LLM-Based Expectations and Fed Communication

Antonio Gargano [*]
University of Houston

Yinru Lei [†]
Hunan University

Bin Li [‡]
Wuhan University

Alberto G. Rossi [§]
Georgetown University

February 1, 2026

Abstract

We introduce a novel approach to measuring monetary policy surprises in FOMC press conferences. We use LLMs to generate counterfactual press-conference responses by the Federal Reserve Chair and measure policy surprises as deviations between realized and simulated answers. Using high-frequency intraday data from 2011 to 2025, we show that equity prices respond strongly to these surprises: hawkish surprises lead to significant declines in stock prices, while dovish surprises increase prices. Importantly, markets respond to the unexpected component of communication rather than to the level of expressed policy stance. Heterogeneity tests show that market responses to policy surprises operate through expectation formation and belief updating. Reactions are strongest when the Chair's answers directly inform the future policy path, namely for forward-looking, policy-related, or statement-linked questions, and are larger, though slower to emerge, during periods of elevated market uncertainty. Responses are also stronger following dovish FOMC statements, when prior expectations are less anchored. Finally, omitting macroeconomic context or the policy statement from the LLM prompt eliminates the return predictability of policy surprises. Our findings demonstrate that LLMs provide a powerful tool for identifying economically meaningful surprises in real-time policy communication.

Keywords: LLMs, FOMC, Return Predictability
JEL Classification: E44, E52, G12

[*]C.T. Bauer College of Business, University of Houston. Email: agargano@bauer.uh.edu.
[†]College of Finance and Statistics, Hunan University. Email: yinrulei@hnu.edu.cn.
[‡]Economics and Management School, Wuhan University. Email: binli.whu@whu.edu.cn.
[§]McDonough School of Business, Georgetown University. Email: agr60@georgetown.edu.

1 Introduction

Central bank communication has become a primary tool of monetary policy. Beyond setting interest rates, the Federal Reserve now actively manages market expectations through forward guidance, press conferences, and public speeches. The information revealed in these communications can move asset prices, shift the term structure of interest rates, and therefore influence real economic activity in both the short- and medium-run (Bernanke and Kuttner, 2005; Nakamura and Steinsson, 2018; Golez and Matthies, 2025). Yet measuring the information content of policy communication remains a fundamental empirical challenge. The difficulty lies in distinguishing what market participants already expected from what genuinely surprised them—a distinction that requires modeling how investors form and update beliefs in real time.

A growing literature demonstrates that FOMC press conferences contain market-moving information beyond the policy statement itself. Gorodnichenko, Pham, and Talavera (2023) show that the Chair's vocal tone affects equity prices and macroeconomic expectations. Curti and Kazinnik (2023) document that facial expressions have independent effects on risk and trading volumes. Gómez-Cram and Grotteria (2022) find that verbal communication during press conferences generates real-time price discovery. All these studies have a common underlying feature: they measure the *content* of communication—its tone, sentiment, or novelty—rather than its *surprise* relative to market expectations. Without a credible counterfactual for what the Chair was expected to say, it is impossible to identify the unexpected component that should, in theory, move prices. Traditional approaches using surveys or interest rate futures are poorly suited to capturing surprises in nuanced oral communication, while dictionary-based text methods require keywords that rarely appear verbatim in spontaneous spoken answers.

This paper proposes a new approach to measuring policy surprises in central bank communication. We use large language models (LLMs) to simulate the Federal Reserve Chair's responses to journalist questions during FOMC press conferences, conditioning the model on the same real-time information set available to market participants. The difference between the simulated response—representing what a well-informed observer would expect the Chair to say—and the Chair's actual response defines our policy surprise measure. Recent work documents that LLMs exhibit human-like patterns of belief

formation and information processing (Bybee, 2023; Kim, Muhn, and Nikolaev, 2024), making them well-suited to this task. FOMC press conferences provide a clean setting for this exercise: the policy statement, released thirty minutes before the Q&A begins, contains the Committee's decision and rationale in a concise format that market participants can fully absorb before the Chair takes questions. We can therefore treat the statement, along with macroeconomic data and prior Q&A exchanges, as conditioning information that anchors both investor expectations and our LLM-generated counterfactuals. We show that this surprise measure predicts intraday stock returns during press conferences: hawkish surprises are associated with equity price declines, while dovish surprises are associated with price increases.

We implement this approach using all FOMC press conferences from 2011 through June 2025, spanning the tenures of three Chairs: Bernanke, Yellen, and Powell. For each press conference, we construct a real-time information set consisting of vintage macroeconomic data, the Summary of Economic Projections (when available), the policy statement released at 14:00 ET, the Chair's introductory remarks, and the cumulative sequence of Q&A exchanges that have already occurred. We then prompt the LLM to generate the Chair's response to each journalist question, conditioning on this information set. Both the simulated and actual responses are classified into five policy stance categories—strongly dovish, dovish, neutral, hawkish, and strongly hawkish—following Hansen and Kazinnik (2023), who show that LLMs classify monetary policy stance with higher accuracy than traditional dictionary approaches. The policy surprise for each answer is defined as the difference between the stance implied by the actual response and the stance implied by the simulated counterfactual.

To link press-conference communication to asset prices, we time-stamp each question and answer using automatic speech recognition applied to the press conference video. This yields the precise start and end time of every statement, allowing us to match policy surprises to high-frequency equity returns. We aggregate surprises cumulatively over the press conference, weighting by answer duration, and estimate local projection regressions relating SPY returns to the cumulative policy surprise at horizons ranging from 1 to 15 minutes.

We start by validating our measure along several dimensions. At the meeting level, the policy stance extracted from actual press-conference answers correlates strongly with the stance implied by

simulated counterfactual answers, confirming that our counterfactual responses capture the systematic component of policy communication. Importantly, the absolute magnitude of our policy surprise is positively associated with both market uncertainty (the VIX) and the conventional monetary policy surprise of Nakamura and Steinsson (2018). These correlations suggest that larger deviations between realized and counterfactual communication occur precisely when policy uncertainty is elevated and when the FOMC decision itself surprises markets.

Our main finding is that equity prices respond systematically to the unexpected component of monetary policy communication. Hawkish policy surprises—cases where the Chair's actual response is more hawkish than the LLM counterfactual—are associated with statistically significant declines in equity prices, while dovish surprises are associated with price increases. The response emerges rapidly after each question and lasts for several minutes. In terms of magnitude, a one-standard-deviation hawkish surprise is associated with a 5–10 basis point negative return on the SPY over the subsequent 5–10 minutes.

Importantly, equity prices respond to the *surprise* component of communication rather than to the *level* of expressed policy stance. When we replace the policy surprise with the policy stance conveyed in the Chair's realized answers, the estimated response is weak and largely insignificant. This finding confirms that markets do not move mechanically in the direction of hawkish or dovish communication; instead, they respond to deviations from expectations.

We conduct several tests to sharpen the economic interpretation of these results. First, we examine heterogeneity across question characteristics. Policy surprises generate stronger market responses when questions are forward-looking, policy-related, or directly tied to the FOMC statement. These are precisely the dimensions along which the Chair's answers most directly inform expectations about the future policy path. When questions address backward-looking or economy-related topics, the response is delayed or attenuated, consistent with markets requiring additional context to assess the policy implications.

Second, we examine how market responses vary with the informational environment. During periods of elevated market uncertainty, as measured by the VIX, policy surprises have larger effects—though the response emerges more slowly, consistent with noise confounding the signal at short horizons

(Bekaert, Hoerova, and Duca, 2013). Following dovish FOMC statements, press-conference surprises generate larger price reactions than following hawkish statements, suggesting that the scope for belief updating is greater when prior expectations are less anchored.

Third, we vary the conditioning information provided to the LLM. When we omit macroeconomic context or exclude the policy statement from the prompt, the estimated return responses are no longer statistically significant at any horizon. This finding demonstrates that both components are necessary for constructing economically meaningful surprises: the LLM-generated counterfactual must be anchored to the same information set that market participants use to forecast the future policy path.

Fourth, we compare our LLM-based approach to two established dictionary-based methods: the fixed-phrase dovish–hawkish dictionary of Neuhierl and Weber (2019) and the compositional term–modifier dictionary of Apel, Grimaldi, and Hull (2022). Policy surprises constructed using these dictionaries generate weaker and less persistent market responses. The fixed-phrase dictionary produces no statistically significant response at any horizon, likely because predefined keywords rarely appear verbatim in spontaneous spoken answers. The term–modifier approach captures some short-horizon variation but fails to generate the persistent responses observed with the LLM-based measure. These comparisons underscore the advantage of semantic understanding over keyword counting in capturing policy signals from oral communication.

In the final part of the paper, we provide several robustness exercises. First, we examine alternative aggregation methods. The baseline analysis aggregates answer-level surprises using a time-weighted cumulative measure, where each answer contributes in proportion to its duration. Using equal-weighted aggregation yields qualitatively similar results, though with smaller magnitudes—suggesting that longer answers carry more policy-relevant information. Replacing the cumulative measure with individual-answer surprises yields insignificant results, consistent with markets integrating information progressively over the press conference rather than responding to isolated statements.

Second, we address the concern that the LLM may draw on information learned during training—including transcripts of the press conferences we study—leading to look-ahead bias. To test this, we restrict the sample to press conferences held after the model's knowledge cutoff date (October

2023), ensuring that simulated answers cannot reflect post-event information. Our results hold in this restricted sample, with magnitudes comparable to the full sample, suggesting that look-ahead bias is not a significant concern.

Third, we examine alternative dependent variables. Beyond equity returns, hawkish policy surprises are associated with increases in the VIX and appreciation of the US Dollar versus the Euro. These results indicate that press-conference surprises are reflected across multiple asset classes, reinforcing the interpretation that our measure captures economically meaningful variation in policy communication.

2 Contributions to the Literature

This paper contributes to several strands of literature. First, we contribute to the literature on the effects of central bank communication on financial markets. Traditionally, this literature has focused on the effect of the formal policy decisions on fundamental economic outcomes such as the yield curve (Kuttner, 2001; Nakamura and Steinsson, 2018), the stock market (Bernanke and Kuttner, 2005; Cieslak, Morse, and Vissing-Jorgensen, 2019) , exchange rates (Gurkaynak et al., 2021), and short-term economic activity (Golez and Matthies, 2025). An increasing body of research indicates that markets respond not only to policy decisions themselves but also to the content and tone of how these decisions are communicated.[1] Since 2011, a press conference has followed all FOMC meetings with the goal to increase the transparency of the Fed's actions and provide further forward guidance. More recent studies examining the incremental information revealed during these events (Gorodnichenko, Pham, and Talavera, 2023; Curti and Kazinnik, 2023; Gómez-Cram and Grotteria, 2022; Narain and Sangani, 2026). We contribute to this literature by proposing a new approach to measuring policy surprises in oral communication that leverages LLMs to construct counterfactual expectations, and we show that this approach outperforms traditional dictionary-based methods.

Second, we contribute to the growing literature applying LLMs to finance and economics. Recent work has used LLMs to classify text (Hansen and Kazinnik, 2023), extract sentiment and information

[1]For example, text-based measures constructed from policy statements, such as policy stance (Gardner, Scotti, and Vega, 2022; Neuhierl and Weber, 2019), sentiment (Ehrmann and Talmi, 2020), semantic similarity (Correa et al., 2021), readibility (Bholat et al., 2019), and transparency (Hansen, McMahon, and Prat, 2018)—have been shown to affect asset prices and market volatility.

from news earnings announcements and corporate disclosures (Siano, 2025; Jha et al., 2025; Lopez-Lira and Tang, 2023), and simulate human decision-making (Bybee, 2023; Kim, Muhn, and Nikolaev, 2024; Bini et al., 2025; Bhagwat et al., 2026). We extend this literature by using LLMs to generate counterfactual expectations—a distinct application that exploits the model's ability to condition on complex information sets and produce contextually appropriate responses. Our finding that LLM-based surprises explain returns while dictionary-based measures do not highlights the value of semantic understanding in extracting economically relevant information from text.

Third, we contribute to the literature on expectation formation and information processing. A long tradition in macroeconomics and finance studies how agents form and update beliefs in response to new information (Romer and Romer, 2004; Coibion and Gorodnichenko, 2015). The challenge of measuring expectations has historically limited empirical work in this area. Our approach offers a scalable method for constructing counterfactual expectations in settings where agents must interpret complex, high-dimensional information—such as policy communication, earnings calls, or regulatory filings—opening new avenues for research on belief formation and market efficiency.

3 Measuring Policy Surprises with LLMs

Simulating belief formation and computing the associated surprises has long been a fundamental challenge in empirical finance and macroeconomics. Expectations depend on how agents interpret complex, high-dimensional information sets and map them into forecasts about future policy and economic conditions. Traditional empirical methods, such as dictionary-based measures, sentiment indices, and topic models, can only provide coarse and low-dimensional summaries of text. LLMs, by contrast, can interpret and generate language in a way that reflects the conditional reasoning and nuanced interpretation involved in forming expectations. A rapidly growing literature documents that LLMs exhibit human-like information processing and patterns of belief formation, e.g. Bertomeu et al. (2025); Bybee (2023); Kim, Muhn, and Nikolaev (2024).

This section describes how we use an LLM to generate real-time simulated answers and associated policy surprises during FOMC press conferences. We begin by outlining the institutional setting (Section 3.1) and our procedure for time-stamping the press conference Q&A (Section 3.2). We then

detail how we construct the real-time information set and use an LLM to simulate Chair responses (Section 3.3), compare simulated and actual answers (Section 3.4), and define our policy surprise measure (Section 3.5).

3.1 The Setting

Our sample includes all FOMC policy statements and press conferences from 2011 through June 2025. The FOMC holds eight scheduled meetings per year. Since 2011, the Committee has communicated its decisions through a policy statement released at 2:00 pm Eastern Time, followed by a press conference at 2:30 pm. Press conferences were initially held quarterly but have accompanied every meeting since 2019. Both the policy statements and the press conference videos were obtained directly from the Federal Reserve's official website.[2]

We focus on FOMC meetings not only for their central role in monetary policy communication (Gorodnichenko, Pham, and Talavera, 2023; Curti and Kazinnik, 2023), but also because they offer a cleaner information environment than alternative settings such as earnings calls.[3] The FOMC statement establishes a common prior, giving all market participants the same baseline assessment and policy decision. Any information revealed during the subsequent press conference therefore adds to this baseline rather than arriving bundled with other simultaneous disclosures. Moreover, because the press conference takes place during regular trading hours and is fully recorded and time-stamped, belief updating can be observed directly and in real time using high-frequency data.

3.2 Time-Stamping the Press Conference Q&A

We begin by processing the raw FOMC press conference transcripts obtained from PDF files and generating a structured dataset in which each observation corresponds to an individual question or answer, accompanied by speaker identity, textual content, and meeting-level metadata.

[2]The statements are available at `https://www.federalreserve.gov/monetarypolicy/fomccalendars.htm`, and the press conference videos can be found at `https://www.federalreserve.gov/videos.htm`.

[3]Earnings conference calls are typically embedded in bundled disclosure events, where earnings announcements, management guidance, and other firm-specific information are released jointly, making it difficult to isolate the market effects of any single disclosure component (Rogers, Skinner, and Van Buskirk, 2009). Earnings announcements and calls also differ in timing, usually occurring outside regular trading hours. As a result, price reactions cannot be cleanly mapped to specific statements, and investors' expectations before such events are rarely well defined.

To recover the precise timing of each statement, we extract audio from the press conference video and generate word-level timestamps using an automatic speech recognition model.[4] The model returns a sequence of segments, each containing a text fragment with start and end times. We align these segments to the PDF transcripts and convert them to absolute clock time relative to the scheduled start of the press conference (14:30 ET), yielding the precise start time, end time, and duration of each statement.[5]

Panel A of Table 1 reports summary statistics for all FOMC press conferences in our sample, broken down by Chair. The distribution of meetings varies across tenures: Bernanke held 12 press conferences, Yellen 16, and Powell 57. Bernanke's conferences are shorter in total questions and answers but feature slightly longer average question durations. Yellen's conferences have both fewer questions and longer answers, consistent with a more elaborative communication style. Powell's conferences, by contrast, involve the most questions but the shortest answers.

Panel B summarizes the composition of questions posed during press conferences. We classify each question into one or more predefined categories—statement-related, economic-related, policy-related, forward-looking, and tone-related—using a structured prompt applied to GPT models. On average, each meeting includes 4-6 questions directly related to the policy statement, with Bernanke receiving the most and Yellen the fewest. Non-statement questions constitute the bulk of the Q&A, averaging 19-20 per meeting and reaching their highest levels under Powell. Economic questions average roughly 5 per meeting; policy questions average around 12, again highest under Powell. Critical-tone and neutral-tone questions each occur about 11-12 times per meeting.

3.3 Constructing a Real-time Information Set and Simulating Chair Responses

To simulate investors' expectation formation during the FOMC press conference, we first construct a dataset that captures the information set available to market participants at the time of the meeting (Romer and Romer, 2004). The dataset draws on three primary sources. The first source is a

[4]The conference video is first converted into a 16-kHz mono audio waveform using FFmpeg to ensure compatibility with the model's input specifications. We extract timing information using the Whisper large-v3-turbo automatic speech recognition model, which generates word-level timestamps and identifies continuous spoken segments by merging adjacent words that exhibit no silence gap.

[5]To assess reliability, we manually timestamp four press conference videos and compare these human-recorded times with the model-generated timestamps. Deviations are small, typically falling within a one-minute margin, indicating that the automated method captures timing with sufficient precision for our analysis.

panel of macroeconomic indicators obtained from Bloomberg, consisting of real-time (vintage-specific) releases of key variables, along with their historical paths, and survey-based forecasts. The final information set consists of quarterly GDP growth, the unemployment rate, nonfarm payrolls, wage growth, headline and core PCE inflation, and the federal funds rate target. These series provide a quantitative characterization of the economic environment investors would have used to form their pre-meeting expectations. The second source is the Summary of Economic Projections (SEP) released on the FOMC website during projection meetings, from which we extract the Committee's forecasts for growth, inflation, unemployment, and the federal funds rate. These projections offer additional guidance on the Committee's policy outlook and therefore form part of the public information set.

Another crucial component of the information set is the policy statement released at 14:00 ET. This statement communicates the Committee's policy decision and baseline assessment and is observed by market participants prior to the press conference. We include its full text as part of the conditioning information for all simulated answers. Following the release of the statement, the Chair delivers the introductory remarks at the press conference, which we also treat as publicly observed information. In addition to these two sources of initial communication, we incorporate the cumulative sequence of Q&A exchanges that have already occurred in the same press conference. Each prior question and its corresponding answer are included in full and ordered chronologically. This yields an expanding information set that mirrors the real-time flow of information observed by market participants.

Next, we generate simulated Chair responses using large language models, following standard best practices for structured, context-grounded prompting. We specify a clear persona—the Chair of the Federal Reserve responding to a reporter at an FOMC press conference—and condition the model on the full real-time information set described above.

The task is defined explicitly: generate a spoken-style response to the current question that is consistent with the available information and the institutional setting of an FOMC press conference. To limit hallucination and ensure internal consistency, we instruct the model not to rely on any external information beyond the supplied context. Models are further instructed to internally reason through the relevant macroeconomic conditions, policy trade-offs, and risks, while withholding this reasoning from the final output (Li et al., 2025; Kim, Muhn, and Nikolaev, 2024).

Finally, we specify the required output format and style. The model is instructed to return only the Chair's final spoken response, without analytical commentary or extraneous text, and in a tone appropriate for an FOMC press conference. The full prompt is reproduced below.

3.4 The Output

The counterfactual answer for each question is generated using GPT-4o-min. To illustrate the nature of the model output, we present three representative examples drawn from our sample. For each case, we report the original question, the simulated answer produced by the model, and the actual response delivered by the Federal Reserve Chair.

Question: *"Paul Kiernan from Dow Jones Newswires. Thanks for the question. The Dow is down as much as 400 points. What I'm hearing is a reluctance to provide more guidance around the future path of rates. And I'm wondering if that reflects a greater lack of consensus on the Committee. And, you know, how much consensus do you want to see around these decisions, and how split are people about this?"* (Chair Powell's press conference, July 31, 2019)

Simulated Answer: *"The Committee is indeed focused on being data-dependent and assessing evolving economic conditions as we move forward. While there may be differing views among members regarding the appropriate path for rates, we remain committed to making decisions based on a wide range of information, including labor market conditions, inflation pressures, and global developments. It is important to maintain a flexible approach, allowing us to respond to new data as it arrives. Consensus is valuable, but ultimately our decisions reflect the best collective judgment of the Committee based on the current economic landscape."*

Actual Answer: *"You know, you're right. There's a range of views on the Committee, but the Committee is unified—completely unified—on our dedication to making the best policy decisions that we can make. And that means people have a responsibility to do their best thinking and to present that thinking, and I wouldn't have it any other way. In terms of the way forward, we will be monitoring the factors that I mentioned, and we laid that out in the postmeeting statement. That's the roadmap we're going to be following going forward. We're going to be data-dependent. We're going to be, as we always are, doing what we believe we need to do to support the economic expansion."*

Model input (prompt structure)

"role": "system"
"content": You are the Chair of the Federal Reserve responding to a reporter at a press conference. Reconstruct what the Chair would reasonably say given ONLY the provided Context (current FOMC statement, structured macro data, and market conditions).

"role": "system"
"content": privately reason through the policy logic behind the Chair's response, but do NOT reveal that reasoning. Then produce the most likely spoken-style answer that addresses the reporter's question under chair's logic; and simply states the underlying logic. Do NOT start or end with greetings or thanks. Output only the final answer in natural spoken prose (no bullet points).

"role": "system"
"content": FOMC statement released at 14:00 (full text)

"role": "system"
"content": Chair's opening remarks at the beginning of the press conference (full transcript)

"role": "system"
"content": Structured macroeconomic data available at the time of the meeting, including recent releases, revisions, and SEP projections when applicable

"role": "user"
"content": Few-shot example question from a past press conference

"role": "user"
"content": Prior question 1 from the same press conference

"role": "assistant"
"content": Actual Chair response to question 1

...

"role": "user"
"content": Prior question $(n-1)$ from the same press conference

"role": "assistant"
"content": Actual Chair response to question $(n-1)$

"role": "user"
"content": Current question n

Model output

"role": "assistant"
"content": Simulated Chair response to question n

Question: *"Sorry, I had a quick follow-up. Also, in your speech in Philadelphia, you did not say that you felt that it would be appropriate for a rate hike to occur in the coming months. Did you intentionally leave that out?"* (Chair Yellen's press conference, June 15, 2016.)

Simulated Answer: *"The decision on whether to raise rates is made on a meeting-by-meeting basis and depends heavily on incoming data and evolving economic conditions. In my Philadelphia speech, I sought to emphasize the importance of monitoring labor market developments and other economic indicators closely before making any policy decisions. It is essential to determine whether the underlying momentum in the economy is sufficient to warrant a rate increase. For that reason, I did not specify a timeline for a potential rate hike. Our approach remains data-dependent and flexible."*

Actual Answer: *"You know, we do need to make sure that there is sufficient momentum. I don't know what the timetable is going to be to gain that assurance. Every meeting is live; there is no meeting that is off the table—no meeting is out in terms of a possible rate increase. But we really need to look at the data, and I can't prespecify a timetable. So I'm not comfortable saying it's in the next meeting or two, but it could be; it's not impossible. It's not impossible that by July, for example, we would see data that led us to believe that we're on a perfectly fine course, that the earlier data was an aberration, and that other concerns had passed."*

Question: *"Hi, Mr. Chairman, it's Darren Gersh, Nightly Business Report. Your forecast doesn't get back to full employment for four years, so could these new bond purchases go on for years? And can you give us a better idea—if you have specifics in mind—on when you'll know it's time to stop?"* (Chair Bernanke's press conference, September 13, 2012.)

Simulated Answer: *"The Committee is committed to closely monitoring economic conditions, and decisions regarding our bond purchases will depend on incoming data about the labor market and broader economic activity. While our forecast suggests that full employment may not be reached for several years, asset purchases are intended to continue only until we observe substantial improvement in labor market conditions. We will look for broad-based gains in employment, sustained job growth, and strengthening economic momentum. The timing of any reduction in purchases will be entirely data-dependent, and we will communicate our assessment transparently as conditions evolve."*

Real Answer: *"Yes. We'll be looking for signs that the economy is strong enough to promote*

improvement and sustained improvement in labor market conditions and declines in unemployment. I mean, we're not going to be able to sustain purchases until we're all the way back to full employment—that's not the objective. The idea is to quicken the recovery, to help the economy begin to grow quickly enough to generate new jobs and reduce the unemployment rate. So that's the criterion we're looking at."

Across all three examples, the simulated responses closely mirror the economic logic and institutional framing of the actual answers. As highlighted in blue, the model consistently reproduces key elements of Federal Reserve communication: an emphasis on collective judgment, meeting-by-meeting decision-making, and explicit labor-market-based criteria for policy adjustment. Importantly, these similarities arise despite noticeable differences in phrasing and rhetorical style, suggesting that the model internalizes the underlying structure of the Federal Reserve's reaction function rather than merely mimicking surface-level language. This feature is central to our empirical strategy, as it allows us to construct counterfactual policy narratives that preserve economic content while abstracting from Chair-specific communication choices.

3.5 Defining the Policy Surprise

Based on the simulated and actual answers, we define our LLM-based policy surprise as the difference between the policy stance implied by the Chair's realized response and the stance implied by the model-generated counterfactual response. This difference captures the unexpected component of the Chair's communication relative to what would be anticipated given the prevailing information available at the time. Constructing this measure requires a reliable and context-sensitive estimate of the policy stance embedded in each answer. Throughout, we treat the counterfactual response as a benchmark for expected communication, so that the surprise isolates deviations in policy-relevant content rather than differences in rhetorical style.

Following Hansen and Kazinnik (2023), we design a prompt that instructs the LLM to classify each answer into five ordered policy-stance categories: strongly dovish, dovish, neutral, hawkish, and strongly hawkish.[6] The prompt provides explicit definitions for each category in terms of the implied

[6] Hansen and Kazinnik (2023) show that LLMs can classify monetary policy stance with significantly higher accuracy than traditional approaches.

direction of the future policy path. We also include a set of few-shot examples from Romer and Romer (2004, 2023) and Hansen and Kazinnik (2023) to illustrate each stance category. For each answer, the LLM returns (i) a categorical label; (ii) a numeric stance score associated with the five categories $\{-1, -0.5, 0, 0.5, 1\}$; and (iii) a brief reasoning statement explaining the classification. This procedure ensures that both actual and simulated answers are evaluated using an identical classification rule, thereby producing the corresponding LLM-based policy surprise.

The following excerpt provides a stylized version of the prompt used to extract policy stance from each press-conference answer.

```
You are an expert in Federal Reserve communication analysis. Read the following answer
from a Federal Reserve press conference.

# Definitions of policy stance categories:

- Dovish: Strongly supports easier policy, stimulus, or lower interest rates.

- Mostly Dovish: Slightly favors easing, highlights growth or employment concerns.

- Neutral: No clear directional bias; descriptive or factual.

- Mostly Hawkish: Slightly favors tightening, mentions inflation vigilance.

- Hawkish: Strongly favors tighter policy or higher rates to contain inflation.

# Step-by-step reasoning (Chain of Thought)

1. Identify whether the statement emphasizes inflation control or growth support.

2. Detect if it suggests tighter (hawkish) or easier (dovish) future policy.

3. Apply definition rules above to avoid confusion between tone and policy direction.

4. Conclude the stance category and score.

Few-shot examples: {Insert Few-shot Examples Here}

Answers: {Insert Answer Here} Classify the monetary policy stance of each answer into
one of five categories.

Respond in valid JSON:

{ "category": "dovish" / "mostly dovish" / "neutral" / "mostly hawkish" / "hawkish",

"score": -1 / -0.5 / 0 / 0.5 / 1,

"reasoning": "brief explanation" }
```

We present several illustrative examples to demonstrate how our policy surprise measure operates in practice by comparing simulated, expectation-based responses with the Chair's actual answers. In the first case, the simulated and actual responses are directionally similar, but the Chair's explicit dismissal of near-term balance-sheet tightening conveys a more dovish signal than implied by the baseline, generating a small dovish surprise. In the second case, the simulated response emphasizes consumer relief and policy flexibility, whereas the Chair forcefully re-centers the discussion on maintaining restrictive rates to restore price stability, producing a large hawkish surprise due to a clear reversal in stance. In the third case, both responses acknowledge ongoing disinflation alongside persistent inflation risks, resulting in close alignment and a neutral surprise. Overall, these examples indicate that the surprise measure captures meaningful differences in policy stance beyond topic coverage, including differences in emphasis and implied reaction functions.

Question: *Victoria Guida with Politico. I wanted to ask about the balance sheet. You all have left open the possibility that you might sell mortgage-backed securities, but we've seen significant slowing in the housing market. Mortgage rates have gone up significantly, and I'm just wondering whether conditions there might affect your plans for how quickly you have the runoff on the MBS side. (Chair Powell's press conference, September 21, 2022.)*

Simulated Answer (Neutral 0): *The housing market is indeed experiencing significant slowing, and we are closely monitoring those conditions as part of our overall assessment. While we have left open the possibility of selling mortgage-backed securities, our decisions regarding the balance sheet will be made on a meeting-by-meeting basis and will be data-dependent. We will consider a wide range of factors, including the state of the housing market and broader economic conditions, before making any decisions. Our primary focus remains on achieving our inflation and employment goals, and we will adjust our approach as necessary based on incoming information.*

Real Answer (Dovish -1): *So we, what we said, as you know, was that we would consider that once balance sheet runoff is well under way. I would say it's not something we're considering right now and not something I expect to be considering in the near term. It's just, it's something I think we will turn to, but that time, the time for turning to it, has not come and is not close.*

Question: *I think consumers were looking for relief on rates when it comes to mortgages—car*

loans. Small businesses want to take out more manageable loans. When you look at the cumulative inflation over the past five years, prices have risen over 20 percent—it's been a rough road. So what is the tipping point, then, for the wait-and-see approach in terms of how much it's going to help versus when it hurts the American consumer? (Chair Powell's press conference, June 18, 2025.)

Simulated Answer (Mostly Dovish -0.5): *The Committee is very aware of the pressures that inflation and interest rates place on consumers and small businesses. Our current approach is to maintain the federal funds rate while we assess incoming data and the evolving economic outlook. We believe this stance allows us to respond effectively to any emerging risks while supporting maximum employment and price stability. As we gather more information on inflation trends and labor market conditions, we will be prepared to adjust our policy as necessary to ensure we are not hindering economic growth or putting undue strain on consumers. The decision-making process will remain data-dependent, and we will continue to monitor the situation closely.*

Real Answer (Hawkish 1): *Well, I mean, we're trying to restore price—the best thing we can do for the public that we serve is restore price stability. If we can, and we will, restore price stability, meaning 2 percent inflation on a durable, sustainable basis—and also maximum employment—and if we restore those things, that's the best thing. So, in the meantime, we have to keep rates high to get inflation all the way down. They're not very high—let's be honest. I would say policy is modestly restrictive. If you look at the economy, it's not performing as though it were under very restrictive monetary policy. …*

Question: *Howard Schneider from Reuters. So I believe, 9 or 10 times, a process that you felt was—I forget the word you used—but gratefully under way, or something like that. Is disinflation still occurring in the U.S. today? (Chair Powell's press conference, March 22, 2023.)*

Simulated Answer (Mostly Hawkish 0.5): *Disinflation is indeed a process that we are observing, but it is important to note that inflation remains elevated. Recent data indicates that while inflation pressures have moderated somewhat since last year, they are still running high. Our commitment to achieving the 2 percent inflation target remains strong, and we will continue to monitor incoming data closely. The path to returning inflation to our target is likely to be bumpy, and our policy decisions will be guided by the evolving economic conditions and the implications for inflation*

and the labor market.

Real Answer (Mostly Hawkish 0.5): *Yes. I mean, what actually happened, …Goods inflation has been coming down now for six months; it's proceeding more slowly than we would have liked, but it's certainly proceeding. …Where we didn't have in February, and we still don't have now, is a sign of progress in the nonhousing services sector. And that is—that's just something that will have to come through softening demand and perhaps some softening in labor market conditions. We don't see that yet. And that's of course 56 percent of the index. So the story is pretty much the same. I will say that the inflation data that we got, to your point, really pointed to stronger inflation.*

4 Assessing the Economic Content of Policy Surprises

This section examines the economic content of deviations between realized and counterfactual press-conference responses. We do so by examining (i) the relationship between actual and simulated policy stances at the meeting level, (ii) heterogeneity in the distribution of press-conference policy surprises across Federal Reserve Chairs, and (iii) the association between our LLM-based measure and standard proxies for uncertainty and monetary policy shocks, including the VIX and the policy surprise measure of Nakamura and Steinsson (2018).

Figure 1 (a) compares the policy stance extracted from actual press-conference answers with the stance extracted from simulated counterfactual answers. For each FOMC press conference, we compute the average stance score from the realized Q&A and plot it on the horizontal axis, while the vertical axis reports the corresponding average stance score implied by the model-generated responses. Each observation, therefore, represents a meeting-level pair of actual and simulated policy stances.

The plot shows that meetings in which the Chair's actual answers were more dovish (or more hawkish) tend to be meetings in which the simulated answers are also more dovish (or more hawkish). The clustering of points around the regression line suggests that the simulation does not merely replicate idiosyncratic speaking styles but instead reflects underlying policy fundamentals. Overall, the strong positive correlation between the actual and simulated stance scores indicates that the LLM-based simulation effectively captures the systematic component of policy communication conveyed during press conferences.

Figure 1 (b) further examines the distributional properties of our press-conference policy surprise across different Federal Reserve Chairs. Bernanke's distribution is highly peaked and tightly concentrated around zero, with thin tails and minimal skewness, indicating that most press-conference surprises are extremely small and that large deviations are rare. Yellen's distribution remains centered near zero but is flatter and slightly more dispersed, with moderately heavier tails, suggesting a higher frequency of moderate surprises. By contrast, Powell's distribution is the least peaked and displays the thickest tails, indicating a substantially higher frequency of large surprises; the modest extension of the right tail reflects a tendency toward occasional large hawkish surprises.

Figure 1 (c) and (d) assess whether the dispersion in our LLM-based press-conference policy surprise is economically meaningful by relating it to two established measures of uncertainty and monetary policy shocks. Panel (c) plots the absolute average policy surprise during the press-conference Q&A against the level of market uncertainty, proxied by the VIX index. The positive relationship indicates that press conferences associated with heightened financial-market uncertainty tend to feature larger deviations between realized and counterfactual policy communication. This pattern is economically intuitive. When uncertainty is elevated, investors are more sensitive to nuanced signals about the policy outlook, and Chairs face greater difficulty in anchoring expectations through communication alone. As a result, small differences in wording, emphasis, or framing during the Q&A can generate larger perceived policy surprises. The positive slope, therefore, suggests that our measure captures meaningful variation in the informational content of policy communication precisely when uncertainty is high, and communication is most consequential.

Panel (d) relates the absolute press-conference policy surprise to the absolute monetary policy surprise constructed by Nakamura and Steinsson (2018). Here again, we find a positive association. Meetings that generate larger conventional policy surprises also tend to be meetings in which the press-conference Q&A exhibits larger deviations between actual and counterfactual policy stances. This result reinforces the interpretation that our LLM-based measure is not merely capturing linguistic noise or stylistic variation, but instead reflects economically relevant information about unexpected policy developments. When the policy decision or its implications surprise markets, the subsequent press conference naturally becomes a key channel through which new information is revealed, clarified,

or inadvertently amplified.

Taken together, these results provide external validation for our approach. The correlation between actual and simulated stances, the heterogeneity across Chairs, and the positive relationships with both market uncertainty and conventional policy surprise measures all indicate that deviations between realized and counterfactual press-conference answers reflect economically meaningful variation in central bank communication.

5 Q&A Policy Surprises and Stock Returns

5.1 Case Studies of Intraday Market Reactions

Before turning to systematic evidence, we present two case studies illustrating how policy communication relates to intraday equity returns during FOMC press conferences. Figure 2 displays minute-by-minute cumulative SPY returns alongside two measures: the policy stance of the Chair's actual answers and the corresponding policy surprise.

The first case study corresponds to the December 12, 2012 press conference, held during a period of aggressive monetary accommodation under Chair Bernanke. In this episode, SPY returns decreased over the press-conference window even though the cumulative policy stance implied by the Chair's answers was dovish (Panel (a)). Viewed through the lens of expressed policy stance alone, this pattern appears puzzling: equity prices dropped despite communication leaning towards accommodative policies. However, the corresponding surprise panel (Panel (b)) reveals that the same interval was dominated by hawkish surprises—the Chair's realized answers were tighter than what market participants would have expected given the prevailing information set. The decline in equity prices coincided not with a hawkish stance in absolute terms, but with communication that was *more* hawkish than anticipated.

The second case study is drawn from the May 4, 2022 press conference under Chair Powell. This meeting occurred amid a rapid tightening cycle as the Federal Reserve responded to elevated inflation. During the press conference, SPY returns increased even though the cumulative expressed policy stance became progressively more hawkish over the sequence of answers. The surprise panel shows that this increase coincided with dovish surprises: the Chair's realized responses were more accommodative than

the counterfactual answers generated by our model. Once again, prices responded to the unexpected component of communication rather than to its raw content.

These two episodes illustrate a central theme of our analysis. Equity prices during FOMC press conferences do not move mechanically in the direction of the expressed policy stance; instead, they respond to the surprise component of policy communication relative to market expectations. We now examine whether this pattern holds systematically across the full sample.

5.2 Intraday Market Responses to Policy Surprises using Local Projections

To assess whether the patterns documented in the case studies hold systematically, we employ a local projection framework in line with Gorodnichenko, Pham, and Talavera (2023). For each horizon h, we estimate:

$$\text{SPYReturn}_{t_a \to t_{a+h}, m} = \beta_0^{(h)} + \beta_1^{(h)} \, \text{PolicySurprise}_{t_0 \to t_a, m} + \lambda_m + \gamma_a + X_t + \varepsilon_m^{(h)}, \tag{1}$$

where the dependent variable $\text{SPYReturn}_{t_a \to t_{a+h}, m}$ is the cumulative return of the SPY ETF from the end of the question a at time t_a to time t_{a+h} in meeting m, where h ranges from 1 to 15 minutes. The key explanatory variable $\text{PolicySurprise}_{t_0 \to t_a, m}$ measures the time-weighted cumulative policy surprise between t_0 and t_a. We aggregate surprises cumulatively rather than at the individual-answer level because market participants plausibly integrate information progressively over the press conference, updating beliefs as each answer reinforces or refines the signals conveyed in earlier responses. Within this cumulative measure, weights are proportional to the duration of each answer, reflecting the intuition that longer responses—which typically elaborate on policy rationale or address complex questions—convey more information than brief clarifications.

The coefficient $\beta_1^{(h)}$ captures the horizon-specific response of equity returns to policy communication surprises, allowing the dynamic effect to vary flexibly across horizons without imposing parametric restrictions on the adjustment path. The specification includes meeting fixed effects λ_m and question-rank fixed effects γ_a, which control for systematic differences across press conferences and for predictable variation in market responses associated with the ordering of questions. The vector X_t includes lagged returns, volatility, and the realized content of the Chair's answers, such as stated

sentiment, policy stance, economic assessments, and the frequency of surprise-related language in the statement (Gómez-Cram and Grotteria, 2022). These measures are constructed cumulatively over the press conference and aggregated as time-weighted averages across answers—in line with the construction of the PolicySurprise variable.[7] Standard errors are clustered at the meeting level to account for within-conference dependence.

The solid line in Figure 3 (a) plots the estimated cumulative response of SPY returns to a one-unit increase in the LLM-based policy surprise measure—the $\beta_1^{(h)}$ coefficients—with shaded areas denoting 90% confidence intervals. It shows that equity prices respond negatively to hawkish policy surprises and positively to dovish surprises, with responses that are statistically significant over short horizons. The adjustment occurs rapidly within the first few answers and accumulates over the course of the press conference, indicating that markets incorporate unexpected information conveyed during the Q&A in real time. In terms of magnitude, the estimated cumulative responses of 0.0005–0.001, corresponding to 5–10 basis points, are comparable to typical minute-by-minute fluctuations in SPY returns and therefore economically meaningful in a high-frequency setting.[8]

By contrast, Panel (b) shows that the response of equity returns to the *level* of policy stance conveyed in the Chair's realized answers is weak and largely statistically insignificant across horizons. An exception arises at the 2-minute horizon, where a small but statistically significant response appears with a sign opposite to the expected one. Taken together, these results indicate that equity prices respond primarily to the unexpected component of monetary policy communication rather than to the direction of expressed policy stance.

As a complement to the local projection plots reported above, Table 3 presents coefficient estimates for our baseline specification described in Equation (1) at the 5-minute (Panel A) and 10-minute (Panel B) horizons to assess the sensitivity of our coefficient estimates to different control variables. Columns (1)–(5) report results from progressively augmented specifications that sequentially introduce control variables summarizing the realized content of the Chair's answers—namely policy stance, economic assessment, and sentiment—as well as statement-based policy surprises D_t. All specifications include

[7] Variable definitions are provided in Table 2.

[8] In our sample, the standard deviation of one-minute SPY returns is approximately 0.00047, implying that a typical five-minute price fluctuation is on the order of 0.001 under standard time-scaling. The estimated cumulative responses, therefore, represent a substantial fraction of normal high-frequency price variation.

lagged returns, past volatility, conference fixed effects, and question rank fixed effects. To facilitate interpretation, SPY returns are multiplied by 100 and expressed in percentage points.

Starting from Panel A, the coefficient on PolicySurprise is negative and statistically significant across all specifications, ranging from -0.069 in column (1) to -0.137 in column (5). These estimates indicate that a one-unit increase in cumulative hawkish surprise is associated with a 7–14 basis point decline in SPY returns over the subsequent five minutes. The magnitude and significance of the coefficient remain stable as we add controls, suggesting that the predictive content of the LLM-based surprise measure is not subsumed by other features of the Chair's communication. Panel B extends the analysis to the 10-minute horizon. The coefficient on PolicySurprise strengthens in magnitude, ranging from -0.129 to -0.188, and remains significant at the 5% level across all specifications.

Overall, these results show that intraday equity price movements during FOMC press conferences are closely related to policy surprises, suggesting that our measure captures economically relevant information in monetary policy communication.

5.3 Heterogeneity in Market Responses

To sharpen the economic interpretation of our results, we conduct two additional tests. First, we condition on question characteristics to identify which dimensions of the Q&A are most consequential for price formation. Second, we examine whether market responses vary with the informational environment. If press conferences operate by shifting beliefs about the future policy path, their effects should be larger when prior expectations are less anchored.

5.3.1 Question Characteristics

We start by conditioning on question characteristics because the Chair's answers are shaped by the questions asked. Each response addresses the specific issue raised rather than offering a free-form policy assessment, so the information observable to markets depends critically on the nature of the questions.

Figure 4 examines how the equity market's response to press-conference policy surprises varies across different types of questions. Panels (a) and (b) distinguish between forward-looking and non-

forward-looking questions. Policy surprises are associated with statistically significant market responses when questions explicitly concern future policy paths. For forward-looking questions, more hawkish surprises are followed by declines in equity prices, whereas more dovish surprises are accompanied by positive returns. By contrast, for non-forward-looking questions, the response of equity returns is statistically significant only at intermediate horizons and is smaller in magnitude. This pattern indicates that market participants primarily react to information that directly refines expectations about future monetary policy, rather than to backward-looking or purely descriptive discussion.

Panels (c) and (d) split the sample by whether questions relate directly to the FOMC statement. For statement-related questions, policy surprises are priced promptly and persistently throughout the press-conference window. For non-statement-related questions, the response emerges only after approximately five minutes. This asymmetry likely reflects differences in how easily markets can interpret the Chair's answers. When questions pertain to the statement, responses are anchored in an existing policy framework, allowing immediate incorporation of surprises into prices. When questions address topics outside the statement, markets may require additional context to assess the policy implications, delaying the response.

Panels (e) and (f) distinguish between policy-related and economy-related questions. Policy-related questions generate statistically significant responses throughout the press-conference window. These questions directly concern monetary policy decisions, so the Chair's answers map readily into expectations about the future policy path. For economy-related questions, the response is insignificant at short horizons but emerges after approximately five minutes and grows at longer horizons. Such questions elicit broader macroeconomic assessments rather than direct policy guidance, requiring markets to infer the policy implications—a process that delays but does not eliminate the pricing of surprises.

Overall, these heterogeneity results underscore the importance of the informational context in which monetary policy communication occurs. When questions are forward-looking, policy-related, or directly tied to the statement, surprise components in the Chair's answers are more likely to induce coordinated revisions in market expectations and to be incorporated into asset prices. When questions are less connected to these focal dimensions, the information is harder to map into the policy outlook, generating weaker or more delayed responses.

5.3.2 Information Environment

If press conferences operate by shifting beliefs about the future policy path, their effects may depend on the prevailing level of market uncertainty. The theoretical prediction, however, is ambiguous. On one hand, higher uncertainty may increase the value of information and amplify responses to policy signals (Andrei, Friedman, and Ozel, 2023; Van Nieuwerburgh and Veldkamp, 2010). On the other hand, elevated volatility generates noise that can obscure the pricing of nuanced information, delaying the observed response.

Panels (a) and (b) of Figure 5 split the sample by the VIX. During low-VIX periods, equity returns respond negatively to policy surprises, with statistically significant effects emerging at short horizons. During high-VIX periods, the point estimates are larger in magnitude, but the response is not statistically significant until approximately six minutes into the press conference. This pattern suggests that policy surprises are incorporated into prices more slowly when market volatility is elevated—consistent with noise obscuring the signal at short horizons. As background volatility dissipates, the cumulative effect of policy communication becomes apparent, and the response strengthens at longer horizons. These results highlight a trade-off between the magnitude and the speed of information incorporation. High-uncertainty environments may amplify the eventual response to policy surprises, but the same volatility that reflects uncertainty also impedes rapid price adjustment.

Panels (c) and (d) split the sample by the tone of the accompanying FOMC statement. Following dovish statements, policy surprises generate large and statistically significant responses throughout the press-conference window. Following hawkish statements, the response is substantially smaller and largely indistinguishable from zero. This asymmetry suggests that the scope for belief updating varies with the baseline policy signal. When the statement is dovish, the future policy path remains uncertain—markets may anticipate eventual tightening but face ambiguity about its timing and pace, leaving room for the Chair's answers to shift expectations. When the statement is hawkish, prior beliefs are more tightly anchored around a restrictive policy stance, limiting the incremental information revealed during the Q&A. These results are consistent with the view that market responses are stronger when prior expectations leave greater scope for updating.

5.4 Conditioning Information and LLM versus Dictionary Methods

This section examines which features of our methodology are important for explaining stock market reactions to press conference Q&As. We consider two dimensions. First, we vary the conditioning information provided to the LLM to assess which elements are necessary for constructing surprises that explain returns. Second, we compare our LLM-based approach to dictionary-based alternatives to evaluate whether semantic understanding offers advantages over keyword counting in capturing policy signals from spoken communication.

5.4.1 Alternative Prompt Specifications

Providing accurate context is important for obtaining reliable responses from LLMs (Bybee, 2023). Our baseline prompt, therefore, includes comprehensive information about the macroeconomic environment and recent policy actions, replicating the information set available to market participants at the time of each press conference. To assess sensitivity to this design choice, we re-estimate the local-projection specifications using alternative prompts that omit macroeconomic context and the latest FOMC statements.

Figure 6 reports local-projection estimates under these restricted prompts. Panel (a) omits macroeconomic background and SEP forecasts; Panel (b) further excludes the policy statement, leaving only the press conference introduction and prior Q&A.

Under both specifications, the estimated return responses are no longer statistically significant at any horizon. This finding demonstrates that both macroeconomic context and the policy statement are necessary for constructing surprises that explain stock market reactions. Our surprise measure is defined as a deviation from a counterfactual expectation, and the economically relevant counterfactual is one formed using the same state variables that market participants use to forecast the future policy path. When either component is omitted, the LLM-generated counterfactual is no longer anchored to investors' information set—and the resulting "surprise" fails to predict returns.

5.4.2 LLM versus Dictionary Methods

To assess whether the predictive power of our surprise measure stems from the LLM's semantic understanding rather than simple keyword detection, we construct alternative policy surprises using two established dictionary-based approaches and compare their ability to explain intraday returns.

First, we follow the dictionary-based approach of Gómez-Cram and Grotteria (2022). For each answer, we first normalize the text by removing extra whitespace to ensure consistent phrase matching. We then apply the dovish–hawkish dictionary of Neuhierl and Weber (2019), which classifies policy-relevant phrases as either dovish or hawkish, and count the number of occurrences of each phrase in the answer text. Summing across phrases yields the total number of dovish and hawkish occurrences within each answer.

Using these counts, we compute an answer-level policy stance as the normalized difference between dovish and hawkish phrase frequencies. Formally, we compute

$$\text{Stance}_a = \frac{D_a - H_a}{D_a + H_a}, \tag{2}$$

where D_a and H_a denote the number of dovish and hawkish phrase occurrences in answer a, respectively; if an answer contains no dictionary phrases, the stance is set to zero.

We apply this procedure both to the Chair's realized answers and to the corresponding counterfactual answers generated by the LLM under the same information set. Policy surprises are defined as the difference between realized and counterfactual stances and then aggregated across the press conference using a time-weighted cumulative scheme, as in our baseline results.

We also construct a policy stance measure based on a compositional term-modifier dictionary following Apel, Grimaldi, and Hull (2022). This dictionary separates economically meaningful policy terms—such as inflation, economic activity, and employment—from modifiers that convey the directional inclination of monetary policy. Both terms and modifiers may include wildcard expressions to allow for flexible morphological variants. For each answer, we split the text into sentences and tokenize each sentence into individual words. For every term-modifier pair in the dictionary, we search within each sentence for occurrences of the term and the corresponding modifier. A term-modifier pairing is counted as valid if the modifier appears within a symmetric window of seven words before or after

the term in the same sentence. Each valid pairing counts at most once per sentence, preventing repeated matches within the same local context. We then sum the number of valid hawkish and dovish term-modifier occurrences across all sentences within an answer. Using these counts, we compute an answer-level policy stance as the normalized difference between dovish and hawkish term-modifier frequencies, as defined in Equation (2).

Figure 7 reports local-projection estimates of the intraday response of SPY returns to press-conference policy surprises constructed using dictionary-based text measures. Panel (a) employs the fixed-phrase dovish–hawkish dictionary of Neuhierl and Weber (2019). Under this specification, we find no statistically significant response of SPY returns across horizons. This null result likely reflects the reliance of fixed-phrase dictionaries on predefined policy keywords, which are less likely to appear verbatim in spontaneous spoken answers that often involve paraphrasing, hedging, and indirect language.

Panel (b) instead uses the compositional term–modifier dictionary of Apel, Grimaldi, and Hull (2022). Using this measure, equity returns exhibit a statistically significant response to policy surprises at short horizons, concentrated within the first eight minutes of the press conference, although the magnitude of the response is smaller than in the baseline LLM-based results. The effect dissipates quickly and does not persist at longer horizons. Although the modifier approach offers greater linguistic flexibility and therefore captures a limited amount of market-relevant variation in early responses, it still falls short of fully capturing the nuanced information conveyed during press conferences.

Overall, the dictionary-based measures appear too restrictive to systematically recover the policy-relevant surprises embedded in spoken communication, underscoring the advantage of the LLM-based approach in this context.

6 Robustness Checks

6.1 Alternative Weighting and Aggregation of the Policy Surprises Measure

Figure 8 repeats our baseline estimates under two alternative specifications. Panel (a) uses a simple mean cumulative measure of policy surprise, constructed by assigning equal weight to each answer and averaging surprise content across responses since the beginning of the press conference. Relative to the

time-weighted aggregation used in the main analysis, this measure treats each answer equally. Empirically, Panel (a) shows that equity returns respond negatively to hawkish surprises and positively to dovish surprises, consistent with our baseline results. This directional consistency across weighting schemes indicates that our main findings are not driven by the particular aggregation method. However, the estimated responses are smaller in magnitude under equal weighting, suggesting that answer duration proxies for information content: longer responses, which receive greater weight in the baseline specification, appear to convey more policy-relevant signals.

Panel (b) examines whether equity prices respond to policy information conveyed in individual surprises rather than cumulative ones. In this "flow" specification (Gorodnichenko, Pham, and Talavera, 2023), the policy surprises are measured at the level of each individual response, relating the surprise content of a single answer to the price movement that immediately follows. Panel (b) shows that the estimated response of SPY returns is not statistically significant at any horizon. The absence of a significant response suggests that the information conveyed in any single answer is too noisy to generate a detectable price reaction on its own. This result confirms the rationale for our baseline specification: market participants integrate information progressively over the press conference, and the cumulative aggregation captures this belief-updating process while filtering out idiosyncratic variation across individual responses.

6.2 Knowledge Cutoff

One potential concern is that the LLM may implicitly rely on information learned during training to simulate the Chair's answers, even when such information would not have been available in real time (Breitung and Muller, 2025; De Kok, 2025). If the model has been exposed to transcripts or summaries of later FOMC press conferences during training, its simulated responses could partially reflect future realizations rather than genuine inference based on the contemporaneous information set, leading to a form of look-ahead bias.

To address this concern, we restrict the sample to press conferences held after October 1, 2023, the reported knowledge cutoff date of the GPT-4o-mini model, ensuring that simulated answers cannot draw on post-event information embedded in the training data. Figure 9 re-estimates the baseline

specification on this restricted sample. Panel (a) shows that the estimated responses remain statistically significant and closely resemble those obtained in the full sample. In particular, equity prices continue to respond negatively to hawkish surprises and positively to dovish surprises during the press conference window. Panel (b) reports corresponding estimates using realized policy stance classifications. In this case, the estimated responses are not statistically significant. Consistent with our baseline results, intraday return responses are primarily driven by unexpected policy information.

Overall, that these effects hold in the post-cutoff sample mitigates concerns that the main results are driven by potential look-ahead bias arising from model memorization of future policy outcomes or press-conference content.

6.3 Alternative Dependent Variables

In the main analysis, we focus on equity returns during the press conference window to capture how policy surprises are incorporated into prices in real time. While returns provide a direct measure of directional price adjustment, they reflect the net outcome of multiple channels and may therefore mask how press-conference communication affects investors' beliefs along other dimensions. To complement the return-based evidence, we examine alternative market outcomes that capture distinct aspects of expectation updating. Specifically, we consider market volatility, measured by changes in the VIX, and exchange rates, measured by changes in the EUR/USD rate, following Curti and Kazinnik (2023). We implement this extension by re-estimating the local-projection specifications with these alternative dependent variables.

Figure 10 reports local-projection estimates of the intraday response of market volatility to press-conference policy surprises. Panel (a) plots the response of the VIX to the time-weighted cumulative policy surprise measure. The estimates show that hawkish policy surprises are followed by increases in the VIX, with the response statistically significant during the first twelve minutes of the press conference. This pattern indicates that unexpected signals of tighter monetary policy are associated with heightened market uncertainty about the future policy path. Panel (b) plots the response of the EUR/USD exchange rate. Hawkish policy surprises are associated with a statistically significant decline in the EUR/USD rate over short horizons, indicating an immediate adjustment in foreign

exchange markets following press-conference communication.

Taken together with the equity return evidence documented earlier, these results show that press-conference policy surprises are reflected across equity prices, market-implied volatility, and exchange rates within the press conference window. Overall, the evidence indicates that the policy surprise measure is informative for understanding high-frequency market responses during FOMC press conferences.

7 Conclusions

Central bank communication is a primary tool of monetary policy. Given its significant influence on financial markets and the real economy, policy announcements and subsequent Q&A sessions are closely monitored by investors and have been widely studied in the academic literature.

This paper proposes a new methodology to identify monetary policy surprises in central bank communication. Conditioning on the same real-time information set available to market participants, we use large language models to generate counterfactual press-conference responses by the Federal Reserve Chair and measure policy surprises as deviations between realized and simulated answers.

Our central finding is that equity markets respond systematically to these surprises during FOMC press conferences. Hawkish policy surprises (i.e., when the Chair's actual response is more hawkish than the LLM counterfactual) are associated with statistically significant and economically meaningful declines in equity prices, while dovish surprises are associated with price increases. The response emerges rapidly after each question and lasts for several minutes. Crucially, these effects are not driven by the absolute stance of the Chair's communication, but by departures from expectations, confirming that markets price surprises rather than tone or direction per se.

Heterogeneity tests confirm that the market response to policy surprises is driven by expectation formation and belief updating. Policy surprises generate stronger market responses when the Chair's answers most directly inform expectations about the future policy path, i.e., when questions are forward-looking, policy-related, or directly tied to the FOMC statement. During periods of elevated market uncertainty, policy surprises have larger effects, but the response emerges more slowly, consistent with noise confounding the signal at short horizons. Finally, following dovish FOMC state-

ments, press-conference surprises generate larger price reactions than following hawkish statements, suggesting that the scope for belief updating is greater when prior expectations are less anchored. Finally, when we omit macroeconomic context or exclude the policy statement from the prompt, the estimated return responses are no longer statistically significant at any horizon.

Overall, our results show that large language models provide a powerful new tool for measuring expectation-based policy surprises in real time, opening the door to a richer analysis of how central bank communication shapes financial markets.

References

Andrei, D., H. Friedman, and N. B. Ozel. 2023. Economic uncertainty and investor attention. *Journal of Financial Economics* 149:179–217.

Apel, M., M. B. Grimaldi, and I. Hull. 2022. How Much Information Do Monetary Policy Committees Disclose? Evidence from the FOMC's Minutes and Transcripts. *Journal of Money, Credit and Banking* 54:1459–90.

Bekaert, G., M. Hoerova, and M. L. Duca. 2013. Risk, uncertainty and monetary policy. *Journal of Monetary Economics* 60:771–88.

Bernanke, B. S., and K. N. Kuttner. 2005. What explains the stock market's reaction to federal reserve policy? *Journal of Finance* 60:1221–57.

Bertomeu, J., Y. Lin, Y. Liu, and Z. Ni. 2025. The impact of generative AI on information processing: Evidence from the ban of ChatGPT in Italy. *Journal of Accounting and Economics* 101782.

Bhagwat, V., J. A. Cookson, C. Dim, and M. Niessner. 2026. The market's mirror: Revealing investor disagreement with LLMs. Working Paper.

Bholat, D., N. Broughton, J. Ter Meer, and E. Walczak. 2019. Enhancing central bank communications using simple and relatable information. *Journal of Monetary Economics* 108:1–15.

Bini, P., L. W. Cong, X. Huang, and L. J. Jin. 2025. Behavioral Economics of AI: LLM Biases and Corrections. Working Paper.

Breitung, C., and S. Muller. 2025. Global Business Networks. *Journal of Financial Economics* 166:104007.

Bybee, J. L. 2023. The Ghost in the Machine: Generating Beliefs with Large Language Models. *Working Paper* .

Cieslak, A., A. Morse, and A. Vissing-Jorgensen. 2019. Stock returns over the FOMC cycle. *Journal of Finance* 74:2201–48.

Coibion, O., and Y. Gorodnichenko. 2015. Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts. *American Economic Review* 105:2644–78.

Correa, R., K. Garud, J. M. Londono, and N. Mislang. 2021. Sentiment in central banks' financial stability reports. *Review of Finance* 25:85–120.

Curti, F., and S. Kazinnik. 2023. Let's face it: Quantifying the impact of nonverbal communication in FOMC press conferences. *Journal of Monetary Economics* 139:110–26.

De Kok, T. 2025. ChatGPT for Textual Analysis? How to Use Generative LLMs in Accounting Research. *Management Science* 71:7888–906.

Ehrmann, M., and J. Talmi. 2020. Starting from a blank page? semantic similarity in central bank communication and market volatility. *Journal of Monetary Economics* 111:48–62.

Gardner, B., C. Scotti, and C. Vega. 2022. Words speak as loudly as actions: Central bank communication and the response of equity prices to macroeconomic announcements. *Journal of Econometrics* 231:387–409.

Golez, B., and B. Matthies. 2025. Fed information effects: Evidence from the equity term structure. *Journal of Financial Economics* 165:103988.

Gómez-Cram, R., and M. Grotteria. 2022. Real-time price discovery via verbal communication: Method and application to Fedspeak. *Journal of Financial Economics* 143:993–1025.

Gorodnichenko, Y., T. Pham, and O. Talavera. 2023. The Voice of Monetary Policy. *American Economic Review* 113:548–84.

Gurkaynak, R., A. H. Kara, B. Kisacikoglu, and L. S. Seok. 2021. Monetary policy surprises and exchange rate behavior. *Journal of International Economics* 130:1033–44.

Hansen, A. L., and S. Kazinnik. 2023. Can ChatGPT Decipher Fedspeak? *Working Paper* .

Hansen, S., M. McMahon, and A. Prat. 2018. Transparency and deliberation within the FOMC: A computational linguistics approach. *Quarterly Journal of Economics* 133:801–70.

Jha, M., J. Quian, M. Weber, and B. Yang. 2025. Chatgpt and corporate policies. Working Paper.

Kim, A. G., M. Muhn, and V. V. Nikolaev. 2024. Financial Statement Analysis with Large Language Models. *Working Paper* .

Kuttner, K. N. 2001. Monetary policy surprises and interest rates: Evidence from the fed funds futures market. *Journal of Monetary Economics* 47:523–44.

Li, K., F. Mai, R. Shen, C. Yang, and T. Zhang. 2025. Dissecting corporate culture using generative AI. *Review of Financial Studies* Forthcoming.

Lopez-Lira, A., and Y. Tang. 2023. Can chatgpt forecast stock price movements? return predictability and large language models. *Working Paper* .

Loughran, T., and B. Mcdonald. 2011. When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance* 66:35–65.

Nakamura, E., and J. Steinsson. 2018. High-Frequency Identification of Monetary Non-Neutrality: The Information Effect. *Quarterly Journal of Economics* 133:1283–330.

Narain, N., and K. Sangani. 2026. The market impact of fed communications: The role of the press conference. Working Paper.

Neuhierl, A., and M. Weber. 2019. Monetary policy communication, policy slope, and the stock market. *Journal of Monetary Economics* 108:140–55.

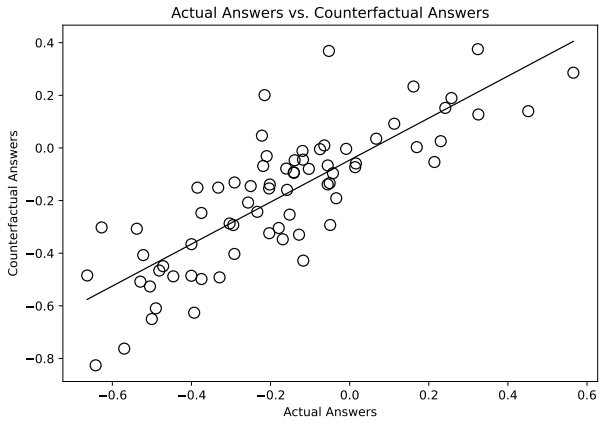
Rogers, J. L., D. J. Skinner, and A. Van Buskirk. 2009. Earnings guidance and market uncertainty. *Journal of Accounting and Economics* 48:90–109.

Romer, C. D., and D. H. Romer. 2004. A New Measure of Monetary Shocks: Derivation and Implications. *American Economic Review* 94:1055–84.

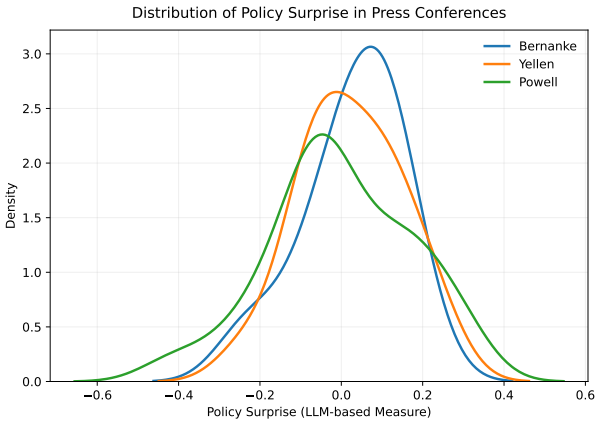
———. 2023. Presidential Address: Does Monetary Policy Matter? The Narrative Approach after 35 Years. *American Economic Review* 113:1395–423.

Siano, F. 2025. The News in Earnings Announcement Disclosures: Capturing Word Context Using LLM Methods. *Management Science* 71:9831–55.

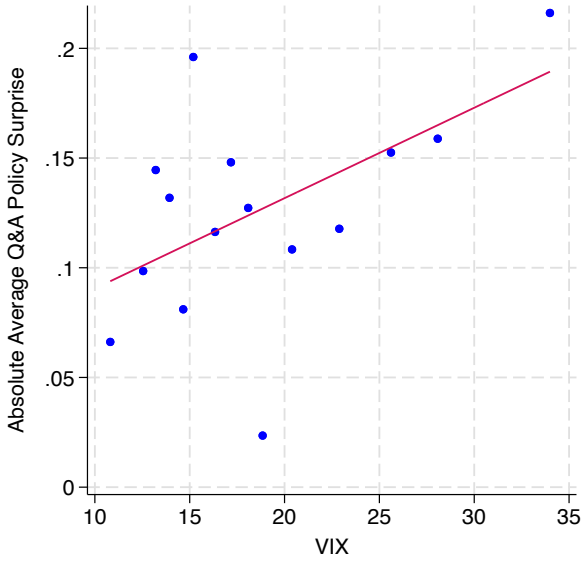
Van Nieuwerburgh, S., and L. Veldkamp. 2010. Information acquisition and under-diversification. *Review of Economic Studies* 77:779–805.



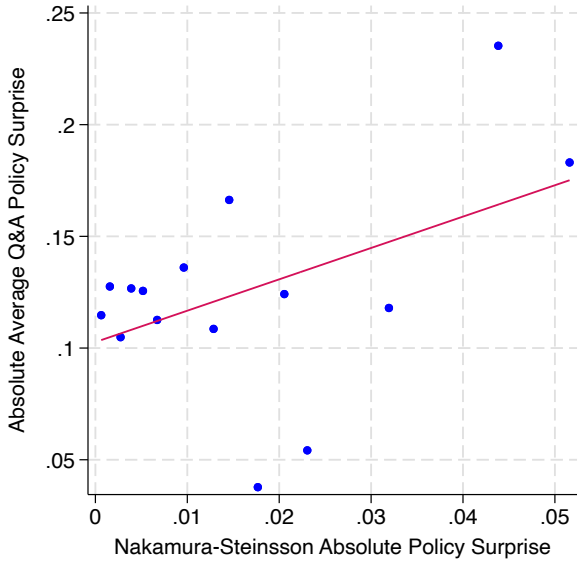
(a)



(b)

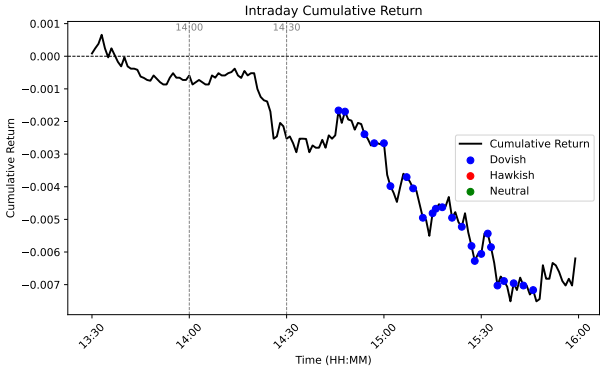


(c)

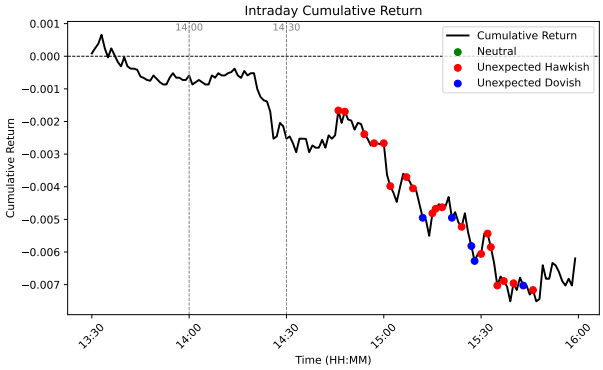


(d)

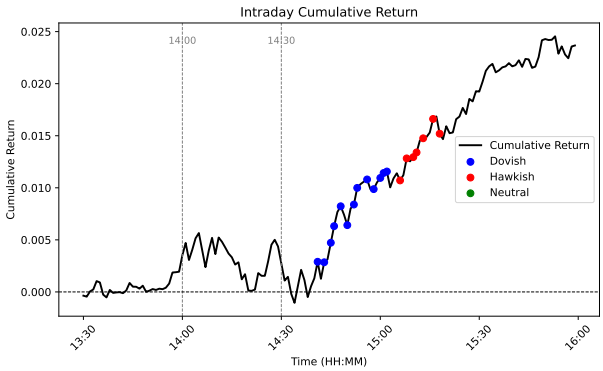
Figure 1: Panel (a) compares the policy stance extracted from actual FOMC press-conference answers with the stance extracted from simulated counterfactual answers generated by the model. The policy stance is computed from the LLM's five-category classification, mapped into numerical scores ranging from -1 (strongly dovish) to $+1$ (strongly hawkish). For each FOMC press conference, we compute the average stance score from the real Q&A and plot it on the horizontal axis, while the vertical axis reports the corresponding average stance score from the counterfactual answers. The fitted regression line summarizes the relationship between the two measures. Panel (b) plots the cross-sectional distribution of our LLM-based policy-surprise measures at the FOMC press-conference level. For each FOMC press conference, we compute the average stance score, which summarizes the policy surprise extracted from the Chair's responses during the question-and-answer session. The plot reports the distribution of press-conference policy surprises by Chair. Panel (c) and (d) show the relationship between the VIX index and the Nakamura-Steinsson Absolute Policy Surprise, respectively, and the absolute average policy surprise during FOMC press conference Q&A sessions, using binned scatterplots with 15 bins.



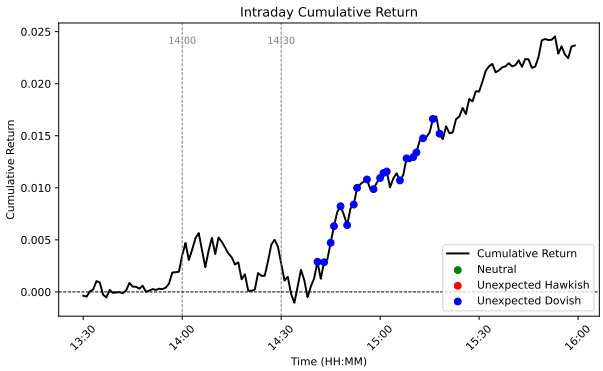
(a) Case 1: Policy stance of actual answers



(b) Case 1: LLM-based policy surprise

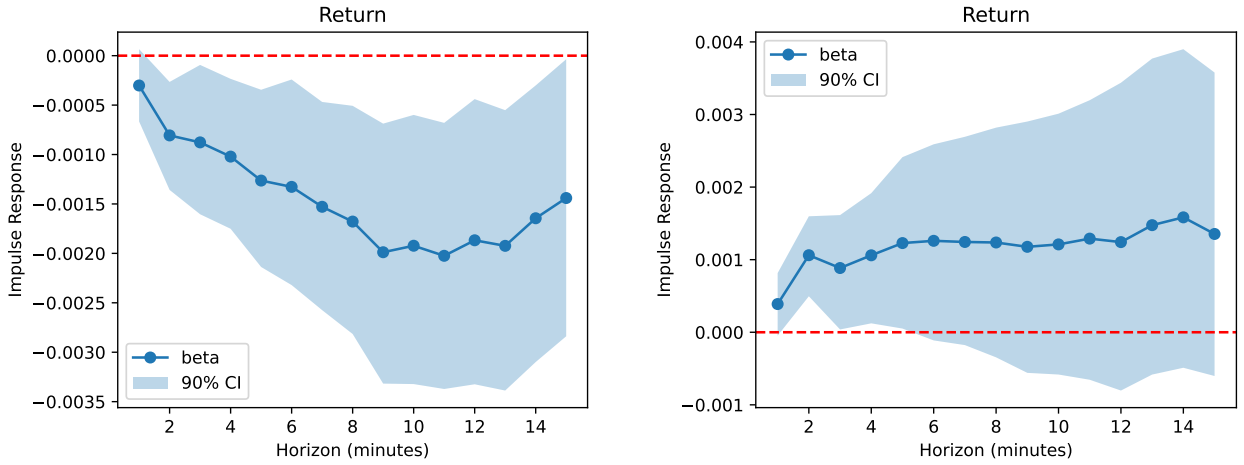


(c) Case 2: Policy stance of actual answers



(d) Case 2: LLM-based policy surprise

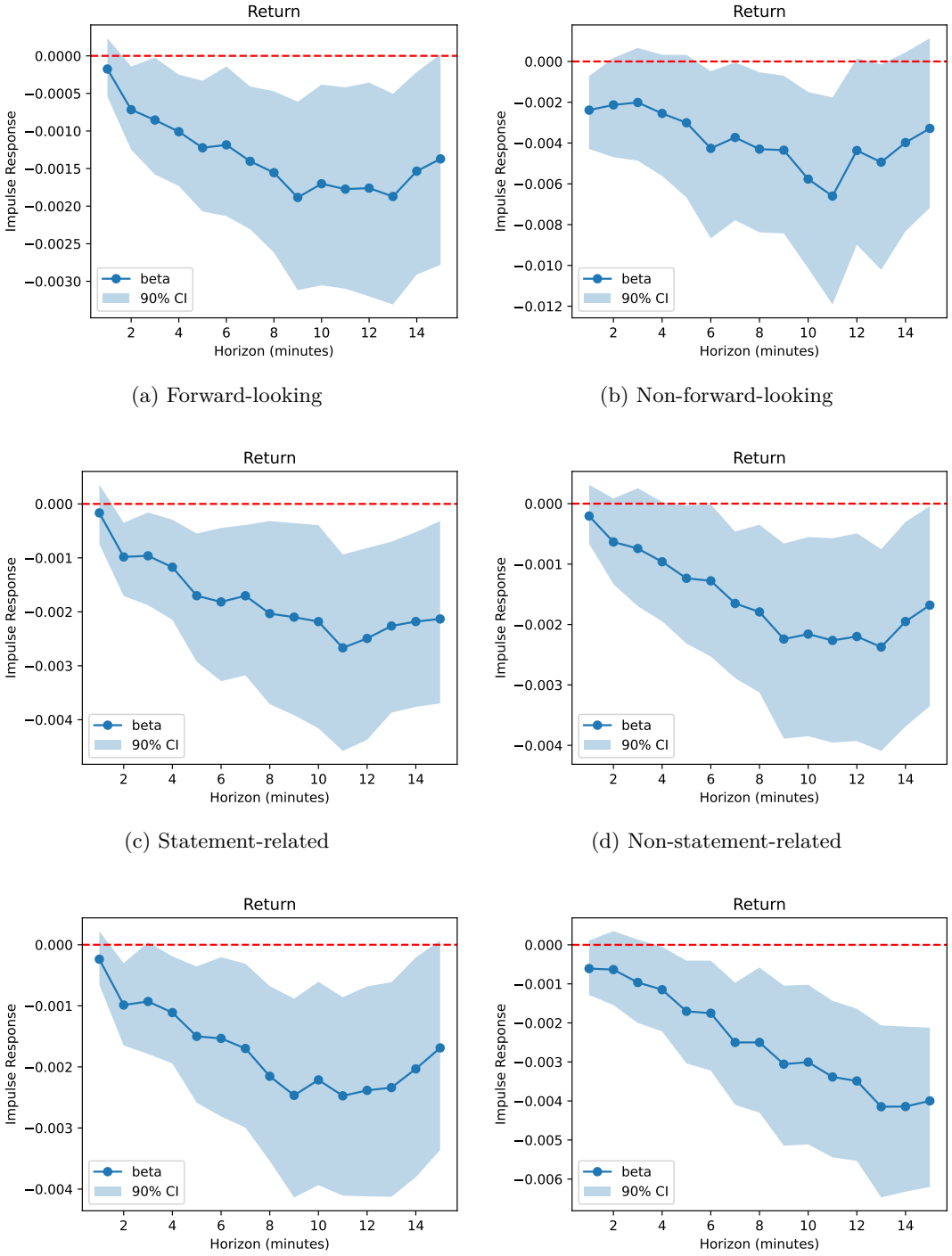
Figure 2: This figure presents the intraday cumulative return of the S&P 500 ETF (SPY) around two example FOMC press conferences. Case 1 corresponds to the December 12, 2012 press conference chaired by Bernanke, and Case 2 corresponds to the May 4, 2022 press conference chaired by Powell. Panels (a) and (c) color each response according to the cumulative policy stance implied by the Chair's actual responses (dovish, hawkish, or neutral). Panels (b) and (d) instead color each response according to the cumulative LLM-based policy surprise, defined as the difference between the policy stance implied by the Chair's actual response and the stance predicted from the real-time information set. Vertical dashed lines mark the release of the 14:00 FOMC statement and the beginning of the 14:30 press conference.



(a) PolicySurprise

(b) PolicyStance in Realized Answers

Figure 3: This figure presents the estimated impulse responses of SPY ETF returns to (a) the LLM-based policy surprise during FOMC press conferences and (b) the policy stance conveyed in the Chair's realized answers. The responses are estimated using a local projection equation (1) that controls for meeting fixed effects, question-rank fixed effects, lagged returns, lagged volatility, sentiment, policy stance, economic assessment, and statement-based policy surprises. Plotting the estimated coefficients $\hat{\beta}_1^{(h)}$ across horizons yields an impulse-response-style representation of how equity prices adjust to unexpected components of policy communication during FOMC press conferences. Shaded areas represent 90% confidence intervals based on bias-corrected and accelerated bootstrap standard errors clustered at the meeting level.



(a) Forward-looking

(b) Non-forward-looking

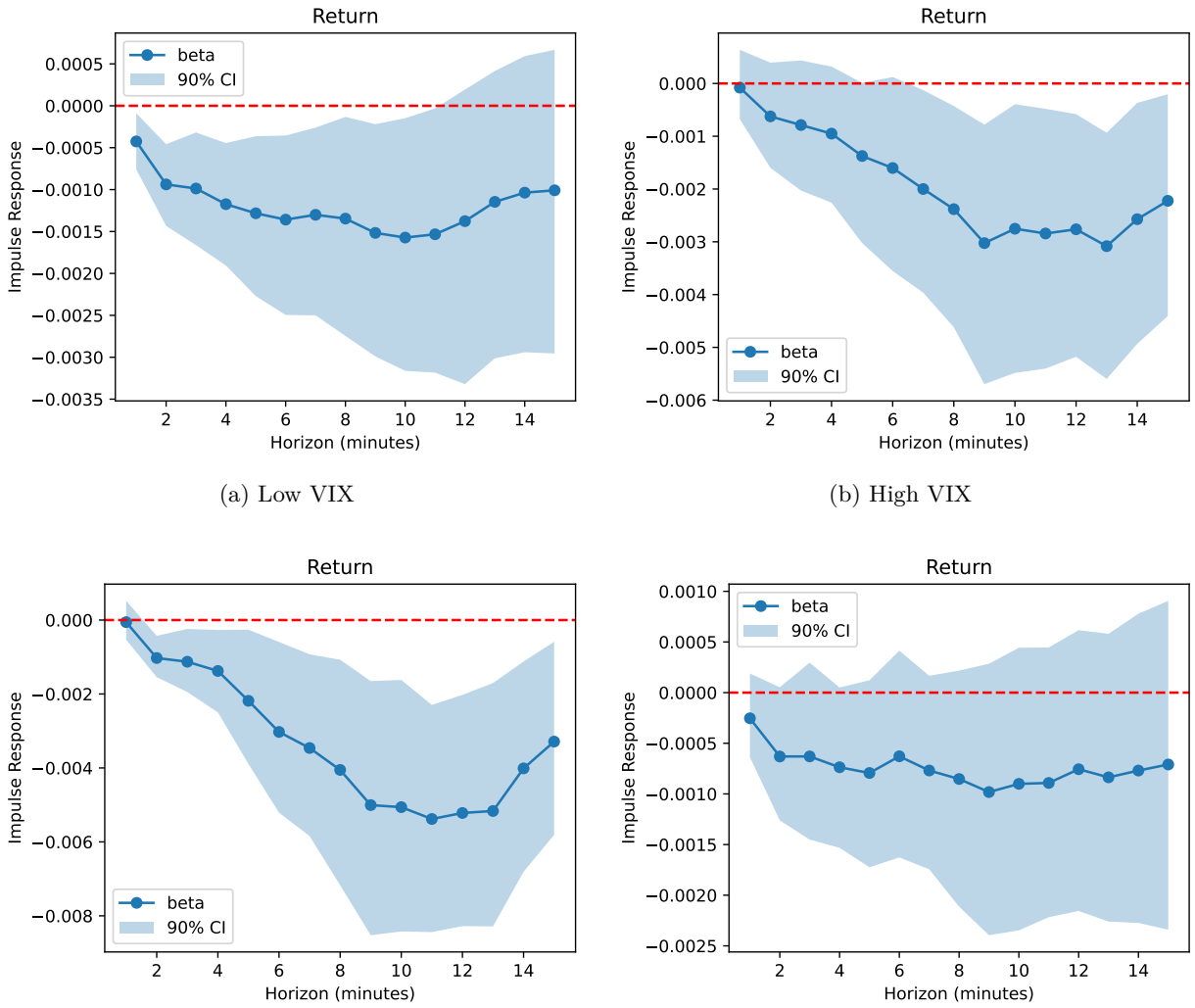
(c) Statement-related

(d) Non-statement-related

(e) Policy-related

(f) Economy-related

Figure 4: This figure presents local-projection estimates of the intraday response of SPY ETF returns to monetary policy surprises during FOMC press conferences, stratified by question characteristics. Panels (a) and (b) distinguish between forward-looking questions and questions that are not explicitly forward-looking. Panels (c) and (d) split the sample by whether the question directly relates to the content of the FOMC statement. Panels (e) and (f) split the sample by question topic, distinguishing between policy-related and economy-related questions. All specifications include the same set of control variables as the baseline specification in Equation (1), as well as meeting fixed effects and question-rank fixed effects.



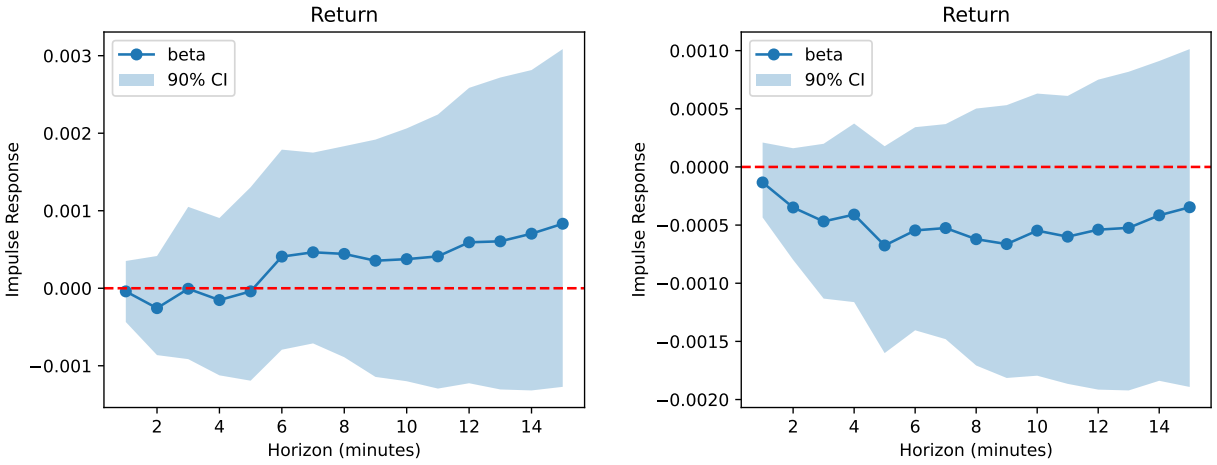
(a) Low VIX

(b) High VIX

(c) Dovish Statements

(d) Hawkish Statements

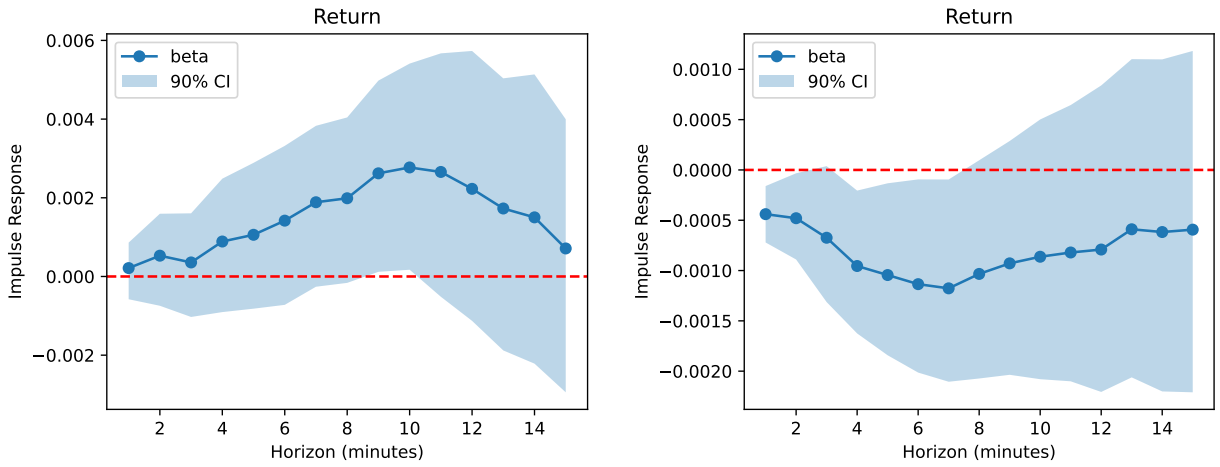
Figure 5: This figure demonstrates local-projection estimates of the intraday response of SPY ETF returns to monetary policy surprise, stratified by market conditions and statement tone. Panels (a) and (b) split the sample by market uncertainty, defined by whether the VIX index is below or above its sample median. We use the level of the VIX measured at 14:30 on the day of the FOMC meeting to classify market uncertainty, ensuring that the classification is predetermined with respect to information revealed during the press conference. Panels (c) and (d) split the sample by statement tone, classified as dovish or hawkish based on the policy shock measure of Nakamura and Steinsson (2018), with positive values indicating a hawkish stance. All specifications include the same control variables, as well as meeting and question-rank fixed effects.



(a) Without Macroeconomic Context

(b) Without Macroeconomic and Policy Context

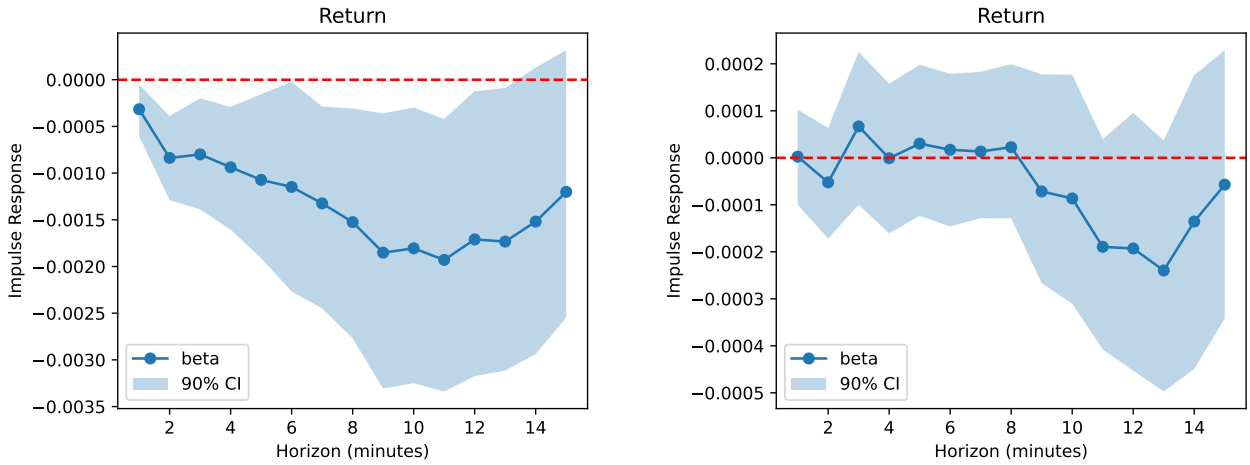
Figure 6: This figure presents local-projection estimates of the intraday response of SPY ETF returns to policy surprises during FOMC press conferences. Panel (a) reports estimates based on a policy surprise measure constructed from LLM-predicted answers that condition only on the press conference introduction and previously asked questions, omitting all background macroeconomic information. Panel (b) further restricts the prompt by excluding both macroeconomic context and policy-related information from the FOMC statement. In both panels, policy surprises are constructed using the time-weighted cumulative aggregation as in the main analysis. All specifications include the same set of control variables as in the main analysis, as well as meeting fixed effects and question-rank fixed effects.



(a) PolicySurprise (NW dictionary)

(b) PolicySurprise (ABH dictionary)

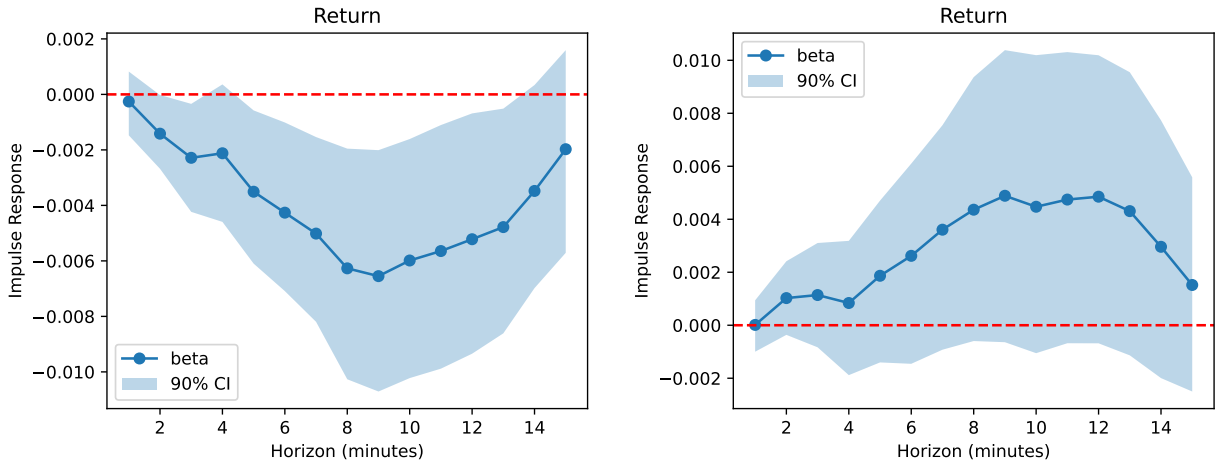
Figure 7: This figure presents local-projection estimates of the intraday responses of SPY ETF returns to press-conference policy surprises constructed using dictionary-based text measures. Following Gómez-Cram and Grotteria (2022), policy surprises are derived by quantifying the policy stance conveyed in the Chair's answers using textual dictionaries. Panel (a) reports results based on the fixed-phrase dovish–hawkish dictionary of Neuhierl and Weber (2019, hereafter the NW dictionary). Panel (b) instead uses a compositional term–modifier dictionary following Apel, Grimaldi, and Hull (2022, hereafter the ABH dictionary), which identifies policy stance through combinations of economically meaningful terms and directional modifiers. In both panels, policy surprises are constructed at the answer level and aggregated across the press conference using a time-weighted cumulative measure. All specifications include the same set of control variables as in the main analysis, as well as meeting fixed effects and question-rank fixed effects.



(a) Mean Cumulative Policy Surprise

(b) PolicySurprise Answer by Answer

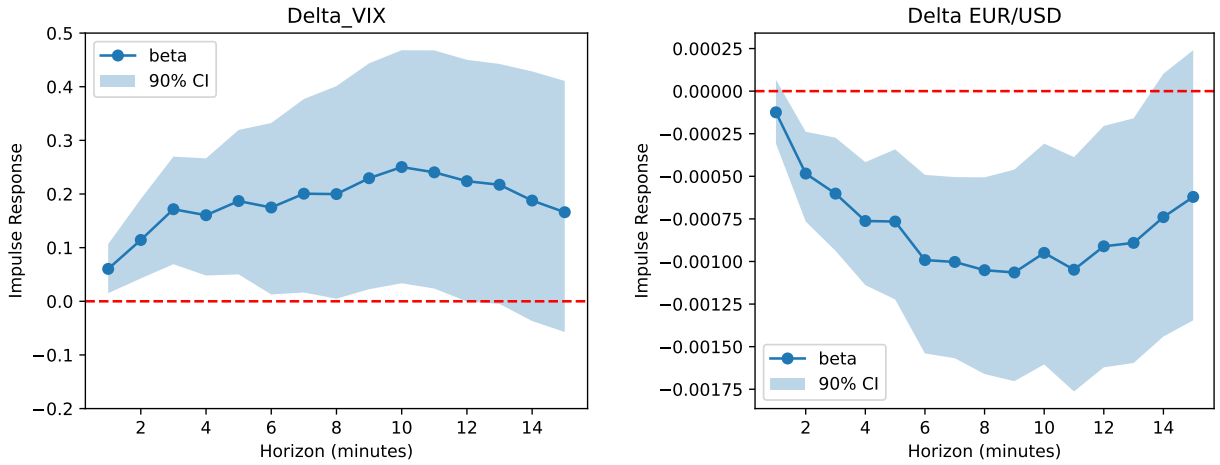
Figure 8: This figure presents local-projection estimates of the intraday response of SPY ETF returns to monetary policy communication in alternative specifications. Panel (a) reports local-projection estimates based on the mean cumulative LLM-based policy surprise aggregated across answers since the beginning of the press conference. Panel (b) reports estimates based on a flow specification that measures the policy stance conveyed in each real-time answer. Both specifications include the same control variables, as well as meeting fixed effects and question-rank fixed effects. Shaded areas represent 90% confidence intervals based on bias-corrected and accelerated bootstrap standard errors clustered at the meeting level.



(a) PolicySurprise

(b) PolicyStance in Realized Answers

Figure 9: This figure presents local-projection estimates of the intraday responses of SPY ETF returns to monetary policy communication during FOMC press conferences, using a sample restricted to meetings occurring after the knowledge cutoff of the GPT-4o-mini model (October 01, 2023). Panel (a) reports estimates based on the LLM-based policy surprise, while panel (b) reports estimates based on the policy stance conveyed in the Chair's realized answers. The specification includes the same control variables as in the main analysis, as well as meeting fixed effects and question-rank fixed effects. Shaded areas represent 90% confidence intervals based on bias-corrected and accelerated bootstrap standard errors clustered at the meeting level.



(a) ΔVIX

(b) ΔEUR/USD

Figure 10: This figure presents local-projection estimates of the intraday responses of financial markets to press-conference policy surprises using alternative dependent variables, following Curti and Kazinnik (2023). Panel (a) plots the response of market volatility, measured by the change in the VIX index, ΔVIX. Panel (b) plots the response of the exchange rate, measured by the change in the EUR/USD rate, ΔEUR/USD. In both panels, policy surprises are measured using the time-weighted cumulative policy surprise constructed from answer-level surprises. All specifications include the same set of control variables as in the main analysis, along with meeting fixed effects and question-rank fixed effects.

Table 1: Summary Statistics of FOMC Press Conferences Q&A Sessions

Panel A: Overall Statistics				
	All	Bernanke	Yellen	Powell
# Meetings (counts)	85.00	12.00	16.00	57.00
# Question per meeting	25.00	24.00	19.00	27.00
Time for each Question (seconds)	24.77	26.39	30.77	22.38
Time for each Answer	80.74	90.47	109.13	71.31
Panel B: Question Analysis				
# Statement-related question per meeting	4.36	6.18	4.07	4.37
# Non statement-related question per meeting	19.49	15.00	13.75	20.47
# Economic-related question per meeting	5.21	5.91	3.27	5.32
# Policy-related question per meeting	11.71	5.36	8.31	13.09
# Forward-looking question per meeting	15.41	11.55	10.63	16.61
# Critical tone question per meeting	11.48	12.00	9.75	11.07
# Neutral tone question per meeting	11.41	9.00	7.25	12.21

This table provides descriptive statistics for the FOMC press conferences in our sample. Panel A reports overall features of the Q&A sessions, including the number of meetings in the sample, the average number of questions and answers per meeting ("#" denotes counts), and the average duration of questions and answers, measured in seconds. These variables summarize the basic structure and intensity of interaction during each press conference. Panel B presents a breakdown of question types. To construct the question categories, we use GPT-based text classification. Each journalist question is processed through a standardized prompt that assigns it to thematic categories based on its semantic content. "Statement-related" and "non-statement-related" questions distinguish whether a question refers directly to the policy statement released at 14:00. "Economic-related" and "policy-related" questions capture inquiries about macroeconomic conditions and monetary policy decisions, respectively, while "forward-looking" questions focus on the future path of policy or the economic outlook. Finally, "critical tone" and "neutral tone" categories classify the sentiment of each question.

Table 2: Definition of Control Variables

Variable	**Definition**
Lagged Returns	Cumulative SPY ETF return over the 30-minute interval immediately preceding time t.
Lagged Volatility	Standard deviation of SPY ETF returns over the 30-minute interval immediately preceding time t.
Sentiment	Text-based sentiment score extracted from the Chair's answers at time t, constructed using the Loughran and Mcdonald (2011) dictionary. Sentiment is defined as the normalized difference between the number of positive and negative words, computed using only dictionary entries active in the corresponding year. The sentiment score is set to zero when no sentiment-classified words are present.
PolicyStance	LLM-based measure of the policy stance conveyed in the Chair's realized answers at time t. Each answer is classified into one of five ordered categories—strongly dovish, dovish, neutral, hawkish, and strongly hawkish—based on the implied direction of the future policy path, following Hansen and Kazinnik (2023). The categorical labels are mapped to numeric scores $\{-1, -0.5, 0, 0.5, 1\}$.
EcoAssessment	LLM-based measure of the Chair's assessment of economic conditions expressed in the realized answers at time t. Each answer is classified into one of five ordered categories reflecting the tone of economic conditions, namely positive, mildly positive, neutral, mildly negative, and negative, based on statements about growth, labor markets, inflation, and supply conditions. The categorical labels are mapped to numeric scores $\{1, 0.5, 0, -0.5, -1\}$.
D_t	The variable D_t captures statement-based policy surprises revealed during the press conference. We first identify surprises in the FOMC statement relative to the previous meeting (Gómez-Cram and Grotteria, 2022). For each statement, we compare the current text to the immediately preceding statement at the sentence level using fuzzy string matching. Sentences are classified as unchanged, changed, added, or removed based on their textual similarity. From sentences that are added, changed, or removed, we extract nouns and verbs to identify substantively modified policy-related content. These keywords are further expanded using a complementary dictionary of press-conference language, yielding a set of terms that summarize the dimensions along which the statement changes at time t. We then measure whether these changed terms are explicitly mentioned by the Chair in the realized press-conference answers.

Table 3: Regression Results: SPY ETF Returns and Policy Surprises

Panel A: Ret_{t+5}	(1)	(2)	(3)	(4)	(5)
PolicySurprise	-0.069*	-0.133**	-0.137**	-0.130**	-0.126**
	(-1.80)	(-2.49)	(-2.48)	(-2.48)	(-2.37)
PolicyStance		0.113	0.136*	0.134*	0.121*
		(1.58)	(1.87)	(1.87)	(1.68)
EcoAssessment			0.097*	0.091	0.094
			(1.65)	(1.49)	(1.53)
Sentiment				0.021	0.018
				(0.48)	(0.41)
D_t					0.013
					(1.04)
PastReturn/Vol	Yes	Yes	Yes	Yes	Yes
Conference FE	Yes	Yes	Yes	Yes	Yes
Question Rank FE	Yes	Yes	Yes	Yes	Yes
Adj. R^2	0.23	0.23	0.23	0.23	0.23
Panel B: Ret_{t+10}					
PolicySurprise	-0.129**	-0.185**	-0.188**	-0.188**	-0.187**
	(-2.15)	(-2.33)	(-2.34)	(-2.26)	(-2.23)
PolicyStance		0.099	0.116	0.116	0.113
		(0.94)	(1.08)	(1.08)	(1.06)
EcoAssessment			0.070	0.070	0.070
			(0.96)	(0.98)	(0.96)
Sentiment				0.001	0.000
				(0.01)	(0.00)
D_t					0.002
					(0.10)
PastReturn/Vol	Yes	Yes	Yes	Yes	Yes
Conference FE	Yes	Yes	Yes	Yes	Yes
Question Rank FE	Yes	Yes	Yes	Yes	Yes
Adj. R^2	0.40	0.40	0.40	0.40	0.40

This table presents regression results relating SPY ETF returns to the LLM-based policy surprise during FOMC press conferences.

$$\text{SPYReturn}_{t_a \to t_{a+5},m} = \beta_0 + \beta_1 \, \text{PolicySurprise}_{t_0 \to t_a,m} + \lambda_m + \gamma_a + X_t + \varepsilon_m.$$

The dependent variable $\text{SPYReturn}_{t_a \to t_{a+h},m}$ denotes the cumulative SPY return over the h-minute interval following answer a at time t. The interval h equals 5 minutes in Panel A and 10 minutes in Panel B. The key independent variable $\text{PolicySurprise}_{t_0 \to t_a,m}$ captures the unexpected component of monetary policy communication, defined as the difference between the policy stance conveyed in the Chair's realized answer and the stance predicted based on the real-time information set. Columns (1)–(5) report results from progressively augmented specifications that sequentially introduce control variables summarizing the realized content of the Chair's answers—namely policy stance, economic assessment, and sentiment—as well as statement-based policy surprises D_t. All specifications include past returns, past volatility, conference fixed effects, and question rank fixed effects as controls. To facilitate interpretation of coefficient magnitudes, SPY returns are multiplied by 100 and expressed in percentage points. The t-statistics based on standard errors clustered at the meeting level are reported in parentheses. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.