market expectations by 2.33. This value is rescaled and translates to an increase in ED2 by 0.07. Hence, the market expects the policy rate to go down in the future, which can be seen as an easing of monetary policy. Passages such as *...decided today to leave its target for the FFR unchanged...*; *...fostering favorable trends in unit costs and prices, and much recent information suggests that these trends have been sustained...* and *...the growth of demand has continued to outpace that of supply...* contributed positively to this prediction. Whereas *...the Committee (FOMC) will need to be especially alert...* contributed negatively. From these sentences, we would generally expect such a reaction. Hence, our model learned to interpret the market reaction accurately.

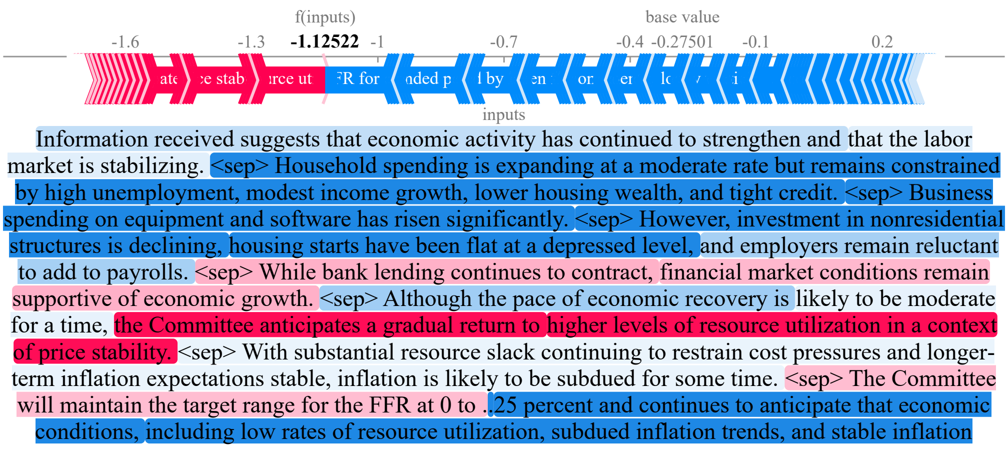
Figure 11: SHAP Values from FOMC Statement 1999-10-05



Note: The figure marks text input red (blue) if it contributed positively (negatively) to the prediction. Firmer colors mark stronger contributions to the prediction. <sep> and <cls> are tokes that we added to flag the end of a sentence and the end of the statement to the model.

In Figure 12, we see a decrease of -1.13, rescaled this translates to a reduction in ED2 by -0.02. This can be seen as a contractionary monetary policy shock. We can see that *...Household spending is expanding at a moderate rate but remains constrained by high unemployment, modest income growth, lower housing wealth, and tight credit. Business spending on equipment and software has risen significantly. However, investment in nonresidential structures is declining, housing starts have been flat at a depressed level,...* contributed negatively to the prediction. Again, keeping the target rate unchanged *...The Committee will maintain the target range for the FFR at 0...* contributed positively to the prediction. Yet, the second part of the phrase *continues to anticipate that economic conditions, including low rates of resource utilization, subdued inflation trends, and stable inflation expectations, are likely to warrant exceptionally low levels of the FFR for an extended period...* is seen as a rather negative signal. We can see that even though the interest rate didn't change, the language is quite strong and provokes negative market reactions.

Figure 12: SHAP Values from FOMC Statement 2010-03-16



Note: The figure marks text input red (blue) if it contributed positively (negatively) to the prediction. Firmer colors mark stronger contributions to the prediction. <sep> and <cls> are tokes that we added to flag the end of a sentence and the end of the statement to the model.