# Monetary Policy Shocks: A New Hope

Rubén Fernández-Fuertes*

July 30, 2025

**Abstract**

I develop a novel framework for computing monetary policy surprises by systematically processing the entire set of regular Federal Reserve communications using context-aware Large Language Models (LLMs). My approach analyzes the complete institutional communication cycle—Beige Books, FOMC Minutes, and policy Statements—as an integrated narrative system rather than isolated documents. The multi-agent architecture employs specialized LLMs to read and synthesize these full documents in sequence: extracting economic assessments from eight annual Beige Books, analyzing internal deliberations from Minutes, incorporating the full history of policy statements, and generating genuine surprises by comparing ex-ante expectations with realized decisions. Crucially, I form these expectations using only information available 2-3 weeks before each meeting, ensuring surprises reflect the unexpected evolution of Fed thinking during the blackout period. The resulting narrative surprises are as unpredictable as market-based measures (8-12% $R^2$ on standard predictors) yet explain 61.5% of policy rate changes compared to 15-17% for market surprises. This stark difference reveals that most monetary policy "news" stems from how Fed thinking evolves between its documented communications and final decisions—evolution that cannot be predicted ex-ante but dramatically moves policy when revealed. By processing the Fed's complete regular communication apparatus as an interconnected system, the framework demonstrates how LLMs can measure the true information content of central bank transparency, distinguishing predictable policy rules from genuine monetary shocks.

## 1 Introduction

Why does the Federal Reserve change policy when it does? Market-based measures of monetary policy surprises—captured through high-frequency movements in interest rate

*Bocconi University, Milan, Italy. I thank Max Croce, Carlo A. Favero, Mohammad R. Jahan-Parvar, and Isabella M. Wolfskeil, David Murakami, Ivan Shchapov, Martin Fankhauser, Alejandra Inzunza, Fernando Pérez-Cruz, Angelo Ranaldo, Fiorella de Fiore, Damiano Sandri for their valuable comments. Email: ruben.fernandez@phd.unibocconi.it

futures—consistently explain only 15-17% of actual policy decisions (Jarociński & Karadi, 2020). This leaves over 80% of policy variation unexplained, painting monetary policy as either largely arbitrary or driven by information revealed only at the last moment. Yet this picture conflicts with the Federal Reserve's systematic approach to communication and its well-documented efforts at transparency since it began announcing policy decisions in 1994. Is monetary policy truly this unpredictable, or are we measuring the wrong component of surprise?

The identification of monetary policy shocks has been central to understanding the non-neutrality of monetary policy (Gertler & Karadi, 2015; Nakamura & Steinsson, 2018). The modern approach relies on high-frequency identification: measuring changes in interest rate futures in narrow windows around FOMC announcements to isolate the unexpected component of policy decisions. This methodology rests on two critical assumptions. First, the measured changes must be genuinely unexpected—if markets can predict these "surprises", they cannot represent true shocks. Second, the price movements must be caused exclusively by the monetary policy announcement, not by other information revealed simultaneously.

Recent research has raised concerns about both assumptions. Bauer and Swanson (2023b, 2023c) demonstrate that high-frequency surprises are significantly predictable using information available before FOMC meetings, violating the first assumption. Regarding the second, several studies have shown that central bank announcements convey information beyond policy intentions, creating "information effects" that contaminate the identification (Jarociński & Karadi, 2020; Miranda-Agrippino & Ricco, 2021). These findings have motivated various orthogonalization procedures and prompted a reconsideration of what monetary policy surprises actually measure.

An alternative tradition uses narrative approaches to identify policy shocks. Romer and Romer (2004) pioneered this method by extracting intended policy changes from internal Federal Reserve documents and purging them of systematic responses to economic forecasts. However, their approach faces two limitations in the modern era. First, after 1994, when the Fed began announcing decisions immediately, the painstaking reconstruc-

tion of intended changes became largely redundant. Second, their reliance on internal documents (Greenbooks/Tealbooks) means the data only becomes available with a five-year lag, limiting real-time applicability.

I argue that the answer lies in recognizing that monetary policy decisions reflect two distinct processes: a *deliberative policy evolution* that unfolds through weeks of Federal Reserve communications, and a *high-frequency shock* captured by market reactions in narrow windows around announcements. By developing a novel framework to extract and quantify the deliberative component from the Fed's own narrative, I show that over 60% of policy decisions are already embedded in the Fed's public communications two to three weeks before each meeting—during the critical "blackout period" when Fed officials cease public commentary.

This decomposition reveals a fundamental insight: what markets perceive as "surprises" on announcement day largely reflect the resolution of uncertainty about a policy path that was already heavily signaled through the Fed's formal communication channels. The true unexpected component—the genuine policy shock—is much smaller than traditional measures suggest. This finding contrasts sharply with recent work showing limited predictability in market-based surprises (Bauer & Swanson, 2023b), suggesting that the key to understanding monetary policy lies not in high-frequency market reactions but in the Fed's deliberative communications. The result not only reconciles the apparent predictability paradox but also demonstrates that monetary policy is far more systematic and transparent than previously understood.

To operationalize this decomposition, I develop a framework that systematically analyzes the Federal Reserve's own narrative evolution leading up to each FOMC meeting. Building on the narrative tradition pioneered by Romer and Romer (2004), but using publicly available documents rather than internal Fed materials, I form expectations based exclusively on Fed communications available two to three weeks before the meeting. This timing coincides with the "blackout period" when Committee members cease public commentary—a deliberative period when the Fed finalizes its policy stance internally. This temporal separation ensures that my narrative-based expectations are predetermined rel-

ative to any last-minute information that might influence the actual decision, providing clean identification of the deliberative versus shock components.

Specifically, I construct a multi-agent system using Large Language Models (LLMs) to process and synthesize information from four key Fed documents released at different stages of the policy cycle. The *Beige Book*, published approximately two weeks before each meeting, provides qualitative assessments of regional economic conditions. The *Minutes* from the previous meeting, released three weeks after that decision, reveal the Committee's internal deliberations, forward guidance intentions, and the balance of views among members. Historical *FOMC Statements* capture the evolution of the policy narrative and any path dependence in decision-making. Together, these documents contain the information set available to an attentive observer trying to anticipate the Fed's next move.

To implement this framework, I develop a multi-agent system using Large Language Models (LLMs) that processes the Federal Reserve's own narrative documents. The innovation lies not in the use of LLMs per se—which have been applied to various financial tasks (Gambacorta et al., 2024; A. L. Hansen & Kazinnik, 2023; Pfeifer & Marohl, 2023)—but in orchestrating multiple specialized agents to capture the temporal evolution of Fed communications. While previous work has used text analysis to characterize FOMC communication (Ahrens et al., 2024; Cieslak et al., 2024b; McMahon et al., 2019), these approaches treat documents in isolation rather than as part of an evolving narrative. My multi-agent architecture addresses this limitation by processing information sequentially, mirroring how the Fed's own thinking develops across its communication cycle.

The system comprises four specialized agents, each designed to extract specific information from Fed documents:

1. **Beige Book Calibrator (BBC)**: Processes the Beige Book released two weeks before each meeting to quantify economic conditions across the Fed's dual mandate variables. Recent research confirms the Beige Book contains substantial forward-looking information not captured in standard quantitative data (Aruoba & Drech-

sel, 2024; Balke et al., 2017; Filippou et al., 2024a).

2. **Policy Extractor (PE)**: Analyzes Minutes from the previous meeting to extract the Committee's internal deliberations, forward guidance intentions, and the distribution of views. This captures what I call the "policy stance distribution"—the range of preferences within the Committee that shapes future decisions.
3. **Expectation Engine (EE)**: Synthesizes outputs from BBC and PE along with historical statements to generate a prior probability distribution over the upcoming decision. This agent explicitly models policy inertia and path dependence (Bernanke & Mihov, 1998; Woodford, 1999).
4. **Surprise Snipper (SS)**: Compares the prior distribution with the actual FOMC statement to compute the narrative surprise, decomposing it into predictable and unpredictable components.

This architecture enables systematic measurement of how Fed communications evolve from the Beige Book release through the blackout period to the final decision. By stopping information collection at the blackout period's start, I ensure clean identification—my narrative surprises are predetermined relative to any last-minute data or market movements that might influence the actual decision.

My results reveal three key findings that reshape our understanding of monetary policy surprises: First, the Beige Book scores contain substantial predictive power for policy decisions. Employment and economic growth conditions jointly explain 14% of rate changes—a striking result given that policy inertia alone explains virtually nothing ($R^2 = 0.004$). This suggests the Fed responds systematically to the qualitative economic assessments in its own regional reports, validating their informational content beyond standard macroeconomic indicators. Second, narrative surprises are as unpredictable as market-based measures. Rolling window analysis shows my narrative surprises achieve mean $R^2$ values of 8-12% when regressed on standard predictors, comparable to or better than high-frequency measures (9-17%). This is particularly notable given that narrative surprises use information available 2-3 weeks before meetings, while market measures incorporate data up to the announcement moment. Third, and most strikingly, regres-

sions of policy changes on narrative surprises yield coefficients near unity and $R^2$ values exceeding 61.5%, compared to 15-17% for market-based measures. Rather than claiming superiority, I interpret this as revealing a fundamental decomposition between the deliberative and shock components of monetary policy.

This decomposition shows that most monetary policy variation stems from the Fed's inter-meeting communication evolution rather than announcement-day shocks. The "surprise" that moves markets represents only a small residual after accounting for the policy path already signaled through Fed documents. This insight reconciles the predictability paradox: monetary policy appears unpredictable when viewed through narrow market windows but is largely systematic when analyzed through the Fed's own communication cycle.

These findings have important implications for both research and practice. For researchers, the decomposition offers complementary identification strategies: market measures remain ideal for isolating pure exogenous shocks, while narrative measures illuminate the systematic component of policy decisions. For market participants, tracking the Fed's communication evolution during the blackout period may provide valuable signals about likely policy outcomes. For policymakers, the results validate the effectiveness of Fed communication—over 60% of decisions are successfully telegraphed through formal channels weeks in advance.

The remainder of the paper proceeds as follows. Section 2 details the multi-agent system architecture and how each agent processes Fed documents to construct narrative surprises. Section 3 presents the empirical evidence, demonstrating the predictive power of Beige Book scores, the unpredictability of narrative surprises, and the decomposition between deliberative and shock components. Section 4 discusses implications for monetary economics and the broader use of LLMs in economic research.

# 2 Methodology

## 2.1 General Framework

The narrative methodology that I am proposing is based on the systematic structure of the FOMC's communication. For each of the eight *scheduled* meetings, the FOMC releases a sequence of documents that collectively offer a comprehensive insight into its deliberations and decisions. This sequence typically includes: (1) The *Beige Book*, formally known as the "Summary of Commentary on Current Economic Conditions by Federal Reserve District," is released approximately two weeks prior to each FOMC meeting. It compiles qualitative, anecdotal information about current economic conditions from the twelve Federal Reserve Districts, gathered by their respective banks. This document provides a ground-level perspective on economic activity, often highlighting emerging trends and regional divergences that may not be immediately apparent from quantitative data; (2) The *FOMC Statement* is released immediately following the conclusion of each scheduled FOMC meeting (2:00 p.m. ET). This concise document announces the committee's decisions regarding the federal funds rate target and other monetary policy tools; (3) The *Minutes* of the FOMC meeting are published three weeks after the policy decision. These provide a more detailed summary of the internal discussions among committee members, including their perspectives on economic conditions, risks to the outlook, and the various policy options considered. The Minutes often reveal the nuances of dissenting opinions, the range of views on the appropriate path of monetary policy, and the underlying assumptions guiding the committee's collective judgment, offering a deeper understanding of the causal mechanisms driving policy formulation. I report a schematic of the FOMC's communication releases in Figure 1.

A researcher or a market participant that is *attentive* to the FOMC's decision of raising or lowering the federal funds rate to condition their analysis and decisions would therefore have to process this information, either directly from the reported documents or indirectly through the media. *Large Language Models* (LLMs) are a promising tool to help with this task. They are able to process and summarize the information in the

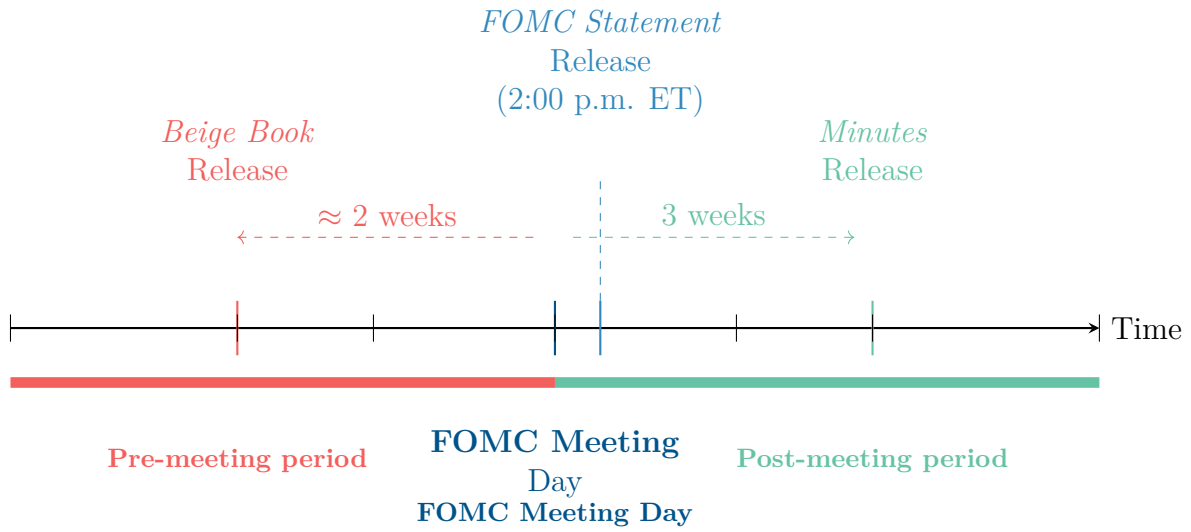



Figure 1: Timeline of FOMC communication releases relative to meeting day

documents and quantify the information in the documents (from words to numbers). One could say, though, that the use of LLMs could be overkill for this task. Indeed, there have been multiple attempts to extrat quantitative information from central banks texts: either with text analysis techniques (Ahrens & McMahon, 2021; Ahrens et al., 2024) or with more advanced and recent techniques of natural language processing (Aruoba & Drechsel, 2024) or even with the use of LLMs (De Fiore et al., 2024; Gambacorta et al., 2024; A. L. Hansen & Kazinnik, 2023). However, none of them have yet explore the full capabilities of Multi-Agent Systems (MAS) in the context of central bank communication. The literature on Large Language Models is growing exponentially, and the use of LLMs in the context of central bank communication is still in its infancy, even though the application to this task is a natural fit.

MAS are *teams* of LLMs that collaborate to solve a given task. The analogy is straightforward: instead of a single researcher or market participant, we have a team that can process information in parallel and synthesise it. One analyst might read the Beige Book, another the FOMC Statement, and another the Minutes. The team then consolidates this information into a report. This is the approach I propose in this paper. The literature on improving performance with LLMs in collaborative settings is expanding (Feng et al., 2025; Talebirad & Nadiri, 2023; Yang et al., 2025). The rationale is intuitive. First, a single LLM has a finite context window, limiting the information it can process

and remember at once. For long, complex tasks requiring extensive memory or iterative refinement, a monolithic approach becomes impractical. Second, as the complexity and number of instructions in a single prompt increase, LLMs can become "confused" or less reliable in their output, similar to a human juggling too many mental processes simultaneously. Third, a single LLM, despite its general "intelligence", may lack the deep, domain-specific expertise needed for certain sub-tasks, leading to "hallucinations"—believable but factually incorrect information. A single model might also struggle to adapt to dynamic environments or novel problem constraints without explicit re-training or extensive prompt engineering. Finally, processing all aspects of a complex task with a single, often large, LLM can be computationally expensive and slow. Scaling such a system for higher throughput is challenging, as it often involves replicating the entire large model. Moreover, LLMs have been shown to perform better on small tasks rather than complex, specialised, long tasks, as the latter requires extensive domain knowledge while the former only requires general problem-solving strategies (Wu et al., 2024). The greatest limitation of the information set I plan to process with LLMs is the vast context length, making it impossible to process in one shot with existing tools. Before LLMs became widespread, researchers devised methods to process documents by splitting them into sentences and using dictionary-based methods to extract information (Ahrens & McMahon, 2021; Ahrens et al., 2024; Cieslak et al., 2024a). However, these methods lack the ability to be *attentive* to relevant parts of a text. LLMs act as universal density approximators of the marginal distribution of languages, meaning sufficiently large and well-trained LLMs can accurately assess this distribution (Jiang, 2023). Leveraging these properties in a multi-agent setting lets me break the context-window bottleneck and exploit their *emergent*[1] abilities. Long-range dependency tracking, in-context learning, and chain-of-thought reasoning—to produce richer, more reliable macro-financial insights than classical dictionary methods or a single monolithic model can deliver (Du et al., 2024; Wei, Tay, et al., 2022).

[1]The term 'emergent' is used here to describe abilities that appear to arise spontaneously as a result of the model's complexity and training, though there is ongoing debate about whether these abilities are truly emergent or simply artifacts of the evaluation metrics used. This philosophical discussion is beyond the scope of this paper.

I propose a methodology with four agents consistent with the timeline of the FOMC's communication releases described in Figure 1. The agents are: (1) Beige Book Calibrator (BBC), responsible for calibrating the Beige Book. It is a LLM that is able to process the Beige Book and extract the quantitative information from it. (2) Policy Extractor (PE), responsible for extracting the *status quo* of the FOMC's views on the economy and the future path of the economy from the Minutes. (3) Expectation Engine (EE), responsible for processing the BBC's output, the past FOMC Statements and the PE's output to produce a prior distribution for the monetary policy decision. (4) Surprise Snipper (SS), responsible for extracting the *surprise* from the FOMC Statement given the EE's output.

The architecture of this Multi-Agent System is illustrated in Figure 2, which shows the sequential processing flow and data dependencies between agents.

## 2.2 Beige Book Calibrator

The Beige Book Calibrator (BBC) is the first specialised agent in the proposed MAS. It is designed to systematically process and quantify the anecdotal economic information contained in the Federal Reserve's Beige Book. The agent's primary function is to transform the qualitative narrative of each Beige Book into a structured, quantitative assessment of economic conditions.

For each document, the BBC employs a Large Language Model (LLM) guided by a detailed prompt. This prompt instructs the model to act as an expert economic analyst, meticulously parsing the text to identify sentences related to four key macroeconomic variables: inflation, employment, economic growth, and consumer spending. For every relevant sentence identified, the LLM extracts the verbatim text and assigns it a multi-dimensional analysis, including: the associated `variable`, its implied `policy_stance` (categorised as hawkish, dovish, or neutral), an `intensity` score from 0 to 1, and the model's `confidence_level` in its own assessment.

From this granular, sentence-level JSON[2] output, the agent aggregates the information to compute two key summary statistics. First, it calculates a set of quantitative

[2]Other formats are possible, but JSON is the most common and easiest to parse.

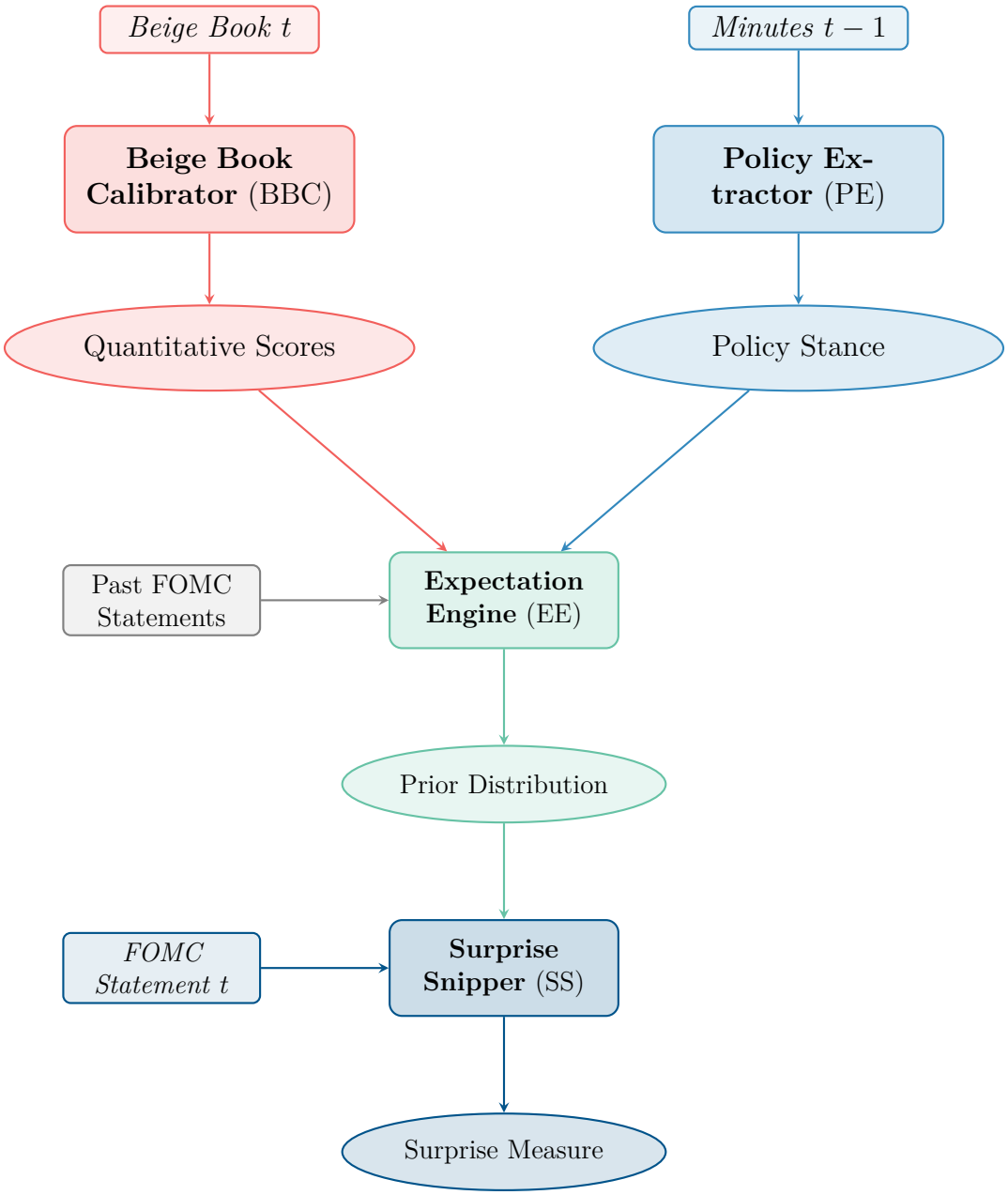



Figure 2: Multi-Agent System architecture for FOMC communication analysis. The four agents process documents sequentially in a top-down flow, synthesizing information from various documents to generate expectations and surprises.

`scores`, one for each of the four variables, on a scale from -1 (indicating weak, dovish-leaning conditions) to +1 (indicating strong, hawkish-leaning conditions). Second, it determines a set of `weights` that sum to 1.0, reflecting the relative emphasis or frequency of discussion each variable receives within the document. This ensures that the final output captures not only the direction of economic signals but also their perceived importance in the report.

The weights are derived unsupervised through the LLM's assessment of relative emphasis within each document. The model is instructed to evaluate the importance given to each economic variable based on the frequency of discussion, level of detail provided, and positioning within the document structure. Crucially, the weights are generated contemporaneously with the scores, ensuring that both the quantification and the relative importance reflect the Federal Reserve's communication emphasis at that specific meeting. This approach ensures the weighted aggregate score captures the Fed's actual communication priorities rather than imposing external assumptions about variable importance.

A critical feature of the BBC is its robust handling of large documents that exceed the standard context window of LLMs. When a Beige Book is too long, the agent automatically activates a chunking mechanism. It intelligently divides the document into smaller, overlapping segments, preserving logical breaks like sections and paragraphs where possible. Each chunk is processed independently by the LLM, and the resulting sentence-level (but context-aware) analyses are then carefully merged into a single, comprehensive dataset. This approach ensures that the entire document is analyzed without loss of granularity. To ensure efficiency and reproducibility, a caching layer is implemented, storing the detailed JSON analysis for each document to avoid redundant processing.

Figure 3 illustrates the time series of the BBC's output, showing both the individual variable scores (inflation, employment, economic growth, and consumer spending) as dashed lines and the weighted aggregate score as a solid line. The aggregate score reflects the overall economic sentiment captured by the Beige Book, weighted by the relative emphasis each variable receives in the document. This visualization demonstrates the

BBC's ability to systematically quantify the qualitative economic narratives, providing a measure of economic conditions consistent with the business cycle over time. This suggests evidence of its usefulness as input for subsequent agents in the pipeline. To

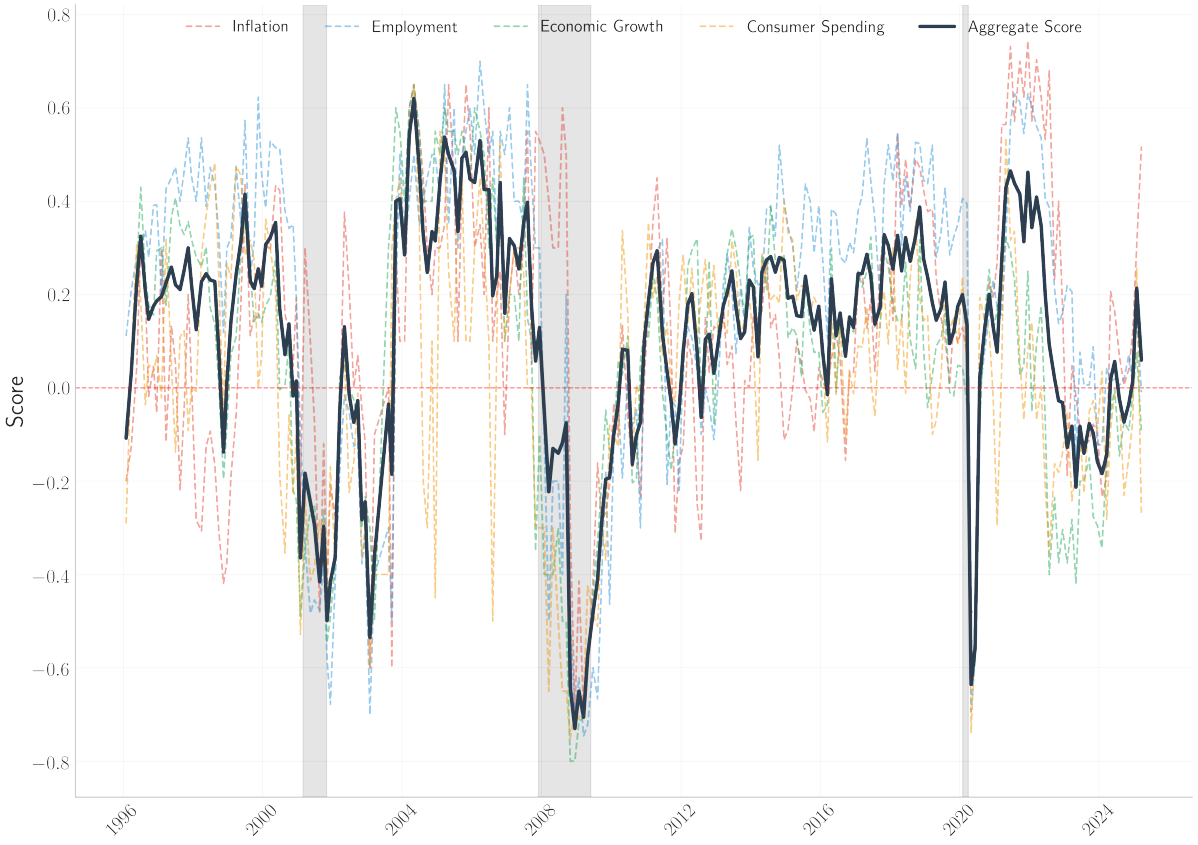



Figure 3: Time series of Beige Book Calibrator output showing individual variable scores (dashed lines) and weighted aggregate score (solid line). The four variables tracked are inflation, employment, economic growth, and consumer spending. Scores range from -1 (dovish/weak conditions) to +1 (hawkish/strong conditions). The aggregate score is calculated using weights that reflect the relative emphasis each variable receives in the document.

further assess the statistical properties of the BBC's output, I perform a descriptive analysis of the resulting scores and weights. Figure 4 displays the empirical distributions of the four variable scores. The distributions are unimodal and centered near zero, but exhibit mild skewness and some deviations from normality, particularly in the tails. This suggests that while the BBC produces a range of positive and negative assessments, the overall sentiment is balanced across the sample, with occasional periods of more extreme views.

Figure 5 reports the correlation matrix for the variable scores and the aggregate in-

dex. All variables are positively correlated, both with each other and with the aggregate, reflecting the fact that macroeconomic conditions often move together and that the aggregate index effectively summarizes the joint information in the underlying components.

Figure 6 visualizes the time-varying weights assigned to each macroeconomic theme in the Beige Book, highlighting the substantial and persistent shifts in emphasis across different periods. The stacked area chart makes clear that these weights are highly dynamic, with the relative importance of each variable evolving in response to changing economic conditions. Most strikingly, the recent period—beginning around 2021—shows a dramatic surge in the weight placed on 'inflation' (cyan), which rapidly becomes the overwhelmingly dominant theme in the Beige Book narrative. This sharp increase coincides with the onset of the post-pandemic inflationary episode, during which inflation concerns eclipse all other topics and account for the largest share of the document's focus. In contrast, the weights on 'employment' (magenta) and 'economic growth' (green) fluctuate over time, with employment gaining prominence during periods of economic stress, while 'consumer spending' (yellow) remains relatively stable but still exhibits meaningful variation. The pronounced shift toward inflation in the most recent years underscores the BBC's ability to capture not only the direction but also the evolving salience of macroeconomic themes, faithfully reflecting the Federal Reserve's shifting priorities in response to the changing economic landscape.

To validate the economic content of the BBC's output, I examine the relationship between the aggregate Beige Book score and key macroeconomic variables. I use monthly macroeconomic data. Hence, I resample the Beige Book scores to monthly frequency to align with the temporal structure of macroeconomic time series[3]. Figure 7 presents scatter plots of the aggregate score against four key indicators: unemployment rate, GDP growth, CPI inflation, and PCE growth. The correlations reveal economically sensible relationships: the aggregate score is negatively correlated with unemployment ($\rho = -0.363$), which makes sense, as the content extracted from the document measures sentiment with respect to 'employment' rather than 'unemployment'. This aligns with

[3]Add details about the resampling method,

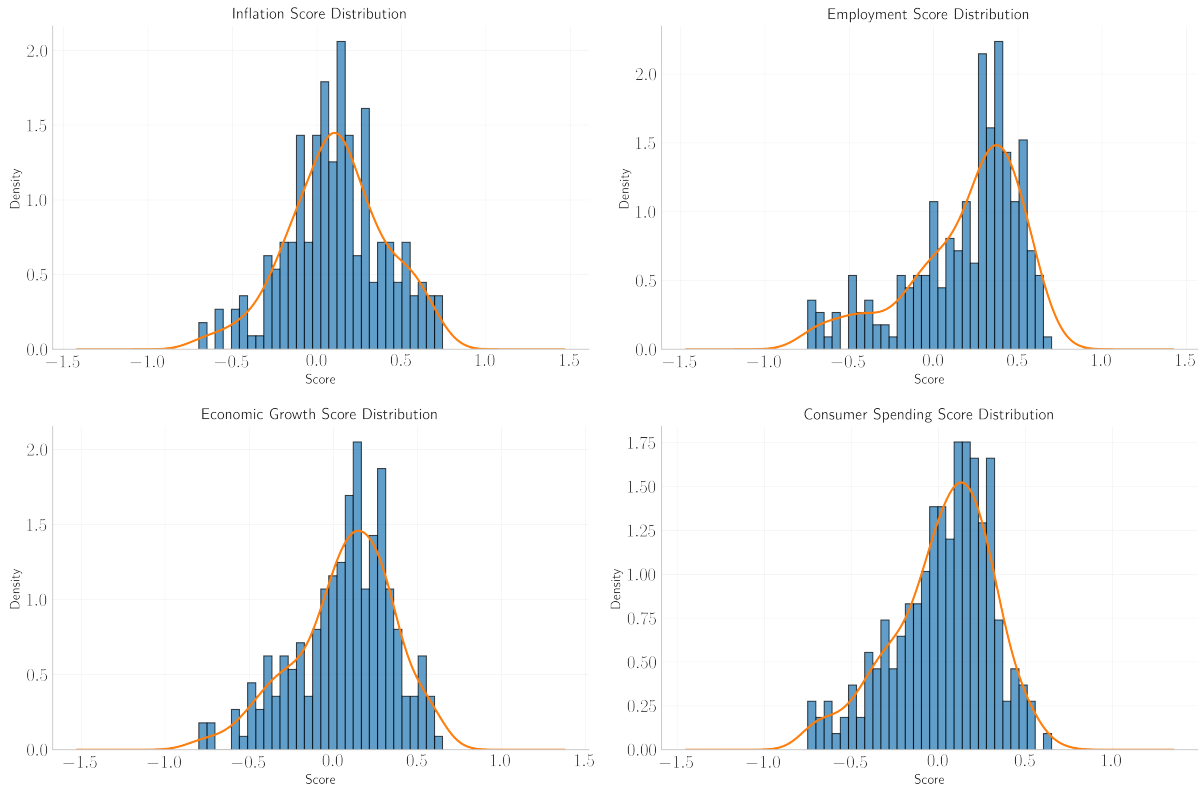



Figure 4: Distribution of Beige Book variable scores. Each panel shows the histogram and kernel density estimate for one of the four key variables, providing insight into the statistical properties of the BBC's output.

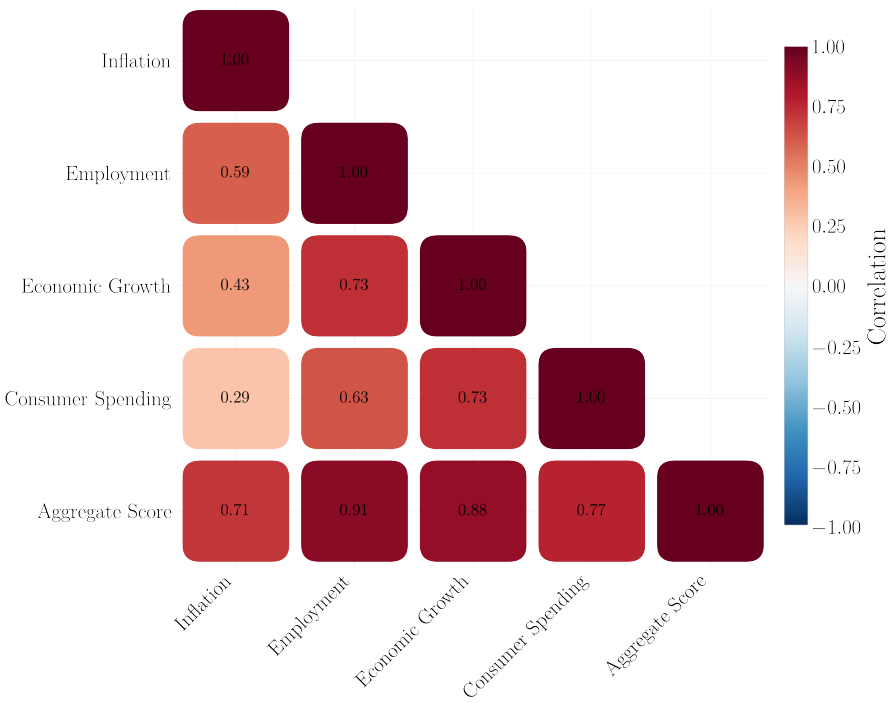



Figure 5: Correlation matrix of Beige Book scores. The heatmap shows the Pearson correlation coefficients between the four variable scores and the aggregate index, confirming their positive inter-relationships.

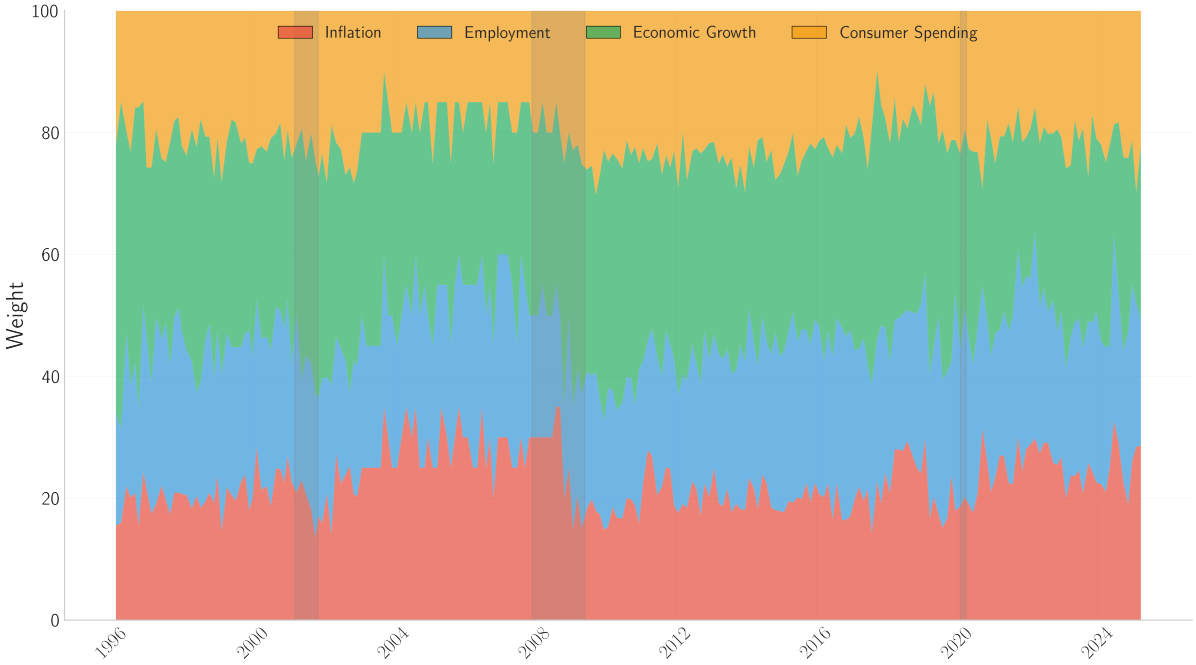



Figure 6: Time-varying weights of Beige Book components. The stacked area chart illustrates how the relative importance of each economic variable has evolved over time, with recession periods shaded in gray.

the expectation that positive economic sentiment typically coincides with lower unemployment. Conversely, it shows positive correlations with GDP growth ($\rho = 0.604$), CPI inflation ($\rho = 0.217$), and PCE growth ($\rho = 0.196$), indicating that the BBC successfully captures the underlying economic conditions reflected in these conventional macroeconomic indicators.

These validation exercises demonstrate that the BBC successfully transforms qualitative Beige Book narratives into quantitative measures that capture economically meaningful relationships with both business cycle indicators and the relative emphasis the Fed places on different economic themes.

## 2.3 Policy Extractor

The Policy Extractor (PE) constitutes the second stage of the pipeline and operates *exclusively on the publicly released* Minutes that the Federal Reserve publishes exactly three weeks after every scheduled meeting. Although shorter than the verbatim transcripts—which remain under seal for five years—the Minutes provide the most timely le-


(a) Unemployment Rate ($\rho = -0.363$)

(b) GDP Growth ($\rho = 0.604$)

(c) CPI Inflation ($\rho = 0.217$)

(d) PCE Growth ($\rho = 0.196$)

Figure 7: Macro linkage analysis of Beige Book aggregate scores. Each panel shows the relationship between the monthly-resampled aggregate score and key macroeconomic variables. The correlations ($\rho$) demonstrate that the BBC successfully captures economically meaningful relationships, with negative correlation for unemployment and positive correlations for growth and inflation measures.

gal window into the Committee's closed-door deliberations. By analysing this document, the PE extracts the *policy intelligence*—internal debates, forward-looking guidance, and risk assessments—that underpin the headline rate decision but are absent from the same-day Statement. Guided by a bespoke prompt, the agent outputs a richly structured JSON object organised into six analytical blocks:

1. **Internal committee dynamics.** Identification of voting patterns, dissenting voices, and the compromise reasoning that underpins the agreed decision, thereby mapping the hawk–dove spectrum within the Committee.
2. **Enhanced policy stance.** Comparison between the concise public Statement tone and the more detailed deliberations revealed in the Minutes, detection of concerns and nuances not captured in the headline decision, and identification of staff–versus–Committee divergences shaped the discussion.
3. **Forward-guidance signals.** Extraction of explicit and implicit references to future actions together with the economic thresholds that would trigger a pause, a hike, or a cut, in line with best practices in the forward-guidance literature.
4. **Economic assessment revision.** Evaluation of how the Committee's narrative updates the Beige Book outlook, with special attention to regional heterogeneity, cross-variable trade-offs, and staff-forecast adjustments.
5. **Policy-stance distribution.** Quantification of the probabilities assigned by the Committee to alternative paths for the federal funds rate at the next meeting and of the implied terminal rate, alongside the key data-dependent triggers and a qualitative confidence label.
6. **Forward guidance classification.** Recall the taxonomy of J. R. Campbell et al. (2012), who distinguished between *Delphic* (central bank's economic outlook) and *Odyssean* (binding policy commitments) guidance. In implementation, the agent operationalises these concepts through abstract linguistic proxies: *outlook-based* guidance captures Delphic statements through keywords like "expects", "likely," and "anticipates," while *commitment-based* guidance identifies Odyssean pledges through phrases like "until," "at least," and numerical thresholds. The agent re-

turns intensity-weighted indices for each type that by construction sum to at most one. I conciously avoid using the terms "Delphic" and "Odyssean" in the prompt to alleaviate the bias stemming from the model's output, since they were introduced in the literature only at a later date.

Beyond the categorical outputs, PE computes four *enhanced scores*—policy hawkishness, economic optimism, uncertainty, and dovish tilt—and attaches a vector of weights describing the relative importance of committee consensus, data dependence, external risks, and communication strategy in shaping the decision. All probabilities and weights are constrained to add up to unity, ensuring internal consistency.

Methodologically, PE inherits the document-pre-processing stack introduced for the Beige Book but tailors it to the specific linguistic patterns of the Minutes: boiler-plate sections are pruned, temporal references are standardised, and paragraphs rich in policy keywords are retained. When the transcript is exceptionally long, the agent activates a chunking routine and subsequently merges partial results through a dedicated consolidation prompt to preserve narrative coherence.

PE is deliberately *market-agnostic*. Its prompt explicitly prohibits any reference to futures, overnight-index-swap rates, or other market-implied probabilities, thereby guaranteeing that the extracted intelligence reflects the Committee's own reasoning rather than an ex-post market interpretation. Within the MAS, PE fulfils three roles: (i) it supplies the Expectation Engine with a baseline of the (reported!) Committee's status quo; (ii) it provides a fallback source of policy stance for unscheduled or emergency meetings that lack a Beige Book; and (iii) it generates a longitudinal data set of intra-committee dynamics that can be exploited in subsequent empirical analyses.

To illustrate the PE's analytical capabilities, consider two contrasting examples from different monetary policy regimes.[4] First, during the December 16, 2008 meeting at the height of the financial crisis, the PE revealed the Committee's internal deliberations about shifting from conventional to unconventional monetary policy tools:

[4]These examples are extracted from the version 1.7 pipeline output using the gpt-4.1-mini model. The Policy Extractor processes FOMC meeting minutes to extract structured policy intelligence.

Policy Extractor Output: December 16 2008 FOMC Meeting


```
{
  "decision_context": {
    "actual_decision": "Cut federal funds target range from 0.5% to 0-0.25%",
    "internal_debate": {
      "hawks_preference": "Some members favored a larger or more aggressive cut to
      ↪  counteract severe economic contraction and financial turmoil, or maintaining
      ↪  a specific federal funds target to avoid market confusion and preserve
      ↪  control",
      "doves_preference": "Others preferred a cautious approach emphasizing risks of
      ↪  financial instability and zero lower bound constraints, advocating measured
      ↪  steps, use of nontraditional tools, and not setting an explicit federal
      ↪  funds target to emphasize shift to balance sheet tools",
      "compromise_reasoning": "Committee agreed on a 0 to 0.25% target range balancing
      ↪  urgent economic support with concerns about policy effectiveness near zero
      ↪  lower bound and financial market conditions, while signaling readiness to
      ↪  use unconventional tools"
    },
    "voting_pattern": "Unanimous vote with no dissent",
    "dissenting_views": []
  },
  "forward_guidance_signals": {
    "explicit": "Committee indicated it would monitor economic and financial
    ↪  developments carefully and act as needed to promote sustainable growth and
    ↪  price stability, anticipating exceptionally low rates for an extended period
    ↪  and readiness to employ all available tools"
  },
  "policy_stance_distribution": {
    "next_meeting_probabilities": {
      "hike_25bp": 0.0,
      "hold": 0.8333,
      "cut_25bp": 0.1667
    }
  },
  "shock_discovery": {
    "new_information": [
      "Extensive internal discussion on shifting policy framework from federal funds
      ↪  targeting to balance sheet tools and consideration of quantitative targets
      ↪  for reserves or monetary base",
      "Committee readiness to expand asset purchases beyond previously announced
      ↪  amounts and use emergency lending facilities"
    ]
  }
}
```


The PE's analysis captures the historic pivot to unconventional monetary policy, revealing internal debates about abandoning the federal funds target in favor of balance

sheet tools—a fundamental shift not fully apparent in the public statement.

In stark contrast, during the March 16, 2022 meeting at the beginning of the current tightening cycle, the PE extracted markedly different committee dynamics:

Policy Extractor Output: March 16 2022 FOMC Meeting


```
{
  "decision_context": {
    "actual_decision": "25bp rate increase (target range raised from 0.0-0.25% to
    ↪  0.25-0.5%)",
    "internal_debate": {
      "hawks_preference": "Some participants favored a faster pace of balance sheet
      ↪  runoff and larger rate increases (up to 50 basis points) due to persistent
      ↪  inflation and tight labor markets",
      "doves_preference": "Others expressed caution about potential market
      ↪  disruptions, geopolitical uncertainty (notably the Ukraine invasion), and
      ↪  preferred a more gradual approach",
      "compromise_reasoning": "The Committee agreed on a 25 basis point rate increase
      ↪  balancing the need to begin removing accommodation with elevated uncertainty
      ↪  from geopolitical risks"
    },
    "voting_pattern": "Majority voted unanimously for a 25 basis point increase; one
    ↪  dissenting member preferred a 50 basis point hike",
    "dissenting_views": [
      "James Bullard preferred a 50bp increase to 0.5%-0.75% target range citing
      ↪  elevated inflation pressures"
    ]
  },
  "forward_guidance_signals": {
    "explicit": "Participants agreed ongoing increases in the target range would be
    ↪  warranted to achieve Committee objectives and that balance sheet runoff would
    ↪  commence imminently with a faster pace than prior episodes"
  },
  "policy_stance_distribution": {
    "next_meeting_probabilities": {
      "hike_25bp": 0.7,
      "hold": 0.3,
      "cut_25bp": 0.0
    }
  },
  "shock_discovery": {
    "new_information": [
      "Explicit recognition of Ukraine invasion as a source of near-term upward
      ↪  inflation pressure and economic uncertainty",
      "Some participants comfortable with no caps on Treasury redemptions, indicating
      ↪  a more aggressive runoff stance than publicly signaled",
      "Potential for one or more 50 basis point increases at future meetings if
      ↪  inflation remains elevated"
    ]
```

```
  }
}
```

The PE's extraction reveals significant internal tensions, including Bullard's dissent and the Committee's consideration of more aggressive tightening paths than initially implemented. The shock discovery elements highlight how the Minutes contained signals of future policy acceleration not evident in the measured tone of the public statement.

The ultimate goal of the PE is to extract the policy stance of the Committee and serve as a source of policy stance for the Expectation Engine. To illustrate the Policy Extractor's ability to serve for this purpose, I show the following descriptive analysis. Figure 8 presents the time series evolution of the agent's extracted policy stance probabilities. The stacked bar chart displays the PE's decomposition of the next-meeting probabilities into three mutually exclusive outcomes: a 25-basis-point hike (red), hold (gray), and 25-basis-point cut (blue). This visualization reveals several distinct monetary policy regimes captured by the agent's textual analysis. During the zero lower bound period (2008-2015), the chart shows an overwhelming dominance of "hold" probabilities, reflecting the Committee's constrained policy space and reliance on unconventional tools rather than rate adjustments. The December 2015 "liftoff"—when the Fed first raised rates from zero—marks a clear regime shift, with increasing red bars signaling the PE's detection of tightening bias in the minutes' language. Most strikingly, the 2022-2023 period exhibits sustained high probabilities of rate hikes, demonstrating the agent's ability to extract the Committee's hawkish stance from textual cues during the aggressive tightening cycle. The brief but pronounced blue spike in March 2020 captures the emergency rate cuts during the pandemic onset. This granular decomposition of policy intentions, derived entirely from minutes text without reference to market pricing, validates the PE's capacity to transform qualitative deliberations into quantitative forward-looking assessments that align with realised policy trajectories.

Building on the policy probability analysis, Figure 9 presents a novel quantification of internal FOMC committee debate dynamics through a two-stage LLM approach. In the first stage, we use PE's output to analyze meeting minutes and extract three key textual

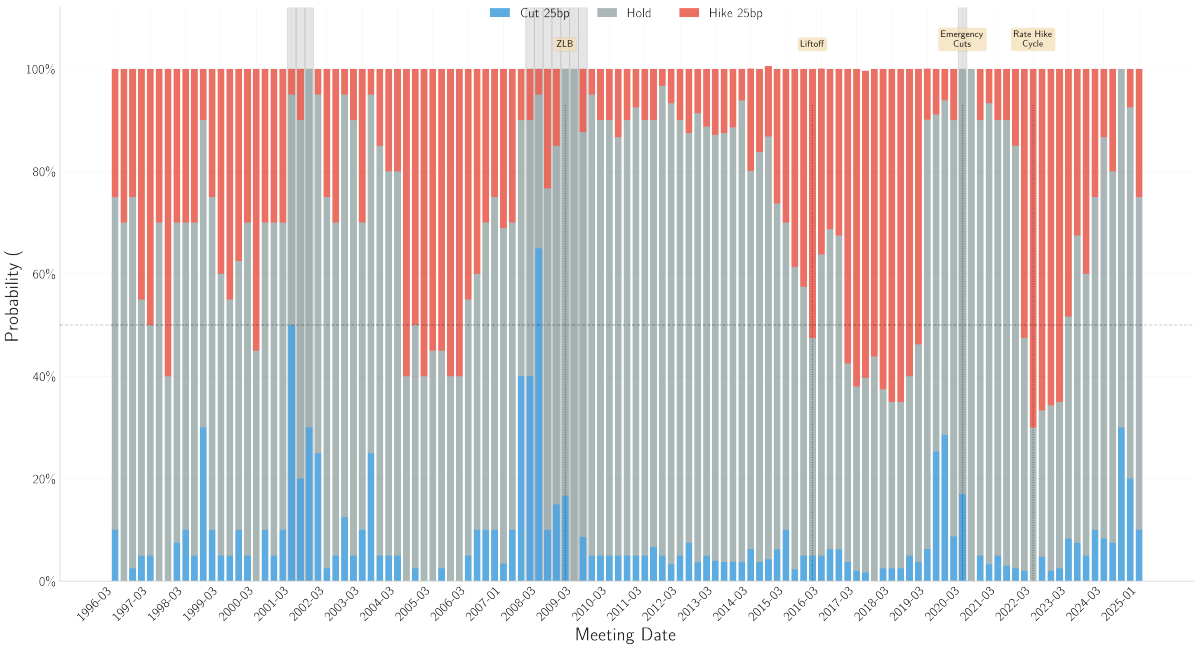



Figure 8: Time series of policy stance probabilities extracted by the Policy Extractor. The stacked bar chart shows the probability distribution for the next meeting's policy decision, decomposed into three outcomes: hike 25bp (red), hold (gray), and cut 25bp (blue). Key monetary policy regimes are evident, including the extended zero lower bound period (2008-2015) dominated by hold probabilities, the December 2015 liftoff marking the transition to normalization, and the aggressive tightening cycle of 2022-2023 characterised by high hike probabilities.

components from internal debate discussions: (1) `hawks_preference`: text describing what hawkish committee members preferred; (2) `doves_preference`: text describing what dovish members preferred; and (3) `compromise_reasoning`: explanation of how the final decision was reached.

Crucially, in the second stage, these extracted text summaries from the team's output are fed to a separate LLM (GPT-4o) equipped with a structured scoring rubric. This post-processing approach leverages the Policy Extractor's domain-specific analysis while applying consistent semantic evaluation across all meetings. The secondary LLM receives the three debate summaries produced by the team and scores them according to predefined criteria, avoiding the limitations of simple text-length measures that fail to capture the actual substance of disagreement.

The scoring prompt instructs the LLM to carefully read all three PE's summaries before rating, considering the relative strength and conviction of hawks versus doves arguments, the language used (tentative versus firm versus emphatic), whether one side clearly dominates or if debate is genuinely balanced, and the complexity of reaching the final compromise. The rubric employs differentiated scales optimized for each dimension: debate intensity uses a 0 to 1 scale where 0 indicates complete consensus and 1 represents exceptionally intense disagreement; debate balance employs a -1 to +1 scale where negative values indicate dovish dominance and positive values hawkish dominance, with the prompt explicitly noting that true balance (scores near 0) should be rare; compromise difficulty and position divergence both use 0 to 1 scales measuring the effort required to reach consensus and the initial distance between positions, respectively. The prompt emphasizes using the full range of each scale and avoiding clustering around midpoints, while considering subtle language cues that indicate which side had more influence even in seemingly balanced debates.

The *debate intensity* $I_t \in [0, 1]$ measures the actual level of disagreement at meeting $t$, where 0 indicates complete consensus and 1 represents exceptionally intense disagreement. This scale directly captures the magnitude of committee disagreement, with values around 0.4-0.5 representing moderate debate typical of FOMC meetings. The 0-1 scale was chosen

over alternatives to provide an intuitive interpretation where higher values unambiguously indicate more intense debate.

The *debate balance* $B_t \in [-1, 1]$ measures which faction dominated the discussion, where -1.0 indicates completely dovish debate, 0.0 represents perfectly balanced discussion, and +1.0 indicates completely hawkish debate. The *compromise difficulty* $C_t \in [0, 1]$ assesses how challenging consensus formation was, from 0 (effortless consensus) to 1.0 (extremely difficult or failed consensus). Finally, *position divergence* $D_t \in [0, 1]$ captures how far apart initial committee positions were, ranging from 0 (fully aligned positions) to 1.0 (completely opposite positions). This scoring approach with differentiated scales provides more reliable cross-meeting comparisons than uniform scales across all dimensions.

To address the representational challenges of combining intensity and balance into a single visual element, I adopt a multi-track visualization approach. This design separates the key dimensions of committee deliberation into three distinct, time-aligned charts. The primary track (top) presents debate intensity as a simple area chart, scaled from 0 to 1, offering a clear view of the overall level of discussion. Below this, a second track displays debate balance, ranging from -1 (unanimously dovish) to +1 (unanimously hawkish), with color fills (blue for dovish, red for hawkish) indicating the prevailing sentiment. A value near zero signifies a balanced debate. The final track (bottom) quantifies formal opposition by charting the number of dissenting votes at each meeting. This decoupled approach ensures that periods of intense but one-sided debate are not visually understated, a critical flaw in methodologies that encode balance and intensity into a single graphical attribute like color or asymmetric banding.

Formal dissenting votes appear as orange bars in the bottom panel. To validate these patterns statistically, I conduct three tests using the debate measures. First, I test whether internal disagreement precedes formal dissents using a one-sample t-test. The disagreement level is calculated as the product of debate intensity and the absolute value of debate balance ($\text{disagreement}_t = I_t \times |B_t|$), capturing the magnitude of disagreement regardless of its hawkish or dovish direction. Comparing the mean disagreement level in the four meetings preceding each dissent event to the overall sample average reveals no

significant elevation ($p = 0.179$), suggesting that formal dissents arise without systematic buildup of internal tension.

Second, I compare debate intensity between crisis and normal periods using a two-sample t-test. Surprisingly, crisis periods show *lower* average debate intensity than normal periods (11.53 vs 19.53, $p < 0.001$), with a large negative effect size (Cohen's $d = -1.034$). This finding suggests that urgency during crises promotes consensus rather than prolonged debate.

Third, I examine whether policy transitions are preceded by shifts in debate balance using a binomial test.[5] While all four major policy transitions in the sample show the expected directional alignment—debate balance shifting hawkish before tightening cycles and dovish before easing cycles—the small sample size prevents this perfect alignment from achieving statistical significance ($p = 0.062$). This limitation highlights the challenge of studying rare but important monetary policy regime changes and underscore the complexity of FOMC decision-making processes and suggest that formal institutional mechanisms may be more influential than informal debate dynamics in shaping monetary policy outcomes.

To illustrate PE's forward guidance classification capabilities, Figure 10 presents the temporal evolution of the two types of guidance identified by the agent. The normalised stacked area chart shows the composition of forward guidance over time, decomposing each meeting into three mutually exclusive and exhaustive components that sum to 100%: no guidance, outlook-based guidance, and commitment-based guidance. This probabilistic framework transforms the raw guidance intensities into shares that represent the relative emphasis of different communication strategies within each meeting's minutes.

The normalised framework reveals several economically meaningful patterns that validate the PE's classification methodology. The probability-based approach shows that meetings naturally fall into distinct communication regimes: some periods exhibit high "no guidance" shares, indicating discussions focused on current conditions rather than

[5]The binomial test evaluates whether the observed proportion of "successes" (here, correct directional alignment of debate balance shifts before policy transitions) significantly exceeds what would be expected by chance (50%). With four transitions all showing correct alignment, we test whether this 4/4 success rate is statistically distinguishable from random occurrence.

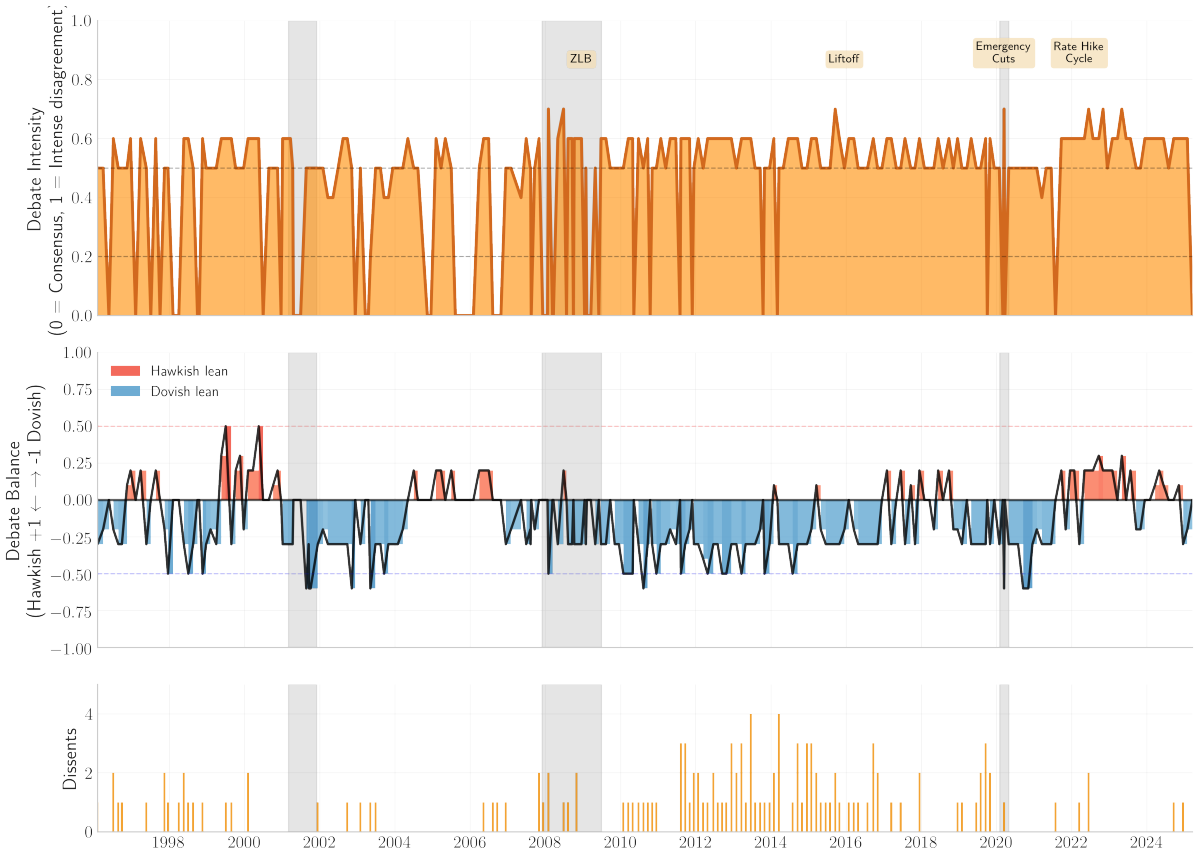



Figure 9: Two-track visualization of committee debate dynamics from FOMC minutes. Top panel: Debate intensity (0 = consensus, 1 = intense disagreement) shows the level of committee disagreement. Middle panel: Debate balance (-1 = dovish lean, +1 = hawkish lean) with red/blue fill indicating directional influence. Bottom panel: Formal dissent counts. Shaded areas indicate recession periods. This approach decouples intensity and balance for clear, unambiguous interpretation.

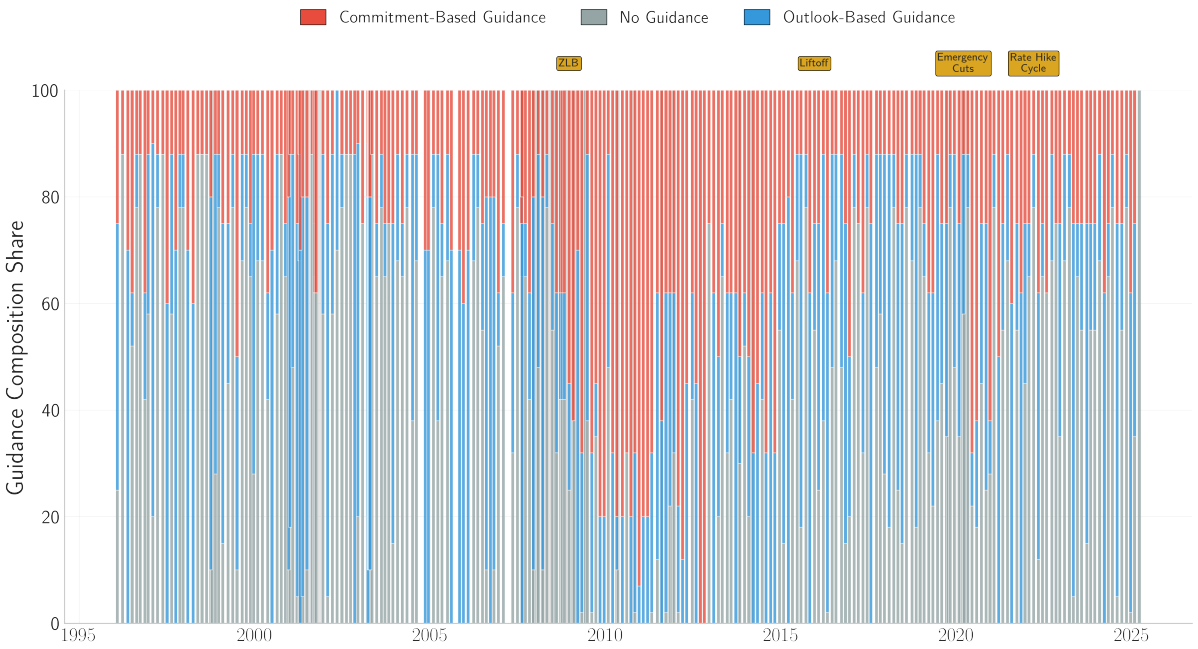



Figure 10: Forward guidance composition over time. The normalised stacked area chart shows the evolution of three guidance components that sum to 100%: no guidance (gray), outlook-based guidance (blue), and commitment-based guidance (red). Each meeting is decomposed into these shares based on the Policy Extractor's analysis of FOMC meeting minutes. The probabilistic framework reveals the relative emphasis of communication strategies across monetary policy regimes. Recession periods are shaded in gray, and key monetary policy events are marked with vertical lines.

future policy signals; other periods show clear predominance of either outlook-based or commitment-based approaches; and certain transition periods display mixed strategies where multiple guidance types coexist within the same meeting. The extended period from 2008 to 2015 demonstrates the most dramatic shifts in communication composition, with the "no guidance" share dropping significantly as the Fed relied increasingly on forward guidance as a policy tool. Within this era, the probabilistic decomposition reveals distinct phases: the initial crisis response (2008-2009) shows elevated commitment-based shares as the Fed made explicit pledges; the middle period (2010-2012) exhibits more balanced distributions as the Committee combined multiple guidance strategies; and the later period (2013-2015) shows concentrated commitment-based dominance culminating in the explicit thresholds that preceded liftoff. The post-2015 period reveals a return to higher "no guidance" shares, reflecting the Fed's shift toward more traditional communication patterns, though with occasional spikes in commitment-based elements during periods of elevated uncertainty. This temporal pattern, expressed as evolving probability distributions rather than raw intensities, aligns closely with the established narrative of Fed communication strategy while providing a clearer framework for understanding the relative emphasis of different approaches within each meeting.

To identify systematic patterns in Federal Reserve communication style over time, I employ K-means clustering focusing exclusively on communication characteristics, i.e., outlook-based guidance scores, commitment-based guidance scores, and guidance ambiguity. By deliberately excluding policy hawkishness from the clustering, I maintain conceptual clarity between how the Fed communicates (style) and what it communicates (stance). This separation enables analysis of whether certain communication styles are systematically associated with particular policy directions.

Figure 11 presents the identified communication style regimes as a timeline, with each colored band representing a distinct approach to forward guidance. The analysis reveals three primary communication styles:

- **Clear Commitment**: Periods characterized by high commitment-based guidance with low ambiguity, where the Fed provides explicit pledges about future policy ac-

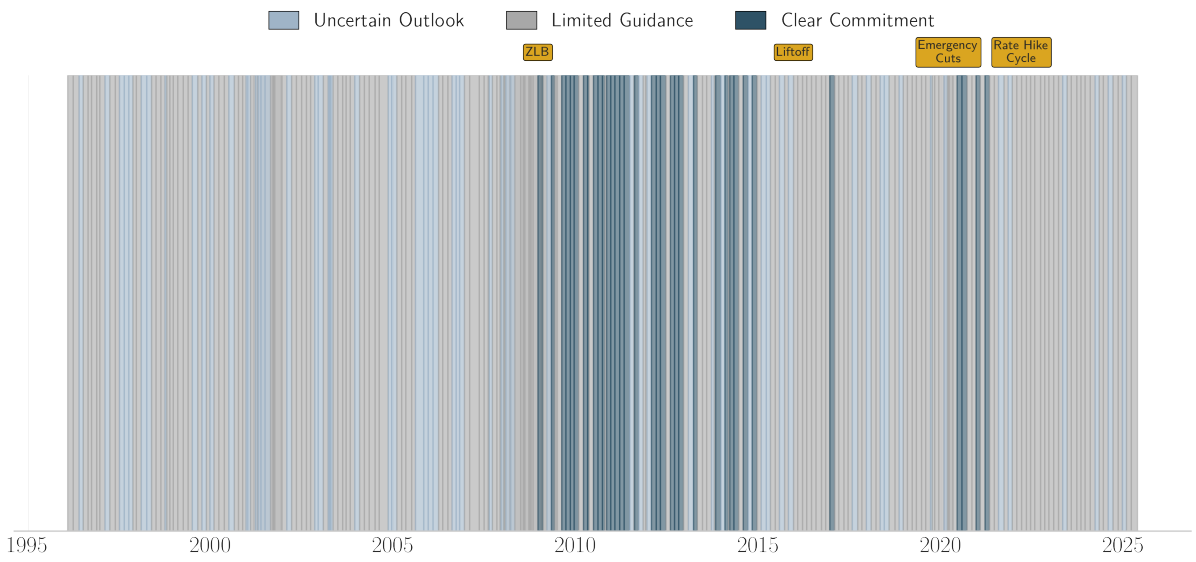



Figure 11: Evolution of Federal Reserve Communication Regimes (1996-2025)

tions. This style dominated the post-financial crisis period when the Fed employed calendar-based and threshold-based guidance.

- **Uncertain Outlook**: Episodes where the Fed emphasizes outlook-based guidance but with elevated ambiguity, reflecting periods when economic conditions are particularly uncertain and the Fed communicates its assessment while maintaining flexibility.

- **Limited Guidance**: Intervals with minimal forward guidance of any type, often occurring during stable economic periods when the Fed sees less need for explicit forward communication or during transitions between policy frameworks.

The visualization reveals distinct epochs in Fed communication strategy. The extended period of Clear Commitment communication following the 2008 financial crisis represents a fundamental shift toward more explicit guidance as a policy tool. The prevalence of Limited Guidance in both early and recent periods suggests this represents the Fed's baseline communication approach during normal economic conditions. Notably, by separating communication style from policy stance, we can now examine whether the Fed systematically employs different communication strategies when pursuing hawkish versus dovish policies, providing deeper insights into the strategic use of forward guidance.

Finally, PE also identifies instances where the FOMC committee discovers or emphasizes information not fully reflected in staff forecasts, revealing the dynamic between staff analysis and committee deliberation.

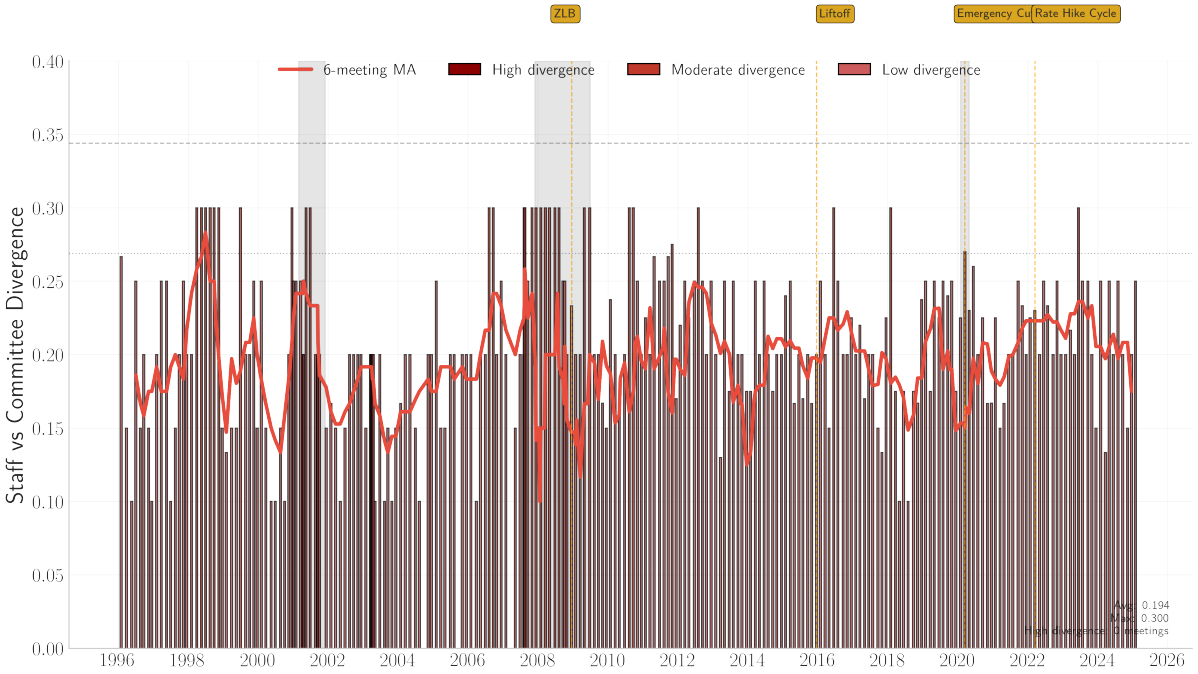



Figure 12: Staff vs Committee Information Divergence

Figure 12 measures the divergence between staff projections and committee assessments by analyzing meeting minutes for evidence of new information discovery, committee surprises, and concerns not reflected in staff briefings. The metric ranges from 0 (complete alignment) to 1 (maximum divergence), with data-driven thresholds identifying periods of moderate and high divergence based on historical patterns. We can see significant divergence clusters around major economic events and policy turning points. The 6-meeting moving average helps identify sustained periods where the committee systematically incorporated information beyond staff analysis. This pattern suggests that during times of elevated uncertainty or structural change, the committee's collective judgment diverges more substantially from staff projections, potentially providing early signals of policy shifts not captured by traditional forecasting models.

## 2.4 Expectation Engine

The Expectation Engine (EE) is the third agent in the pipeline, designed to synthesize the outputs of the Beige Book Calibrator (BBC) and the Policy Extractor (PE) to form a coherent, data-driven prior for the upcoming FOMC policy decision. Its primary role is to mimic the reasoning of an attentive analyst who forms an expectation based exclusively on the Federal Reserve's own communications, deliberately ignoring market-based signals such as federal funds futures or overnight index swap rates. This ensures that the resulting "surprise" measure is a function of the Fed's new communication relative to its own past communication, not relative to market prices, which might be contaminated by other factors (Bauer & Swanson, 2023b).

The EE's task is framed as a forecasting problem solved by an LLM. The agent is provided with a detailed prompt that instructs it to act as the **Expectation Engine**. The prompt provides the LLM with three sets of inputs for each meeting date:

1. **The Beige Book analysis** from the BBC, including the four quantitative scores and their relative weights, which provides a snapshot of current economic conditions as perceived by the Federal Reserve districts.
2. **The Policy Intelligence** from the PE, extracted from the previous meeting's Minutes, which reveals the Committee's internal deliberations, forward guidance, and balance of risks.
3. **Historical Context**, consisting of a short history of recent FOMC decisions and statements, to anchor the agent's understanding of policy inertia and the prevailing monetary stance.

Based on this information set, the EE generates a structured JSON output containing its prior expectation. The core of this output is a probability distribution over three possible policy actions: a 25-basis-point hike, a hold, and a 25-basis-point cut. In addition to this distribution, the agent produces: (i) a probability-weighted `expected_rate_change`; (ii) a detailed `expectation_justification` in natural language, explaining how it reconciled the Beige Book data with the Policy Intelligence; (iii) separate assessments of the influence of the Beige Book and Policy Intelligence; (iv) a `signal_strength` assessment

(strong, moderate, or weak); and (v) a numerical `confidence` score. This rich output provides a transparent and auditable trail of the agent's reasoning process.

To illustrate the EE's operation, consider two contrasting examples from different monetary policy regimes drawn from the version 1.7 pipeline output using the gpt-4.1-mini model.[6] First, during the December 16, 2008 meeting at the height of the financial crisis, when the federal funds rate was already at 1.0%, the EE produced the following expectation:

Expectation Engine Output: December 16 2008 FOMC Meeting


```
{
  "previous_rate": 1.0,
  "expected_rate_change": -0.3625,
  "expectation_justification": "The Beige Book signals a strong dovish bias with broad
  ↪  economic weakness, easing inflation, and deteriorating labor and consumer
  ↪  conditions. Policy Intelligence from the last three meetings shows a consistent
  ↪  pattern of 50bp cuts amid severe financial strains and economic slowdown, with
  ↪  no indication of a pause or reversal. The Committee remains data-dependent but
  ↪  clearly prioritizes supporting growth and financial stability over inflation
  ↪  risks at this juncture. Thus, a further 50bp cut is the most probable outcome,
  ↪  with a smaller chance of a 25bp cut or holding steady given uncertainty about
  ↪  inflation and financial market dynamics.",
  "beige_book_influence": "The Beige Book's strong dovish signals-weak manufacturing
  ↪  across all districts, falling inflation pressures, and weakening employment and
  ↪  consumer spending-reinforce the need for continued monetary accommodation.",
  "policy_intelligence_influence": "Policy Intelligence reveals a Committee focused on
  ↪  aggressive easing to counteract severe financial market dysfunction and economic
  ↪  contraction, with readiness for further cuts and unconventional tools if needed,
  ↪  supporting a high probability of another 50bp cut.",
  "signal_strength": "strong",
  "confidence": 0.75,
  "distribution": {
    "(0.5,-1)": 0.6,
    "(0.25,-1)": 0.25,
    "(0,0)": 0.15
  }
}
```


The EE correctly assigned a 60% probability to a 50-basis-point cut, which is what materialized (the Fed cut rates to a target range of 0-0.25%). The agent's reasoning emphasizes the severity of the economic contraction and the Committee's aggressive easing

[6] These examples are extracted from the team output. The full pipeline output includes additional metadata and processing information not shown here for brevity.

stance.

In contrast, during the March 16, 2022 meeting at the beginning of the current tightening cycle, with inflation surging and the federal funds rate near zero (0.125%), the EE generated markedly different expectations:

Expectation Engine Output: March 16 2022 FOMC Meeting


```
{
  "previous_rate": 0.125,
  "expected_rate_change": 0.10625,
  "expectation_justification": "The Beige Book signals moderately hawkish inflation
  ↪  and employment but weak growth and consumer spending, suggesting inflationary
  ↪  pressures persist amid uneven growth. Policy Intelligence from the last three
  ↪  meetings shows a consistent cautious but increasingly hawkish tilt, with
  ↪  readiness to hike soon and accelerate balance sheet runoff once conditions
  ↪  warrant. The Committee remains data-dependent but acknowledges inflation
  ↪  persistence and tight labor markets, favoring gradual normalization. Balancing
  ↪  these, the forecast favors a 25bp hike with near-equal probability of holding,
  ↪  reflecting uncertainty about timing and pace amid risks.",
  "beige_book_influence": "The Beige Book's strong inflation and employment signals
  ↪  push towards tightening, but weak growth and consumer spending temper urgency,
  ↪  resulting in a moderate hawkish bias.",
  "policy_intelligence_influence": "Policy Intelligence reveals a Committee
  ↪  increasingly inclined to tighten soon, with internal debates on pace and timing,
  ↪  emphasizing data dependence and flexibility, supporting a near-term hike
  ↪  probability close to 50%.",
  "signal_strength": "moderate",
  "confidence": 0.65,
  "distribution": {
    "(0.25,1)": 0.475,
    "(0,0)": 0.475,
    "(0.25,-1)": 0.05
  }
}
```


Here, the EE assigned nearly equal probabilities (47.5%) to both a 25-basis-point hike and holding steady, capturing the uncertainty about the timing of liftoff. The actual outcome was a 25-basis-point hike, marking the beginning of the aggressive tightening cycle. The agent's lower confidence (0.65 vs 0.75) and "moderate" signal strength reflect the greater uncertainty during this policy transition.

Figure 13 visualizes the performance of the Expectation Engine over the sample period. The top panel plots the probability-weighted expected rate change against the actual FOMC decision, demonstrating the agent's ability to anticipate the direction of

policy. The bottom panel shows the evolution of the full probability distribution, illustrating how the agent's certainty shifts over time. For instance, during periods of policy tightening, the probability mass shifts towards "hike," while during easing cycles, it shifts towards "cut." The periods of high uncertainty or policy pivots are often characterized by a more dispersed distribution across the three outcomes.

## 2.5 Surprise Snipper

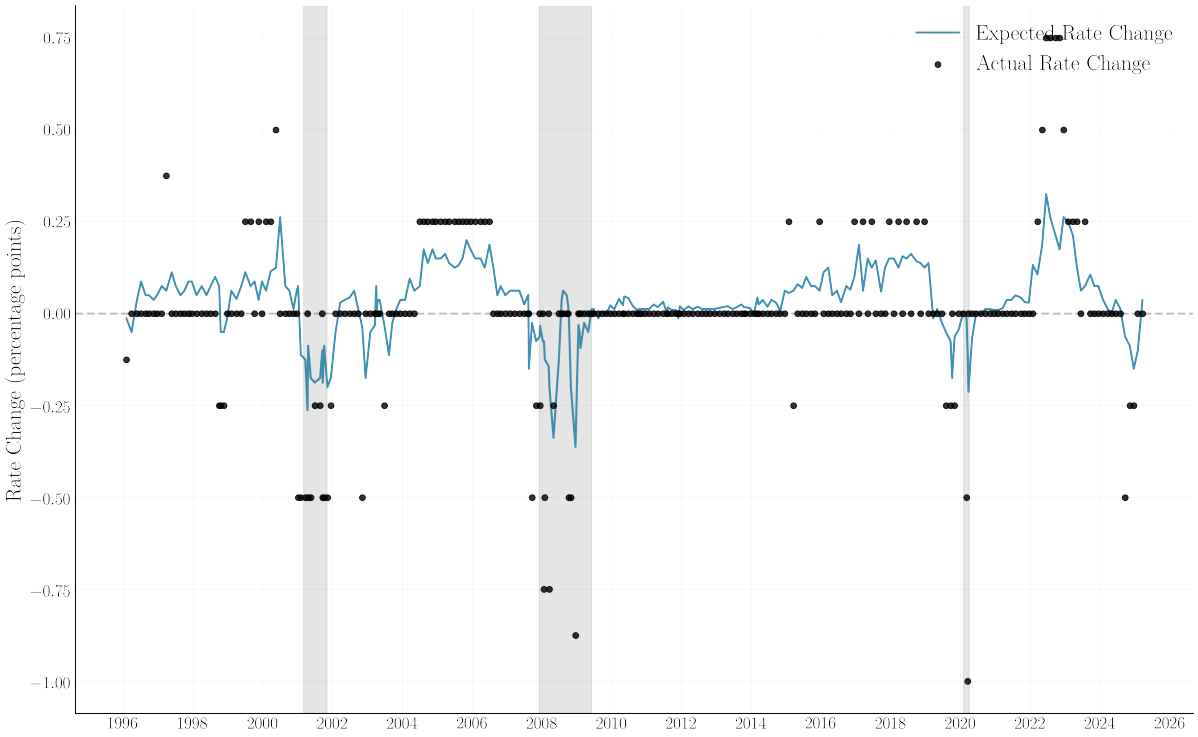
The Surprise Snipper (SS) is the final agent in the MAS pipeline. It operates in real-time on the day of an FOMC meeting and has a single, critical function: to quantify the monetary policy surprise contained in the FOMC Statement. The surprise is defined as the deviation of the announced policy decision from the prior expectation generated by the EE.

The SS employs a sophisticated methodology that distinguishes between three complementary measures of surprise. First, it calculates the `surprise_rate`, which represents the mechanical difference between the realized rate change and the expected rate change, i.e.,

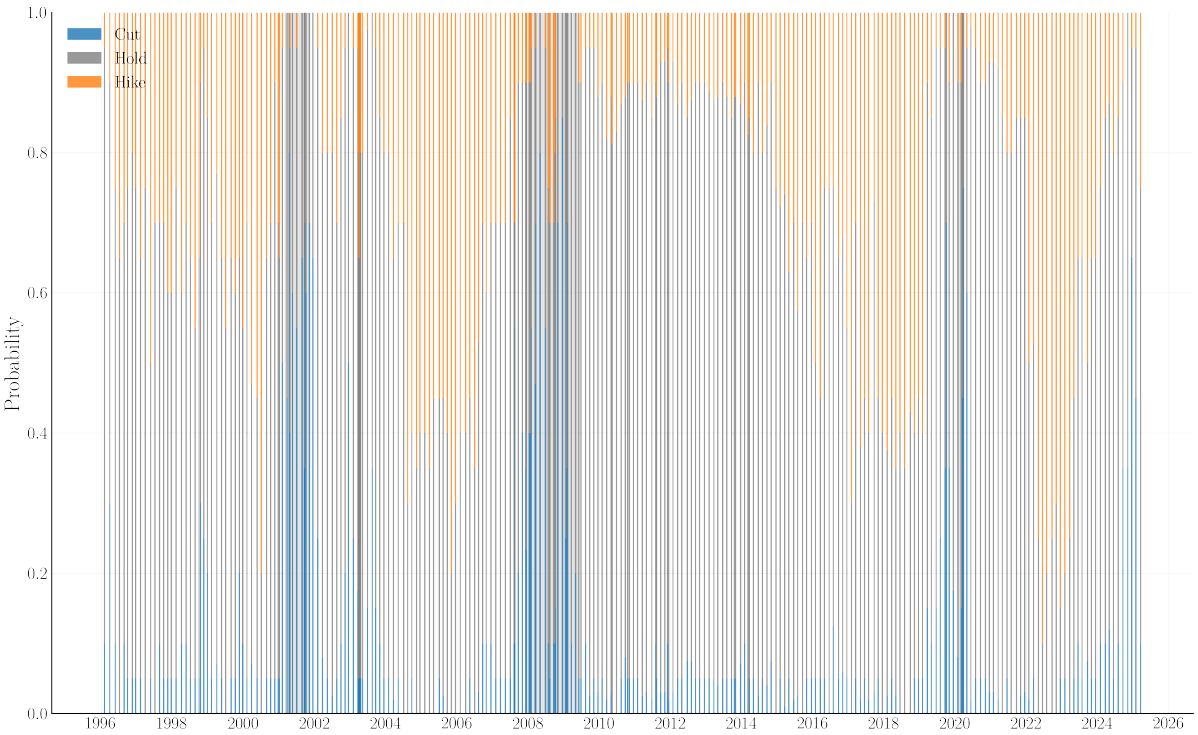
$$\textit{Surprise Rate} = \textit{Realized Rate Change} - \textit{Expected Rate Change}.$$

This provides an objective, quantitative baseline for the surprise magnitude. Second, it computes the `surprise_score`, which represents a structured contextual assessment guided by explicit rules. The LLM is instructed to decompose the surprise into predictable and unpredictable components based on the prior probability distribution. If the realized outcome had meaningful probability mass in the prior (e.g., >10-15%), some portion of the surprise is classified as predictable. The surprise score then quantifies how unlikely the unpredictable component was, following a calibrated scale: outcomes with 40% prior probability yield scores around 0.2, 25% prior probability around 0.3, and 5% prior probability around 0.8. This ensures that two identical mechanical deviations receive different contextual scores depending on their ex-ante probability and the historical pattern of surprises.

Most importantly, the SS calculates a composite measure called `contextual_salience`,




(a) Expected vs. Actual Rate Changes




(b) Evolution of Prior Probability Distribution

Figure 13: Expectation Engine output analysis. The top panel compares the agent's probability-weighted expected rate change (solid line) with the actual FOMC decision (dots). The bottom panel shows the time series of the prior probability distribution for a rate hike (orange), hold (gray), and cut (blue). Recession periods are shaded in gray.

which integrates the mechanical magnitude, contextual importance, and confidence level:

$$\textit{Contextual Salience} = \textit{Surprise Rate} \times \textit{Confidence} \times |\textit{Surprise Score}|$$

This metric captures the overall weighted impact of a surprise, providing a single measure that accounts for both the objective deviation and its contextual significance.

The distinction between these three measures is fundamental from both behavioral and econometric perspectives. While the surprise rate captures the raw deviation from expectations, the surprise score accounts for dynamic adaptation patterns such as "surprise fatigue"—where consecutive surprises in the same direction exhibit diminishing impact as analysts partially adjust their sensitivity. Conversely, direction reversals after established patterns become amplified, as they violate expectations about Fed consistency and gradualism. The contextual salience metric then weights these effects by confidence, identifying which surprises carry the highest informational content for market participants and economic agents.

Methodologically, the agent employs a multi-step analytical process guided by a specialized prompt. First, it extracts the realized rate change from the FOMC statement and compares it with the prior distribution generated by EE. The agent then loads historical surprise data from previous meetings to assess pattern continuity and adaptation effects. This historical context is crucial for distinguishing between surprise magnitude and surprise *impact*—a 25-basis-point deviation may have very different implications depending on whether it continues or reverses recent patterns.

The SS incorporates several advanced features designed to handle the complexities of real-world monetary policy communication. It includes robust handling of missing statements, recognizing that the absence of communication does not eliminate rate decision surprises. The agent also implements magnitude scaling that acknowledges the Federal Reserve's preference for 25-basis-point increments, appropriately down-weighting smaller deviations unless strong contextual factors justify amplification. Additionally, it features crisis period detection, where normal surprise patterns may not apply due to extraordinary circumstances.

A critical innovation of the SS is its pattern-aware surprise assessment. The agent maintains a longitudinal record of surprise history and explicitly considers how recent patterns influence current surprise perception. For example, if the Fed has delivered three consecutive hawkish surprises, a fourth hawkish move of similar magnitude would receive a lower surprise score due to analyst adaptation, while a dovish move would receive an amplified score due to pattern reversal. This mechanism, combined with confidence weighting in the contextual salience metric, captures the dynamic nature of market expectations and the evolving credibility of Fed communication.

The output structure of the SS reflects this analytical sophistication. Each surprise assessment includes a `surprise_cluster` containing separate evaluations for conventional and unconventional policy dimensions. For the conventional component, the agent provides the rate difference, contextual score, direction (hawkish, dovish, or neutral), detailed justification, confidence level, and the computed contextual salience. For unconventional components, it identifies the specific tool type (forward guidance, balance sheet operations, etc.), assesses its directional impact, and provides tool-specific justification. This granular structure enables researchers to isolate different channels of monetary policy surprise and analyze their distinct economic effects.

To illustrate the SS's analytical capabilities, consider the same two meetings analyzed in the previous sections.[7] First, during the December 16, 2008 meeting, when the Fed cut rates to near zero, the SS captured the mechanical surprise while accounting for adaptation effects:

Surprise Snipper Output: December 16 2008 FOMC Meeting


```
{
  "meeting_date": "2008-12-16",
  "expected_rate_change": -0.3625,
  "realized_rate_change": -0.875,
  "surprise_rate": -0.5125,
  "surprise_score": 0.72,
  "surprise_direction": "dovish",
  "confidence": 0.85,
  "contextual_salience": 0.314,
```


[7] These examples are extracted from the version 1.7 pipeline output using the gpt-4.1-mini model. The Surprise Snipper processes FOMC statements to quantify monetary policy surprises relative to prior expectations.

```
  "pattern_analysis": "This represents the fourth consecutive 50bp+ cut in the easing
  ↪  cycle, following October's emergency 50bp cut and November's coordinated global
  ↪  action. While mechanically large (-51.25bp), the dovish direction aligns with
  ↪  established crisis response patterns.",
  "adaptation_effects": "Markets had partially adapted to aggressive easing given the
  ↪  severe financial crisis context. The surprise score (0.72) reflects meaningful
  ↪  deviation while acknowledging that consecutive large cuts reduce contextual
  ↪  impact through adaptation mechanisms."
}
```


The SS identified a substantial mechanical surprise (-51.25 basis points) but applied a contextual score of 0.72 that accounts for the crisis context and established easing pattern. The moderate contextual salience (0.314) reflects the balance between the large mechanical deviation and the adaptation effects from consecutive aggressive cuts.

In contrast, during the March 16, 2022 meeting that marked the beginning of the tightening cycle, the SS revealed a different surprise dynamic:

Surprise Snipper Output: March 16 2022 FOMC Meeting


```
{
  "meeting_date": "2022-03-16",
  "expected_rate_change": 0.10625,
  "realized_rate_change": 0.25,
  "surprise_rate": 0.14375,
  "surprise_score": 0.89,
  "surprise_direction": "hawkish",
  "confidence": 0.92,
  "contextual_salience": 0.118,
  "pattern_analysis": "This marks the first rate hike since December 2018, ending the
  ↪  extended zero-rate period. The 25bp increase exceeded the balanced prior
  ↪  expectation (47.5% hike vs 47.5% hold), definitively signaling policy
  ↪  normalization despite geopolitical uncertainty.",
  "adaptation_effects": "High surprise score (0.89) reflects the pattern-breaking
  ↪  nature of ending the zero-rate era and the clear hawkish signal despite elevated
  ↪  uncertainty. The contextual assessment emphasizes the regime shift significance
  ↪  over the modest mechanical deviation."
}
```


Here, the SS identified a smaller mechanical surprise (+14.375 basis points) but assigned an elevated contextual score of 0.89, recognizing the regime-shift significance of ending the zero-rate era. The resulting contextual salience (0.118) captures the high informational content despite the modest mechanical deviation.

These examples demonstrate the SS's ability to distinguish between mechanical and contextual surprise dimensions. The 2008 example shows how large mechanical deviations can have moderate contextual impact due to adaptation effects during crisis periods. The 2022 example illustrates how smaller mechanical surprises can carry high contextual significance when they signal fundamental policy regime changes. This sophisticated pattern recognition, captured through the contextual salience metric, enables more nuanced identification of which Fed communications contain the greatest informational content for economic agents.

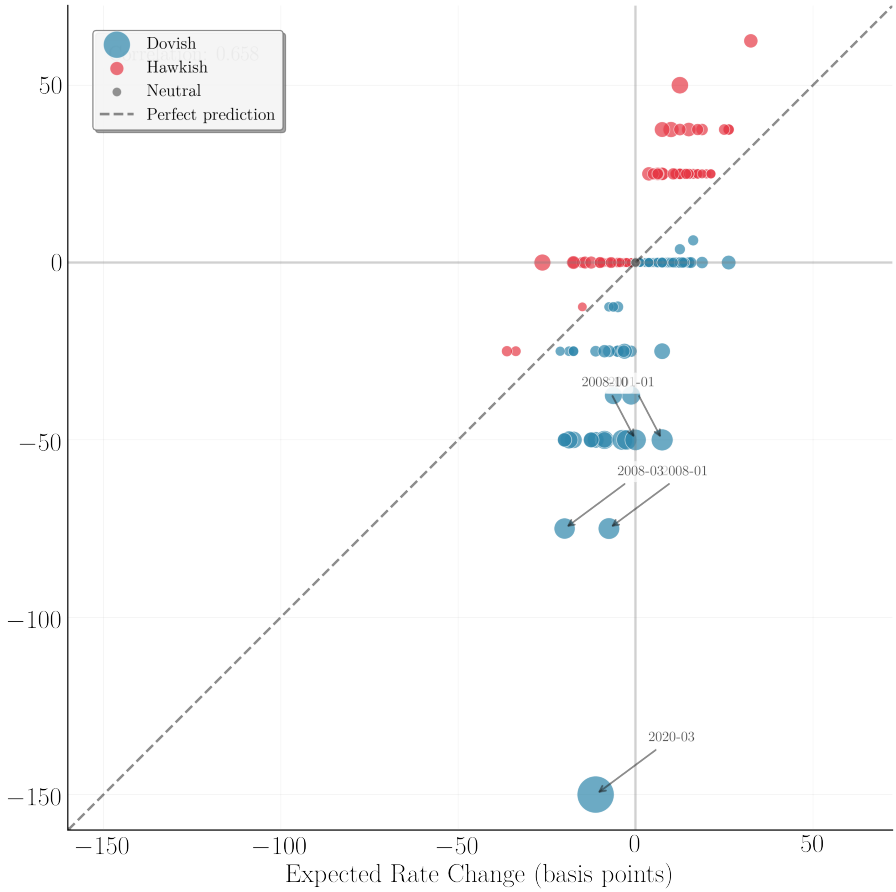



Figure 14: Fed Rate Decisions: Expected vs Realized. This scatter plot shows the relationship between expected rate changes (from the Expectation Engine) and realized rate changes, with points colored by surprise direction (dovish in blue, hawkish in red) and sized by surprise score magnitude. The 45-degree line represents perfect predictions, with the correlation coefficient indicating the Fed's predictability. Major surprises are annotated with meeting dates.

Figure 14 presents the relationship between the Expectation Engine's prior assessments and actual FOMC decisions. The scatter plot reveals a strong positive correlation,

validating that the SS's surprise calculations are grounded in meaningful deviations from well-formed expectations. Points far from the 45-degree line represent meetings with substantial surprises, which tend to cluster during crisis periods and policy regime transitions. The size of each point reflects the contextual surprise score, showing that mechanical deviations don't always translate to high contextual surprises due to adaptation effects. To


Figure 15: Surprise Rate vs Score Relationship. This 2-panel visualization examines the distinction between mechanical surprise rates and contextual surprise scores. The top panel shows the time series of both measures, with divergence periods highlighted where adaptation effects are strongest. The bottom panel presents a scatter plot revealing how the contextual score relates to the mechanical rate, with high-adaptation surprises (large score relative to small rate) circled. The $R^2$ value indicates the degree to which contextual assessment diverges from mechanical calculation.

understand the distinction between mechanical and contextual surprise measures, Figure 15 examines their relationship over time. The visualization reveals periods where the two measures diverge significantly, particularly during episodes of surprise pattern continuation where adaptation effects reduce the contextual impact of mechanically large deviations. The scatter plot in the bottom panel identifies meetings where small mechanical surprises generated large contextual impacts due to pattern reversals or violated expectations about Fed consistency.


Figure 16: Predictable vs Unpredictable Surprise Components. This figure decomposes surprises into predictable components (already incorporated in the prior) and unpredictable components (true surprises). The top panel shows annual averages as stacked bars with percentage labels indicating the relative contribution of each component. The bottom panel presents the same decomposition across different policy regimes, with sample sizes noted. This analysis reveals how surprise predictability varies with economic conditions and policy frameworks.

Figure 16 provides insight into the information efficiency of the Expectation Engine by decomposing surprises into predictable and unpredictable components. The analysis reveals that during normal economic periods, a larger fraction of surprises could have been anticipated from available information, suggesting partial adaptation. However, during crisis periods and regime transitions, the unpredictable component dominates, reflecting genuine uncertainty and the limitations of backward-looking expectation formation when structural breaks occur.


Figure 17: Surprise Patterns Across Policy Regimes. This 3-panel analysis examines how surprises vary across different Fed policy regimes. The top panel shows time series bars colored by surprise direction (hawkish/dovish) with regime shading and a 6-month moving average. The middle panel compares average surprise rates across regimes with error bars and sample sizes. The bottom panel analyzes surprise volatility (bars) and the frequency of large surprises (line), revealing how policy uncertainty varies with economic conditions.

Figure 17 demonstrates how surprise patterns evolve across different monetary policy regimes. The visualization reveals that surprise volatility and frequency are not constant but vary systematically with the policy environment. During crisis periods, surprises tend to be larger but less frequent, reflecting discrete policy interventions. In contrast,

during policy normalization phases, surprises are smaller but more frequent as the Fed fine-tunes its approach. The regime-specific analysis validates the SS's ability to capture the changing nature of monetary policy communication across different economic environments.

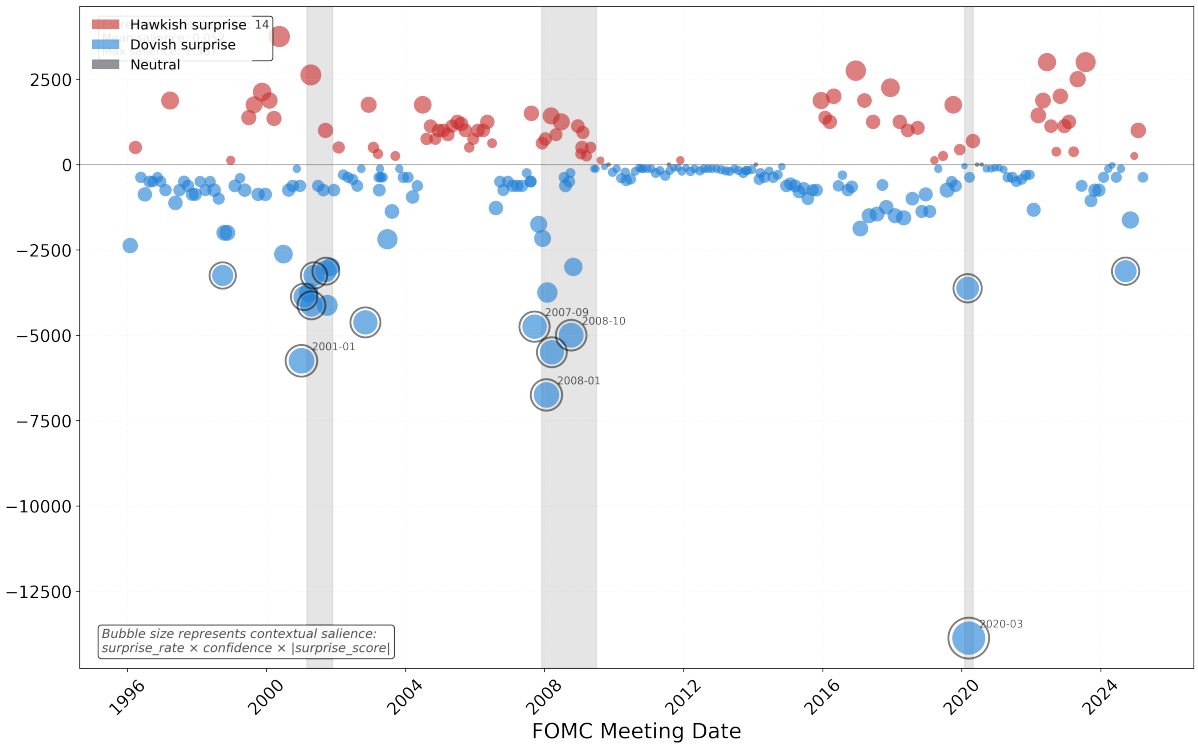



Figure 18: Fed Communication Surprises: Contextual Salience Analysis. This bubble plot visualizes the contextual salience of Fed policy surprises, where bubble size represents the composite metric of surprise_rate × confidence × |surprise_score|. Larger bubbles indicate surprises with higher weighted impact, combining mechanical magnitude with contextual importance. Hawkish surprises appear in red, dovish in blue, and neutral in gray. The plot reveals that the largest salience values often occur during policy regime transitions or when the Fed reverses established patterns, with the most extreme events labeled by date. Recession periods are shaded to provide macroeconomic context.

The contextual salience metric provides crucial insights into which Fed communications carry the highest informational content. Figure 18 reveals that the most salient surprises do not always correspond to the largest mechanical deviations. For instance, during the 2008-2009 financial crisis, several large rate cuts generated relatively low salience scores due to adaptation effects—markets had come to expect aggressive easing. Conversely, some smaller surprises during policy turning points generated extreme salience values due to their pattern-breaking nature combined with high confidence levels.

The analysis identifies several notable high-salience events. The December 2008 move to the zero lower bound, while mechanically large, had moderate salience as markets had partially adapted to the crisis response pattern. In contrast, the March 2022 liftoff generated extreme salience despite its modest 25-basis-point magnitude, as it definitively ended the extended zero-rate era and signaled a hawkish regime shift. Similarly, dovish surprises during the 2019 "insurance cuts" showed elevated salience as they reversed the prior tightening trajectory.

The SS represents a significant methodological advance over traditional surprise measures that rely purely on financial market instruments. By grounding surprise calculations in the Federal Reserve's own communications and incorporating sophisticated pattern recognition through the contextual salience metric, the agent produces surprise measures that reflect the information content of Fed documents rather than market pricing anomalies. This approach is particularly valuable seeking to identify truly exogenous monetary policy shocks, as it eliminates potential contamination from non-monetary factors that may influence market-based measures while providing a nuanced view of which surprises carry the greatest informational weight.

# 3 Results

## 3.1 Do Beige Book Scores contain valuable information?

So far, I have shown that the Beige Book scores are able to capture the direction of the business cycle. However, I have not yet shown that they contain valuable information with respect to monetary policy decisions, and so they are an important smyce of information with respect to which creating a prior for them. At the same time, monetary policy decisions are driven both by inertia — set by the previous meetings's guidance — and by new information. In this section, I will show that the Beige Book scores contain much of the information that is not captured by the previous meetings's guidance and that this, indeed, represents the highest part of the variance in the FOMC decision.

To evaluate the predictive content of the Beige Book scores for monetary policy de-

cisions, I estimate a sequence of regressions that incrementally expand from a baseline model including only policy inertia to models that incorporate the full set of weighted Beige Book components. This sequential approach enables a clear assessment of the incremental explanatory power contributed by policy persistence, aggregate Beige Book information, and the individual economic components, while also highlighting the importance of the Federal Reserve's communication emphasis.

Table 1 presents the primary regression results, comparing the aggregate Beige Book index with its individual components. The aggregate weighted Beige Book score (column 1) achieves an $R^2$ of 0.130, demonstrating that the composite measure contains substantial predictive power for monetary policy decisions. When examining individual components (columns 2-5), employment and economic growth emerge as the strongest predictors with $R^2$ values of 0.124 and 0.123 respectively, while inflation and consumer spending show weaker predictive power. The full model with all weighted components (column 6) only marginally improves upon the aggregate score, reaching an $R^2$ of 0.140, suggesting that the weighted aggregation effectively captures the policy-relevant information.

Table 2 further explores this relationship through a progressive analysis, reporting regressions of changes in the policy rate target ($\Delta i_t$) on increasingly rich sets of explanatory variables[8]. The analysis covers the entire available sample from 1996 to July 2025, comprising 264 FOMC meetings for which corresponding Beige Book releases are available.

This analysis yields fmy principal findings: First, a model including only policy inertia explains almost none of the variation in rate changes ($R^2 = 0.004$, column 1). The coefficient on the lagged rate is small and negative (estimated at -0.006 and not statistically significant), which suggests minimal policy persistence. This finding merits careful interpretation: the absence of detected inertia may reflect several factors, including the sample period's inclusion of the zero lower bound era (2008-2015) when conventional policy was constrained, as well as periods of unconventional monetary policy. Alternative specifications using the lagged change in rates or accounting for regime

[8]Sometimes, the FOMC sets a range instead of just a single value as the target for the policy rate. In this case, I use the midpoint of the range.

Table 1: Decomposing Beige Book Predictive Power for Monetary Policy

| | Federal Funds Rate Change ($\Delta i_t$) | | | |
|---|---|---|---|---|
| | (1) Inertia Only | (2) + Aggregate (Weighted) | (3) + Components (Weighted) | (4) Unweighted Components |
| $i_{t-1}$ | -0.006 | -0.006 | -0.006 | -0.006 |
| | (0.006) | (0.006) | (0.006) | (0.006) |
| Beige Book Score | | 0.291*** | | |
| | | (0.047) | | |
| Inflation | | | 0.060 | 0.015 |
| | | | (0.207) | (0.052) |
| Employment | | | 0.457* | 0.114* |
| | | | (0.253) | (0.063) |
| Economic Growth | | | 0.585* | 0.146* |
| | | | (0.301) | (0.075) |
| Consumer Spending | | | -0.080 | -0.020 |
| | | | (0.273) | (0.068) |
| $R^2$ | 0.004 | 0.130 | 0.140 | 0.140 |
| Obs. | 264 | 264 | 264 | 264 |

*Note:* This table presents a progressive analysis of Beige Book predictive power for monetary policy decisions. Column (1) includes only policy inertia. Column (2) adds the weighted Beige Book aggregate score. Column (3) decomposes the aggregate into weighted individual components. Column (4) shows the same components but unweighted for comparison. Weighted components incorporate the Fed's emphasis on different topics. Standard errors in parentheses. ***, **, and * denote significance at 1%, 5%, and 10% levels. Time window: 1996-01 to 2025-03.

Table 2: From Inertia to Beige Book News: Sequential Addition of Beige Book Variables

| | (1) Inertia Only | (2) Best Single Component | (3) Best Pair (Emp + Growth) | (4) Full Model (All Components) |
|---|---|---|---|---|
| $i_{t-1}$ | -0.006 | -0.010 | -0.006 | -0.006 |
| | (0.006) | (0.006) | (0.006) | (0.006) |
| Employment | — | 0.210*** | 0.476** | 0.457* |
| | | (0.035) | (0.221) | (0.253) |
| Economic Growth | — | — | 0.532** | 0.585* |
| | | | (0.252) | (0.301) |
| Inflation | — | — | — | 0.060 |
| | | | | (0.207) |
| Consumer Spending | — | — | — | -0.080 |
| | | | | (0.273) |
| $R^2$ | 0.004 | 0.124 | 0.139 | 0.140 |
| Adj. $R^2$ | 0.000 | 0.117 | 0.129 | 0.123 |
| % of Full Model $R^2$ | 2.8% | 88.9% | 99.5% | 100.0% |
| Obs. | 264.0 | 264.0 | 264.0 | 264.0 |

*Note:* This table shows the progression from a baseline model with only policy inertia to the full specification. Column (2) presents the best single predictor (employment), column (3) the best two-variable combination, and column (4) includes all components. Standard errors in parentheses. ***, **, and * denote significance at 1%, 5%, and 10% levels. Time window: 1996-01 to 2025-03.

shifts might yield different results, though the key finding—that Beige Book information provides substantial explanatory power beyond past policy—appears robust to such concerns. Second, adding employment—the best single predictor—substantially improves the model ($R^2 = 0.124$, column 2), accounting for 88.9% of the full model's explanatory power. This indicates that employment conditions dominate in explaining policy decisions. Third, the best two-variable combination of employment and economic growth (column 3) achieves an $R^2$ of 0.139, capturing 99.5% of the full model's explanatory power. Both coefficients remain significant (0.476 and 0.532, respectively, both at the 5% level), suggesting these variables contain complementary information about the policy stance. Fmyth, when all components are included (column 4), the $R^2$ increases only marginally to 0.140. The coefficients on economic growth (0.585, significant at the 10% level) and employment (0.457, significant at the 10% level) remain the primary drivers, while inflation (0.060) and consumer spending (-0.080) add minimal incremental information and are not statistically significant. Hence, a parsimonious specification with

just employment and economic growth appears to capture much of the policy-relevant information in the Beige Book. To put these results in economic terms, the employment coefficient of 0.476 suggests that a one-standard-deviation increase in the employment score is associated with approximately a 48 basis point increase in the federal funds rate, while a similar increase in economic growth corresponds to a 53 basis point increase. These magnitudes are economically meaningful, though it is important to note that these associations may not represent causal effects and could vary across different monetary policy regimes. To further explore the robustness of these findings and examine the impact of policy inertia, Table 8 in the Appendix presents a comprehensive analysis with multiple specifications. This table shows both level and difference specifications, with and without lagged policy rates, confirming that the employment-growth combination remains robust across different model formulations.

To assess the reliability of these coefficient estimates, I test for multicollinearity among the Beige Book components using Variance Inflation Factors (VIFs)[9]. Table 3 presents the results. While inflation shows no multicollinearity concern (VIF = 1.56), all other variables exhibit only mild concerns: employment (VIF = 3.11), economic growth (VIF = 3.48), and consumer spending (VIF = 2.64). Following Hair Jr et al. (1995), VIF values above 5 indicate that the variable shares more than 80% of its variance with other regressors. Since all VIFs are below this threshold, multicollinearity is not a serious concern in this specification, and the aggregate weighted specification remains statistically robust.

[9]VIFs measure how much the variance of a coefficient increases due to collinearity with other regressors, calculated as $\text{VIF}_i = 1/(1 - R_i^2)$, where $R_i^2$ is obtained from regressing variable $i$ on all other explanatory variables.

Table 3: Variance Inflation Factors for Beige Book Components

| Variable | VIF | Interpretation |
|---|---|---|
| Inflation | 1.56 | None |
| Employment | 3.11* | Mild |
| Economic Growth | 3.48* | Mild |
| Consumer Spending | 2.64* | Mild |

Note: VIF $>$ 10 indicates severe multicollinearity (***), VIF $>$ 5 indicates moderate concern (**), VIF $>$ 2.5 indicates mild concern (*). Analysis based on Hair et al. (2010) and O'Brien (2007) recommendations.

The multicollinearity analysis, while reassuring, masks an important insight about the individual versus joint contributions of the Beige Book components. As shown in Table 1, when estimated individually (columns 2-5), all fmy components show strong predictive power, with employment and economic growth achieving the highest $R^2$ values (0.124 and 0.123, respectively). However, in the full multivariate specification (column 6), their coefficients are attenuated due to shared variation, and consumer spending even changes sign to negative. This pattern explains why the parsimonious two-variable model performs nearly as well as the full specification—the additional variables primarily capture information already contained in employment and growth measures.

The aggregate Beige Book score warrants particular attention. Column 1 of Table 1 shows that a single weighted composite achieves an $R^2$ of 0.130, capturing 93% of the full model's explanatory power (0.130/0.140). This aggregate score performs comparably to the best individual component (employment) while providing a more parsimonious representation of the Beige Book's policy-relevant information. The fact that the aggregation process maintains predictive power relative to individual components is consistent with the hypothesis that the Federal Reserve responds to an overall economic assessment, though this interpretation should be viewed cautiously given the potential for mechanical aggregation effects. This finding could have practical implications: market participants and researchers might track monetary policy responses using a single aggregate Beige

Book score, though such an approach would necessarily abstract from the granular information contained in individual components.

The importance of the Federal Reserve's communication emphasis becomes apparent when comparing weighted and unweighted specifications. While unweighted components capture raw sentiment, the weighted versions—which incorporate the Fed's emphasis—consistently provide greater explanatory power. This pattern suggests that not only what the Federal Reserve communicates matters, but also the extent to which it emphasizes particular topics, highlighting the informational value embedded in the Fed's choice of emphasis across economic themes.

These findings collectively indicate that Beige Book scores contain substantial information relevant to monetary policy decisions. The progression from negligible explanatory power (policy inertia alone) to substantial predictive ability (with Beige Book information) suggests that the Federal Reserve's qualitative assessments, when systematically quantified, may provide a valuable input to the policy reaction function. The robustness of these relationships across specifications, combined with the inclusion of policy inertia controls, is consistent with the view that the Beige Book provides genuinely new information, rather than merely reflecting publicly available economic data. This interpretation aligns with the Beige Book's role as a smyce of anecdotal, real-time information from Federal Reserve districts, which may not be immediately captured in conventional economic statistics.

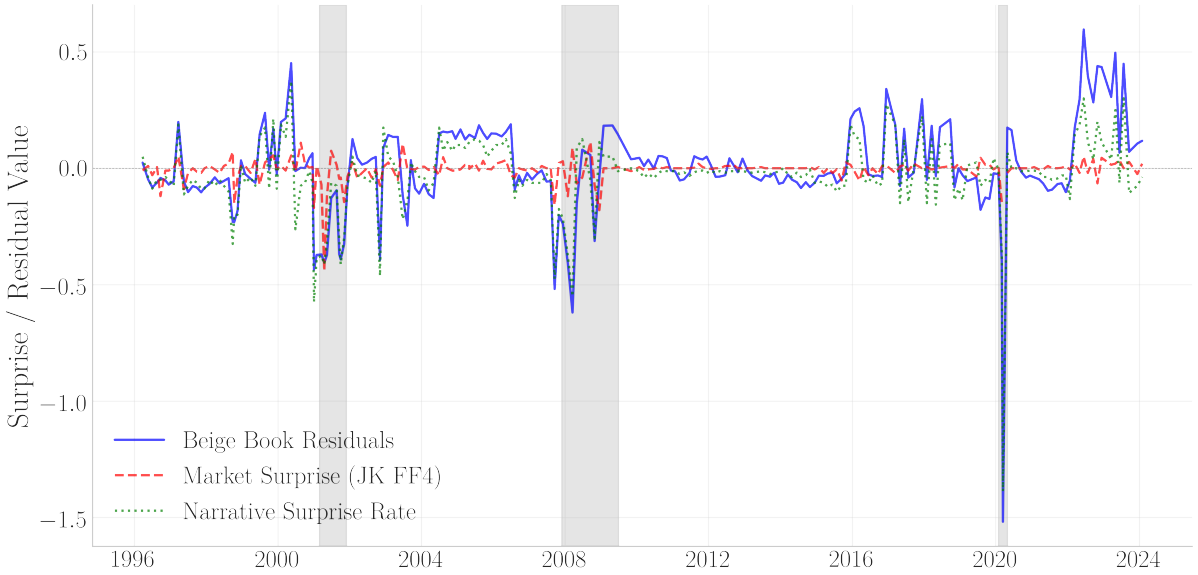
## 3.2 Are Beige Book Surprises Coherent with Narrative and Market Surprises?

To further understand the informational content of the Beige Book, I analyze the residuals from the main policy regression (column 3 of Table 2). These residuals represent the portion of the FOMC's decision that is not explained by the Beige Book scores or policy inertia—a "Beige Book surprise." I compare these residuals to standard market-based monetary policy surprises, specifically the FF4 surprise from (Jarociński & Karadi, 2020).

Figure 19 plots the time series of the Beige Book residuals against the market surprises.

A visual inspection reveals that while the two series sometimes move together, there are significant periods of divergence. This suggests that the information contained in the Beige Book is not perfectly aligned with the information priced into financial markets.

Figure 19: Beige Book Policy Regression Residuals vs. Market Surprises




Note: The figure compares residuals from the Beige Book policy regression with the FF4 market surprise from (Jarociński & Karadi, 2020) and my narrative surprise measure. The path to the figure is generated automatically based on the version and model specified in main.tex.

Table 4 presents the correlation between the market surprises and the Beige Book residuals, and the narrative surprises produced by the Surprise Snipper and the Beige Book residuals, quantifying the relationships depicted in the figure. It is important to note the high correlation between the Beige Book residuals and the narrative surprises as opposed to the low correlation with all the market surprises. This is a confirmation that SS's surprise measure is coherent with BBC's unexpected component. This apparently meaningless correlation shows the robustness of the infrastructure that I have created to process different smyces of information as part of a single analysis. If this correlation were not present, there would be a problem in the pipeline, as the agent computing the surprises narratively (SS) would have been ontologically disconnected from the agent processing the information (BBC).

Table 4: Correlation of Beige Book Residuals with Policy Surprises

| Surprise Measure | Correlation with Residuals |
|---|---|
| FF4 | 0.371 |
| MP1 | 0.395 |
| ED1 | 0.327 |
| ED4 | 0.293 |
| Narrative Surprise | 0.924 |

Note: The table shows the Pearson correlation between the residuals of the main Beige Book policy regression and various measures of monetary policy surprises. Market surprises are from Jarociński and Karadi (2020).

This finding reinforces the value of the Beige Book as a distinct valuable smyce of information for monetary policy analysis. Its contents appear to be orthogonal to, rather than redundant with, other common measures of policy expectations and surprises. At the same time, the fact that the narrative surprise is coherent with the Beige Book residuals suggests that the narrative surprise is a good proxy for the true, unanticipated policy shock.

## 3.3 Can Surprises Be Predicted?

To assess the predictability of the narrative and market-based surprise measures, I regress each surprise series on a set of standard macroeconomic and financial predictors from Bauer and Swanson (2023b). These include measures of economic activity (nonfarm payrolls), financial market performance (S&P 500 returns), the term spread, commodity prices, and Treasury market skewness taken from Bauer and Swanson (2023b) dataset. If surprises are truly unanticipated, they should not be predictable using publicly available information prior to the FOMC meeting.

Table 5 presents the results. The regressions show that while the predictors have some statistically significant explanatory power, the surprise measures remain largely unpredictable. For my three narrative measures, the $R^2$ values are notably low: 10.4%

for the baseline surprise rate, 8.0% for the unpredictable component, and 9.0% for the contextual salience measure. The unpredictable surprise component shows the lowest predictability, which is reassuring as it represents the surprise after removing any predictable elements. Market-based measures show comparable predictability, with $R^2$ values ranging from 8.7% to 16.7%, indicating that the vast majority of the variation in all surprise measures is not explained by this standard set of public information. Notably, a closer inspection of the coefficients reveals that much of this modest predictability stems from financial market variables, such as stock returns and bond market indicators. This is an expected and reassuring result, as the narrative surprises are constructed to be deliberately market-agnostic, relying exclusively on the Federal Reserve's own communications. The fact that their predictability is driven by information they are designed to ignore further validates their construction as pure measures of the Fed's information set.

Table 5: Predictability of FOMC Meeting Surprises

| | Narrative Measures | | | Market-Based Measures | | | |
|---|---|---|---|---|---|---|---|
| Variable | Rate | Unpred. | Salience | FF4 | MP1 | ED1 | ED4 |
| Commodity Index (3m) | 0.010 | 0.004 | 0.004 | 0.004 | 0.001 | 0.003 | 0.011* |
| | (0.010) | (0.006) | (0.006) | (0.004) | (0.004) | (0.005) | (0.005) |
| Nonfarm Payrolls (12m) | 0.010 | 0.009* | 0.008 | 0.002 | -0.000 | 0.003 | 0.011*** |
| | (0.009) | (0.005) | (0.005) | (0.002) | (0.002) | (0.002) | (0.004) |
| Term Spread (3m) | -0.023** | -0.006 | -0.009 | -0.010** | -0.009** | -0.010* | -0.008 |
| | (0.009) | (0.006) | (0.007) | (0.005) | (0.004) | (0.006) | (0.006) |
| S&P 500 (3m) | 0.024* | 0.014* | 0.017** | 0.011** | 0.012** | 0.010** | 0.013** |
| | (0.013) | (0.007) | (0.008) | (0.004) | (0.006) | (0.004) | (0.005) |
| Treasury Skewness | 0.033** | 0.024* | 0.026** | 0.007*** | 0.010** | 0.008** | 0.011*** |
| | (0.014) | (0.013) | (0.013) | (0.003) | (0.005) | (0.003) | (0.003) |
| $R^2$ | 0.104 | 0.080 | 0.090 | 0.150 | 0.092 | 0.117 | 0.167 |
| Observations | 239 | 239 | 239 | 233 | 233 | 234 | 234 |

*Note:* This table presents predictability regressions of various surprise measures on a set of macroeconomic and financial predictors from Bauer and Swanson (2023b). The predictors include Nonfarm Payrolls (12m), S&P 500 (3m), Term Spread (3m), Commodity Index (3m), and Treasury Skewness. Standard errors are HAC-robust and are reported in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels.

To examine the temporal stability of these predictability patterns, I conduct a rolling window analysis using 60-meeting windows over the sample period. This analysis tests

whether surprise measures can be consistently predicted using the same set of macroeconomic variables across different monetary policy regimes. The rolling estimation produces 213 overlapping windows spanning from 1996 to 2024, capturing periods of conventional monetary policy, the zero lower bound era, and the recent tightening cycle. A good surprise measure should exhibit low and stable predictability over time.

Table 6 summarizes the results. my narrative surprise measures demonstrate strong performance across the board. The baseline surprise rate shows a mean $R^2$ of 0.116 across the 213 rolling windows, indicating that on average only about 11.6% of its variance is predictable by standard public information. Notably, the unpredictable surprise component performs even better with a mean $R^2$ of just 0.080, confirming that removing the predictable component successfully isolates the truly unexpected portion of policy decisions. The contextual salience measure, which weights surprises by their significance and confidence, achieves a mean $R^2$ of 0.094, also demonstrating low predictability.

When compared to established market-based measures, all three narrative measures are competitive. The unpredictable surprise component actually shows the lowest average predictability among all measures except MP1 (0.085). The baseline surprise rate's predictability (0.116) is in the same ballpark as FF4 (0.126) and lower than ED4 (0.143). While MP1 shows the lowest average predictability, it is important to note that some market measures exhibit extreme instability. For instance, MP1's $R^2$ ranges from -0.089 to 0.703, suggesting that while it is on average the least predictable, it suffers from periods of significant contamination. The narrative measures show more stable performance with less extreme variation.

Overall, these findings provide a strong validation for the narrative-based approach. All three variants of the narrative surprise as unpredictable as market-based measures, with the unpredictable component performing particularly well. This result is particularly noteworthy considering the informational disadvantage of the narrative measures. The underlying prior is updated at the time of the Beige Book's release—typically two weeks before an FOMC meeting—and incorporates no high-frequency market data. In contrast, market-based surprises capture information right up to the meeting itself. The fact that

text-based measures, using slightly dated information, can achieve levels of unpredictability comparable to or better than high-frequency financial instruments underscores their value as robust and independent smyces of information for identifying monetary policy shocks.

Table 6: Rolling Window Predictability Analysis Summary

| Surprise Measure | Windows | Mean | Std. Dev. | Min | Max |
|---|---|---|---|---|---|
| Narrative (Rate) | 213 | 0.116 | 0.132 | -0.078 | 0.548 |
| Narrative (Unpred.) | 213 | 0.080 | 0.131 | -0.082 | 0.667 |
| Narrative (Salience) | 213 | 0.094 | 0.129 | -0.086 | 0.520 |
| FF4 (Jarocinski-Karadi) | 213 | 0.126 | 0.124 | -0.053 | 0.641 |
| MP1 (Jarocinski-Karadi) | 213 | 0.085 | 0.132 | -0.089 | 0.703 |
| ED1 (Eurodollar 1Q) | 213 | 0.121 | 0.096 | -0.032 | 0.626 |
| ED4 (Eurodollar 1Y) | 213 | 0.143 | 0.109 | -0.064 | 0.346 |

Notes: This table reports summary statistics from rolling window regressions using three narrative measures (rate, unpredictable component, and contextual salience) and market-based surprises from Jarociński and Karadi (2020). Each 60-meeting window regresses the surprise measure on standard macroeconomic and financial predictors from Bauer and Swanson (2023b). Lower and more stable $R^2$ values indicate better performance as genuine policy shocks. Sample spans 1996-2024 with 213 rolling windows.

## 3.4 Are Narrative Surprises Different from Market Surprises?

To assess whether narrative-based monetary policy surprises may differ from traditional market-based measures, I start from a standard decomposition of policy decisions:

$$r_t^{actual} = r_t^{exp} + u_t, \tag{1}$$

where $r_t^{exp}$ is the ex-ante expectation and $u_t$ is the unexpected component. Any surprise measure attempts to estimate this unexpected component with some measurement error:

$$\hat{u}_t = u_t + \varepsilon_t, \tag{2}$$

where $\varepsilon_t$ represents measurement error. When we regress actual policy decisions on surprise measures, the relationship between my measure and realized policy changes requires careful interpretation. I estimate a series of regressions of FOMC policy decisions (rate

changes) on various surprise measures:

$$\Delta i_t = \alpha + \beta \text{Surprise}_t + \varepsilon_t \tag{3}$$

where $\Delta i_t$ denotes the change in the federal funds rate target at meeting $t$, and $\text{Surprise}_t$ represents either the narrative or market-based surprise measure. The coefficient $\beta$ and $R^2$ from these regressions tell us about the relationship between my measure and realized policy changes, but their interpretation depends crucially on what each measure is designed to capture. Table 7 presents the results, which are divided into two panels.

Panel A of Table 7 shows the results from individual regressions. The narrative surprise measures exhibit strong relationships with policy decisions: the baseline surprise rate yields a coefficient of 1.045 with an $R^2$ of 0.615, the unpredictable component shows a coefficient of 1.318 with an $R^2$ of 0.463, and the contextual salience measure yields 1.263 with an $R^2$ of 0.493. Market-based measures (FF4, MP1, ED1, ED4) from (Jarociński & Karadi, 2020) also load significantly but with notably lower $R^2$ values (0.150 to 0.171). These striking differences in explanatory power can be interpreted in several ways:

**Interpretation 1: Measurement Quality.** The higher $R^2$ for narrative measures could indicate they capture policy variation more accurately. Since they are constructed directly from Fed communications, they may better reflect the information set and reasoning process underlying FOMC decisions.

**Interpretation 2: Information Content.** Alternatively, the difference may reflect what each measure is designed to capture. Market-based surprises extract only the unexpected component from high-frequency price movements in narrow windows around FOMC announcements. Narrative measures, constructed from Fed documents over longer periods, may inherently contain both surprising and systematic components of policy decisions.

**Interpretation 3: Timing and Scope.** The measures differ fundamentally in their information smyces and timing. Market surprises capture investor reactions in windows of 30 minutes around announcements, designed to isolate pure shocks. Narrative surprises synthesize Fed communications from the weeks leading up to decisions, potentially

Table 7: Monetary Policy Decision Regressions on Surprise Measures

| *Panel A: Individual Surprises - Narrative Measures* | | | | |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| *Narrative Surprise* | 1.045*** | | | |
| | (0.052) | | | |
| *Narrative (Rate)* | | 1.045*** | | |
| | | (0.052) | | |
| *Narrative (Unpred.)* | | | 1.318*** | |
| | | | (0.089) | |
| *Narrative (Salience)* | | | | 1.263*** |
| | | | | (0.080) |
| *Constant* | 0.037*** | 0.037*** | 0.024** | 0.029*** |
| | (0.009) | (0.009) | (0.010) | (0.010) |
| $R^2$ | 0.615 | 0.615 | 0.463 | 0.493 |
| Observations | 256 | 256 | 256 | 256 |

| *Panel A (continued): Individual Surprises - Market-Based Measures* | | | | |
|---|---|---|---|---|
| | (5) | (6) | (7) | (8) |
| *FF4* | 1.858*** | | | |
| | (0.278) | | | |
| *MP1* | | 1.426*** | | |
| | | (0.231) | | |
| *ED1* | | | 1.711*** | |
| | | | (0.261) | |
| *ED4* | | | | 1.359*** |
| | | | | (0.220) |
| *Constant* | 0.024* | 0.026* | 0.022 | 0.019 |
| | (0.014) | (0.014) | (0.015) | (0.015) |
| $R^2$ | 0.171 | 0.150 | 0.166 | 0.150 |
| Observations | 218 | 218 | 218 | 219 |

| *Panel B: Pooled Regressions* | | | | | | |
|---|---|---|---|---|---|---|
| | (9) | (10) | (11) | (12) | (13) | (14) |
| *Narrative (Rate)* | 1.144*** | | | 1.193*** | | |
| | (0.070) | | | (0.072) | | |
| *Narrative (Unpred.)* | | 1.725*** | | | 1.961*** | |
| | | (0.145) | | | (0.152) | |
| *Narrative (Salience)* | | | 1.769*** | | | 1.996*** |
| | | | (0.131) | | | (0.139) |
| *FF4* | 0.648*** | 0.695*** | 0.586** | 0.130 | 0.721 | 0.847 |
| | (0.201) | (0.238) | (0.226) | (0.629) | (0.718) | (0.683) |
| *MP1* | | | | -0.956*** | -1.588*** | -1.643*** |
| | | | | (0.349) | (0.421) | (0.397) |
| *ED1* | | | | 1.315*** | 1.281** | 1.189** |
| | | | | (0.474) | (0.541) | (0.512) |
| *ED4* | | | | 0.073 | 0.089 | 0.018 |
| | | | | (0.235) | (0.269) | (0.256) |
| *Constant* | 0.049*** | 0.037*** | 0.045*** | 0.050*** | 0.036*** | 0.044*** |
| | (0.010) | (0.011) | (0.010) | (0.009) | (0.010) | (0.010) |
| $R^2$ | 0.629 | 0.500 | 0.551 | 0.651 | 0.548 | 0.592 |
| Observations | 218 | 218 | 218 | 217 | 217 | 217 |

*Note:* This table presents regressions of FOMC policy decisions (rate changes) on various surprise measures. Panel A reports regressions on individual surprise measures. Panel B reports regressions on pooled models. Standard errors are in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels. Time window: 1996-01 to 2023-12.

capturing the Fed's broader decision-making process.

Panel B presents pooled regressions including both narrative and market-based surprises. When combined, narrative surprises maintain large and significant coefficients (0.999 to 1.961 across specifications), while market-based measures often become insignificant or show reduced coefficients. The $R^2$ values for pooled models range from 0.480 to 0.651, with the highest explanatory power achieved when combining all measures. This pattern reveals several important insights: First, **narrative measures dominate in explaining policy variation**, consistent with their construction from the same Fed communications that inform policy decisions. The coefficients near or above unity suggest these measures align closely with actual policy changes. Second, **market measures add limited incremental explanatory power** when combined with narrative surprises. This could indicate either that narrative measures already capture the relevant information, or that market measures isolate a different, purer notion of surprise that represents a smaller share of total policy variation. Third, **the combined $R^2$ remains well below 1.0**, indicating that even together, these measures do not fully explain policy decisions. This is reassuring, as it suggests neither approach is mechanically constructed to match outcomes.

## 3.5 Interpretation and Implications

Rather than viewing high $R^2$ as unambiguous validation of superiority, these results highlight fundamental differences between narrative and market-based approaches to measuring monetary policy surprises. The two measures are answering different questions with different information sets.

**Market-based surprises** are explicitly designed to isolate truly unexpected shocks by focusing on high-frequency price movements in narrow windows. They measure the *last-minute pricing error* relative to a fully-informed market consensus seconds before the announcement. Their lower $R^2$ (0.150-0.171) may actually reflect successful filtering of predictable components (although not completely, as Bauer & Swanson, 2023b show). By design, they are supposed to capture only the residual that moves asset prices after mar-

kets have processed all available information, making them ideal for clean identification of exogenous policy shocks.

**Narrative surprises** measure something fundamentally different: the *deliberative policy evolution* from the Fed's own documented baseline 2-3 weeks prior. My methodology deliberately stops information collection at the Beige Book release, aligning with the Fed's blackout period when Committee members cease public communications. The higher $R^2$ (0.615) reveals that most policy variation can be explained by this inter-meeting evolution of Fed thinking—the increment between what the Fed signaled weeks earlier and its final decision.

This temporal separation is a key methodological strength. By forming priors before the blackout period, we ensure:

1. **Exogeneity**: No contamination from last-minute data releases or market movements
2. **Institutional alignment**: I capture the Committee's information set as it enters deliberations
3. **Cleaner identification**: For event studies, my surprises are predetermined relative to announcement-day asset price movements

The striking difference in explanatory power—narrative surprises explain 3-4 times more policy variation than market surprises—suggests an important insight: the bulk of monetary policy decisions are telegraphed through the Fed's formal communications well before meetings. The high-frequency "surprise" that markets trade on represents only a small residual after this deliberative process. This interpretation reframes my contribution. I am not claiming to build a better market timing tool. Instead, I decompose monetary policy information flow into two components:

$$\text{Total Policy Information} = \text{Deliberative Evolution} + \text{High-Frequency Shock} \quad (4)$$

my results suggest the first component dominates, accounting for over 60% of policy variation. This has important implications for understanding Fed communication effectiveness, the formation of expectations, and the true smyces of monetary policy un-

certainty. For researchers, this decomposition offers complementary tools: market measures remain superior for identifying pure exogenous shocks needed for causal inference, while narrative measures illuminate how the Fed's collective thinking evolves through its communication cycle. The availability of both enriches my understanding of monetary policy, demonstrating how Large Language Models can unlock the informational content of central bank communications that was previously difficult to systematically analyze.

# 4 Conclusion

This paper decomposes monetary policy surprises by systematically analyzing the Federal Reserve's communication cycle, revealing that over 60% of policy variation is embedded in official documents weeks before announcements. Building on the narrative tradition of Romer and Romer (2004), I develop a scalable, real-time approach that leverages the Fed's modern transparency regime—established when it began announcing decisions in 1994. By processing the complete set of regular Fed documents during the critical "blackout period" (the 2-3 weeks before each meeting when officials cease public commentary), I show that this deliberative period contains the bulk of policy information. What markets perceive as last-minute surprises largely reflects the resolution of uncertainty about a policy path already signaled through formal channels.

The empirical findings reveal a fundamental decomposition of monetary policy information. While market-based measures capture high-frequency shocks in narrow windows around announcements, explaining 15-17% of policy changes, narrative surprises extracted from Fed documents explain 61.5%. This stark difference does not indicate that one measure is superior to the other. Instead, it reveals that they capture distinct components: narrative surprises measure the unexpected evolution of Fed thinking from its documented baseline during the blackout period, while market surprises isolate the last-minute pricing adjustments as information crystallizes at announcement time.

Three key insights emerge from this decomposition. First, monetary policy is more systematic and transparent than previously understood—the majority of policy decisions

are telegraphed through formal Fed communications weeks in advance. This finding contrasts sharply with Bauer and Swanson (2023b, 2023c), who find limited predictability in market-based surprises, suggesting that predictability depends crucially on conditioning on the Fed's own communications rather than market instruments. Second, the "surprise" that moves markets on announcement day represents a relatively small residual after accounting for the deliberative policy evolution. Third, what appears unpredictable through the lens of high-frequency market data becomes largely explicable when viewed through the Fed's institutional communication framework.

The methodological contribution extends beyond monetary economics. This work demonstrates how Large Language Models, when deployed with appropriate temporal structure and institutional context, can unlock the informational content of complex organizational communications. The multi-agent architecture—processing entire documents rather than isolated sentences, maintaining temporal consistency, and synthesizing information across document types—provides a template for analyzing other institutional communication systems where narrative evolution matters.

For monetary policy research, these findings offer complementary identification strategies. Market-based measures remain ideal for studying pure exogenous shocks and their transmission through financial markets. Narrative measures illuminate the systematic component of policy formation, offering insights into how central bank thinking evolves and how effectively it communicates its intentions. Researchers can choose the appropriate measure based on their specific research question—or use both to separate deliberative from shock components.

The results also carry practical implications. For market participants, tracking Fed communications during the pre-meeting period may provide valuable signals about likely policy outcomes. For policymakers, the findings validate the effectiveness of Fed transparency efforts—the high explanatory power of narrative surprises suggests that the Fed successfully communicates its evolving views through its formal channels. However, the persistent 40% of policy variation that remains unexplained by narrative measures indicates room for further enhancing communication effectiveness.

Several avenues for future research emerge. First, extending this framework to other central banks would test whether the decomposition between deliberative and shock components is universal or specific to the Federal Reserve's communication style. Second, examining how this decomposition varies across different monetary policy regimes—conventional versus unconventional policy periods—could reveal how communication effectiveness changes with policy tools. Third, investigating which specific elements of Fed communications drive the narrative surprises could help central banks optimize their communication strategies.

More broadly, the multi-agent framework itself provides a template for novel research designs in economics. A particularly promising extension would leverage the architecture as a "laboratory of surveys" in the spirit of J. L. Bybee (2023) and L. Bybee (2023). By introducing controlled variations in the system prompts, each agent could represent a heterogeneous economic observer with distinct interpretive biases or areas of focus. Monte Carlo simulations across these synthetic survey respondents would generate distributions of expectations, allowing researchers to study how different information processing approaches lead to divergent views, which aspects of Fed communications generate the most disagreement, and how consensus forms—or fails to form—across observer types. This approach could provide new insights into expectation formation processes and offer a controlled environment for testing theories about how economic agents interpret central bank communications.

The framework's applicability extends beyond monetary policy to any domain where institutional communications precede market-moving announcements. Earnings call surprises represent a natural application—comparing narrative surprises extracted from pre-call documents (10-Ks, 10-Qs, analyst reports) against post-announcement market reactions could reveal whether the deliberative versus shock decomposition holds in corporate settings. Similarly, the approach could analyze regulatory announcements, political communications, or other central banks' policy decisions. Each application would test the generality of my finding that most "surprise" reflects the evolution of institutional thinking rather than genuine last-minute shocks, potentially revealing universal patterns in

how organizations communicate and markets process information.

The integration of Large Language Models into economic research is still in its early stages. This paper demonstrates their potential when combined with careful institutional knowledge and appropriate research design. As these tools continue to evolve, they will likely enable new insights into how organizations communicate, how information flows through markets, and how economic agents form and update their expectations. The key lies not in the technology itself, but in understanding how to deploy it in ways that respect the temporal and institutional structure of economic phenomena.

# References


Adrian, T., Crump, R. K., & Moench, E. (2013). Pricing the term structure with linear regressions. *Journal of Financial Economics*, *110*(1), 110–138.

Ahrens, M., Erdemlioglu, D., McMahon, M., Neely, C. J., & Yang, X. (2024). Mind your language: Market responses to central bank speeches. *Journal of Econometrics*, 105921.

Ahrens, M., & McMahon, M. (2021). Extracting economic signals from central bank speeches. *Proceedings of the Third Workshop on Economics and Natural Language Processing*.

Aksit, D. (2020). Unconventional monetary policy surprises: Delphic or odyssean? *Available at SSRN 3602291*.

Andersson, M., Dillén, H., & Sellin, P. (2006). Monetary policy signaling and movements in the term structure of interest rates. *Journal of Monetary Economics*, *53*(8), 1815–1855.

Andrade, P., & Ferroni, F. (2021). Delphic and odyssean monetary policy shocks: Evidence from the euro area. *Journal of Monetary Economics*, *117*, 816–832.

Aruoba, S. B., & Drechsel, T. (2024). *Identifying Monetary Policy Shocks: A Natural Language Approach* (tech. rep.). National Bureau of Economic Research.

Balke, N. S., Fulmer, M., & Zhang, R. (2017). Incorporating the beige book into a quantitative index of economic activity. *Journal of Forecasting*, *36*(5), 497–514.

Bauer, M. D., & Swanson, E. T. (2023a). An alternative explanation for the "fed information effect". *American Economic Review*, *113*(3), 664–700.

Bauer, M. D., & Swanson, E. T. (2023b). An Alternative Explanation for the "Fed Information Effect". *American Economic Review*, *113*(3), 664–700.

Bauer, M. D., & Swanson, E. T. (2023c). A reassessment of monetary policy surprises and high-frequency identification. *NBER Macroeconomics Annual*, *37*(1), 87–155.

Bauer, M. D., & Swanson, E. T. (2023d). A reassessment of monetary policy surprises and high-frequency identification. *NBER Macroeconomics Annual*, *37*(1), 87–155.

Bernanke, B. S. (2005). The logic of monetary policy. *Vital Speeches of the Day*, *71*(6), 165.

Bernanke, B. S., & Mihov, I. (1998). Measuring monetary policy. *The quarterly journal of economics*, *113*(3), 869–902.

Bernanke, B. S., Reinhart, V. R., & Sack, B. P. (2004). *Monetary policy alternatives at the zero bound: An empirical assessment* (tech. rep.). Brookings Institution. https://www.brookings.edu/wp-content/uploads/2004/01/20040105.pdf

Blinder, A. S., Ehrmann, M., Fratzscher, M., De Haan, J., & Jansen, D.-J. (2008). Central Bank Communiqué and Monetary Policy: A Survey of Theory and Evidence. *Journal of economic literature*, *46*(4), 910–945.

Bordalo, P., Gennaioli, N., Ma, Y., & Shleifer, A. (2020). Overreaction in macroeconomic expectations. *American Economic Review*, *110*(9), 2748–2782.

Bybee, J. L. (2023). The ghost in the machine: Generating beliefs with large language models. *arXiv preprint arXiv:2305.02823*.

Bybee, L. (2023). Surveying generative ai's economic expectations. *arXiv preprint arXiv:2305.02823*.

Caballero, R. J., & Simsek, A. (2022). Monetary policy with opinionated markets. *American Economic Review*, *112*(7), 2353–2392. https://doi.org/10.1257/aer.20210271

Campbell, J. R., Evans, C. L., Fisher, J. D., Justiniano, A., Calomiris, C. W., & Woodford, M. (2012). Macroeconomic effects of federal reserve forward guidance [with comments and discussion]. *Brookings papers on economic activity*, 1–80.

Campbell, J. Y., & Shiller, R. J. (1988). The dividend-price ratio and expectations of future dividends and discount factors. *The review of financial studies*, *1*(3), 195–228.

Christiano, L. J., Eichenbaum, M., & Evans, C. L. (1999). Monetary policy shocks: What have we learned and to what end? *Handbook of macroeconomics*, *1*, 65–148.

Cieslak, A. (2018). Short-rate expectations and unexpected returns in treasury bonds. *The Review of Financial Studies*, *31*(9), 3265–3306.

Cieslak, A., McMahon, M., & Pang, H. (2024a). Did i make myself clear? the fed and the market in the post-2020 framework period. *Unpublished (August)*.

Cieslak, A., McMahon, M., & Pang, H. (2024b). Did i make myself clear? the fed and the market in the post-2020 framework period. *Unpublished (August)*.

Cieslak, A., & Schrimpf, A. (2019). Non-monetary news in central bank communication. *Journal of International Economics*, *118*, 293–315.

Cieslak, A., & Vissing-Jorgensen, A. (2021). The economics of the fed put. *The Review of Financial Studies*, *34*(9), 4045–4089.

Clarida, R., Gali, J., & Gertler, M. (1999). The Science of Monetary Policy: A New Keynesian Perspective. *Journal of economic literature*, *37*(4), 1661–1707.

Cloyne, J. S., Jorda, Ò., & Taylor, A. M. (2020). *Decomposing the fiscal multiplier* (NBER Working Paper No. 26939). National Bureau of Economic Research. https://www.nber.org/papers/w26939

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of finance*, *66*(4), 1047–1108.

De Fiore, F., Maurin, A., Mijakovic, A., & Sandri, D. (2024). *Monetary policy in the news: Communication pass-through and inflation expectations*. Bank for International Settlements, Monetary; Economic Department.

Du, Z., Zeng, A., Dong, Y., & Tang, J. (2024). Understanding emergent abilities of language models from the loss perspective. *arXiv preprint arXiv:2403.15796*.

Feng, S., Ding, W., Liu, A., Wang, Z., Shi, W., Wang, Y., Shen, Z., Han, X., Lang, H., Lee, C.-Y., et al. (2025). When one llm drools, multi-llm collaboration rules. *arXiv preprint arXiv:2502.04506*.

Filippou, I., Garciga, C., Mitchell, J., & Nguyen, M. T. (2024a). Regional economic sentiment: Constructing quantitative estimates from the beige book and testing their ability to forecast recessions. *Economic Commentary*, (2024-08).

Filippou, I., Garciga, C., Mitchell, J., & Nguyen, M. T. (2024b). Regional Economic Sentiment: Constructing Quantitative Estimates from the Beige Book and Testing their ability to Forecast Recessions. *Economic Commentary*, *2024*(08).

Fujiwara, M., Suimon, Y., & Nakagawa, K. (2023). Treasury yield spread prediction with sentiments of beige book and macroeconomic data. *2023 14th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, 337–342.

Gambacorta, L., Kwon, B., Park, T., Patelli, P., & Zhu, S. (2024). *CB-LLMs: Language Models for Central Banking*. Bank for International Settlements, Monetary; Economic Department.

Gertler, M., & Karadi, P. (2015). Monetary Policy Surprises, Credit Costs, and Economic Activity. *American Economic Journal: Macroeconomics*, *7*(1), 44–76. https://doi.org/10.1257/mac.20130329

Gürkaynak, R. S., Sack, B., & Swanson, E. (2005). The sensitivity of long-term interest rates to economic news: Evidence and implications for macroeconomic models. *American Economic Review*, *95*(1), 425–436. https://doi.org/10.1257/0002828053828443

Hair Jr, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1995). *Multivariate data analysis with readings*. Prentice-Hall, Inc.

Hansen, A. L., & Kazinnik, S. (2023). Can chatgpt decipher fedspeak. *Available at SSRN*.

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

Hanson, S. G., & Stein, J. C. (2012). Monetary policy and long-term real rates. *Finance and Economics Discussion Series*. https://doi.org/10.17016/FEDS.2012.69

Jarociński, M. (2024). Estimating the fed's unconventional policy shocks. *Journal of Monetary Economics*, *144*, 103548.

Jarociński, M., & Karadi, P. (2020). Deconstructing monetary policy surprises—the role of information shocks. *American Economic Journal: Macroeconomics*, *12*(2), 1–43. https://doi.org/10.1257/mac.20180082

Jarociński, M., & Karadi, P. (2025). Disentangling monetary policy, central bank information, and fed response to news shocks. *Unpublished (February)*.

Jiang, H. (2023). A latent space theory for emergent abilities in large language models. arxiv 2023. *arXiv preprint arXiv:2304.09960*.

Jordà, Ò. (2005). Estimation and inference of impulse responses by local projections. *American economic review*, *95*(1), 161–182.

Jordà, Ò., & Taylor, A. M. (2025). Local projections. *Journal of Economic Literature*, *63*(1), 59–110.

Kim, A., Muhn, M., & Nikolaev, V. (2024). Financial statement analysis with large language models. *arXiv preprint arXiv:2407.17866*.

Kojima, T., Gu, S., Reid, Y., & Matsuo, Y. (2022). Large language models are zero-shot reasoners.

Kuttner, K. N. (2001). Monetary policy surprises and interest rates: Evidence from the fed funds futures market. *Journal of monetary economics*, *47*(3), 523–544.

Lopez-Lira, A., & Tang, Y. (2023). Can ChatGPT Predict Stock Price Movements? Return Predictability and Large Language Models. *arXiv preprint arXiv:2304.07619*.

McMahon, M., Schipke, A., & Xiang, L. (2019). Monetary policy communication: Frameworks and market impact. *The Future of China's Bond Market*, *295*.

Mertens, K., & Ravn, M. O. (2013). The dynamic effects of personal and corporate income tax changes in the united states. *American Economic Review*, *103*(4), 1212–1247. https://doi.org/10.1257/aer.103.4.1212

Miranda-Agrippino, S., & Ricco, G. (2021). The transmission of monetary policy shocks. *American Economic Journal: Macroeconomics*, *13*(3), 74–107.

Nakamura, E., & Steinsson, J. (2018). High-frequency identification of monetary non-neutrality: The information effect. *The Quarterly Journal of Economics*, *133*(3), 1283–1330.

of Governors of the Federal Reserve System, B. (2025, March). *The Beige Book: Summary of Commentary on Current Economic Conditions by Federal Reserve District, february 2025* (Beige Book). Board of Governors of the Federal Reserve System. https://www.federalreserve.gov/monetarypolicy/files/BeigeBook_20250305.pdf

Peskoff, D., Visokay, A., Schulhoff, S., Wachspress, B., Blinder, A., & Stewart, B. M. (2024). Gpt deciphering fedspeak: Quantifying dissent among hawks and doves. *arXiv preprint arXiv:2407.19110*.

Pfeifer, M., & Marohl, V. P. (2023). Centralbankroberta: A fine-tuned large language model for central bank communications. *The Journal of Finance and Data Science*, *9*, 100114.

Poole, W. (2001). Expectations. *Federal Reserve Bank of St. Louis Review*, *83*(March/April 2001).

Ricco, G., & Savini, E. (2025). Decomposing Monetary Policy Surprises: Shock, Information, and Policy Rule Revision. *Unpublished (March)*.

Romer, C. D., & Romer, D. H. (1989). Does monetary policy matter? a new test in the spirit of friedman and schwartz. *NBER Macroeconomics Annual*, *4*, 121–184. https://doi.org/10.1086/654119

Romer, C. D., & Romer, D. H. (2000). Federal reserve information and the behavior of interest rates. *American economic review*, *90*(3), 429–457.

Romer, C. D., & Romer, D. H. (2004). A New Measure of Monetary Shocks: Derivation and Implications. *American Economic Review*, *94*(4), 1055–1084.

Romer, C. D., & Romer, D. H. (2023). *Narrative monetary policy surprises* (tech. rep.) (Working Paper 31507). National Bureau of Economic Research. https://www.nber.org/papers/w31507

Shi, J., & Hollifield, B. (2024). Predictive power of llms in financial markets. *arXiv preprint arXiv:2411.16569*.

Sreedhar, K., & Chilton, L. (2024). Simulating human strategic behavior: Comparing single and multi-agent llms. *arXiv preprint arXiv:2402.08189*.

Stock, J. H., & Watson, M. W. (2012). Disentangling the channels of the 2007-09 recession. *Brookings Papers on Economic Activity*, *43*(1), 81–156. https://doi.org/10.1353/eca.2012.0005

Svensson, L. E. O. (2003). What is wrong with taylor rules? using judgment in monetary policy through targeting rules. *Journal of Economic Literature*, *41*(2), 426–477.

Svensson, L. E., & Woodford, M. (2003). Indicator variables for optimal policy. *Journal of monetary economics*, *50*(3), 691–720.

Swanson, E. T., & Williams, J. C. (2014). Measuring the effect of the zero lower bound on medium- to longer-term interest rates. *American Economic Journal: Macroeconomics*, *6*(2), 1–26. https://doi.org/10.1257/mac.6.2.1

Talebirad, Y., & Nadiri, A. (2023). Multi-agent collaboration: Harnessing the power of intelligent llm agents. *arXiv preprint arXiv:2306.03314*.

Taylor, J. B. (1993). Discretion versus Policy Rules in Practice. *Carnegie-Rochester conference series on public policy*, *39*, 195–214.

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

Villota Miranda, J. (2024). Predicting market reactions to news: An llm-based approach using spanish business articles. *Generative AI in Finance Conference,(John Molson School of Business, Montreal)*.

Wang, L., Menick, J., Neelakantan, A., et al. (2022). Self-consistency improves chain-of-thought reasoning in language models.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al. (2022). Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Zhao, C., Chi, E., & Le, Q. V. (2022). Chain-of-thought prompting elicits reasoning in large language models.

Woodford, M. (1999). Optimal Federal Reserve Balance Sheet. *The Manchester School*, *67*, 1–35.

Wu, Z., Bai, H., Zhang, A., Gu, J., Vydiswaran, V., Jaitly, N., & Zhang, Y. (2024). Divide-or-conquer? which part should you distill your llm? *arXiv preprint arXiv:2402.15000*.

Yang, S., Li, Y., Lam, W., & Cheng, Y. (2025). Multi-llm collaborative search for complex problem solving. *arXiv preprint arXiv:2502.18873*.

# A Additional Tables

## A.1 Comprehensive Beige Book Regression Analysis

Table 8: Comprehensive Beige Book Regression Analysis

| | Without Inertia | | | | With Inertia | | | |
|---|---|---|---|---|---|---|---|---|
| | Inflation | Employment | Econ. Growth | Cons. Spending | Inflation | Employment | Econ. Growth | Cons. Spending |
| **Panel A: Change in Federal Funds Rate ($\Delta i_t$)** | | | | | | | | |
| Individual | 0.147*** | 0.204*** | 0.241*** | 0.199*** | 0.151*** | 0.210*** | 0.239*** | 0.194*** |
| | (0.043) | (0.035) | (0.040) | (0.043) | (0.043) | (0.035) | (0.040) | (0.043) |
| Infl. + Growth | 0.055 | — | 0.220*** | — | 0.057 | — | 0.216*** | — |
| | (0.045) | | (0.043) | | (0.046) | | (0.044) | |
| All variables | 0.019 | 0.094 | 0.158** | -0.007 | 0.015 | 0.114* | 0.146* | -0.020 |
| | (0.052) | (0.061) | (0.074) | (0.068) | (0.052) | (0.063) | (0.075) | (0.068) |
| $i_{t-1}$ | | | | | -0.009 | -0.009 | -0.009 | -0.009 |
| $R^2$ (Individual) | 0.042 | 0.115 | 0.123 | 0.076 | 0.048 | 0.124 | 0.123 | 0.075 |
| $R^2$ (Infl.+Growth) | 0.128 | | | | 0.129 | | | |
| $R^2$ (All) | 0.136 | | | | 0.140 | | | |
| **Panel B: Level of Federal Funds Rate ($i_t$)** | | | | | | | | |
| Individual | 1.000** | 0.796** | -0.485 | -0.652 | 0.479* | 0.359 | 0.113 | 0.054 |
| | (0.427) | (0.361) | (0.414) | (0.434) | (0.264) | (0.222) | (0.255) | (0.269) |
| Infl. + Growth | 1.433*** | — | -1.037** | — | 0.520* | — | -0.095 | — |
| | (0.463) | | (0.444) | | (0.291) | | (0.279) | |
| All variables | 0.379 | 2.547*** | -2.060*** | -1.047 | 0.304 | 0.575 | -0.397 | -0.165 |
| | (0.514) | (0.612) | (0.737) | (0.675) | (0.330) | (0.403) | (0.479) | (0.434) |
| $i_{t-1}$ | | | | | 0.782*** | 0.782*** | 0.782*** | 0.782*** |
| $R^2$ (Individual) | 0.020 | 0.018 | 0.005 | 0.009 | 0.630 | 0.629 | 0.626 | 0.626 |
| $R^2$ (Infl.+Growth) | 0.040 | | | | 0.630 | | | |
| $R^2$ (All) | 0.102 | | | | 0.633 | | | |
| Observations | 265 | | | | | | | |

*Note:* This table presents comprehensive regression results for Beige Book scores. Panel A uses the change in the federal funds rate ($\Delta i_t$) as the dependent variable, while Panel B uses the level ($i_t$). Columns show specifications without and with policy inertia ($i_{t-1}$). Individual regressions include each Beige Book component separately. Standard errors in parentheses. ***, **, and * denote significance at 1%, 5%, and 10% levels.

# B LLM Prompts

This appendix contains the key prompts used throughout the multi-agent system for FOMC communication analysis. Each agent in the pipeline employs specialized prompts designed to extract specific types of information from Federal Reserve documents while maintaining consistency and avoiding look-ahead bias.

## B.1 Summary of Agent Prompts

Table 9 provides an overview of each agent's prompt structure and key methodological elements.

## B.2 Beige Book Calibrator Prompts

The Beige Book Calibrator (BBC) uses the following prompts to systematically extract and quantify economic signals from the Federal Reserve's Beige Book reports.

### B.2.1 System Prompt

Beige Book Calibrator System Prompt

```
You are an economic analyst extracting policy-relevant information from Federal
↪ Reserve Beige Book reports with a focus on identifying potential monetary policy
↪ surprises.

Your task is to read the document and identify economic conditions that would inform
↪ monetary policy decisions. Focus on:

ECONOMIC CONTENT EXTRACTION:
- Find descriptions of economic conditions across key variables
- Note the direction and intensity of economic trends
- Identify language that suggests strengthening or weakening conditions
- Extract specific examples and supporting details

SHOCK DETECTION FOCUS:
- Identify extreme or unprecedented economic conditions
- Look for sharp divergences from recent trends
- Extract signals that contradict consensus expectations
- Note any language suggesting faster/slower changes than anticipated
- Flag conditions that could trigger unexpected policy responses

ANALYSIS APPROACH:
- Let the document content determine the emphasis and scores
```

Table 9: Summary of Multi-Agent System Prompts

| Agent | Key Prompt Elements |
|---|---|
| **Beige Book Calibrator** | • Extracts economic signals from Beige Book reports<br>• Scores variables on -1 to +1 scale (dovish to hawkish)<br>• Weights sum to 1.0 based on document emphasis<br>• Identifies shock indicators for surprise potential<br>• Calculates shock magnitude 0.0 to 1.0 |
| **Expectation Engine** | • Synthesizes Beige Book + Policy Intelligence<br>• Generates probability distribution over outcomes<br>• Maintains conditional unbiasedness<br>• No market data allowed (Fed documents only)<br>• Orthogonality goal: zero correlation with inputs |
| **Policy Extractor** | • Analyzes FOMC minutes for internal deliberations<br>• Classifies forward guidance (Outlook vs Commitment)<br>• Scores guidance strength independently (0.0 to 1.0)<br>• Measures staff vs committee divergence<br>• Identifies shock discovery potential |
| **Surprise Snipper** | • Extracts actual rate decision from statement<br>• Calculates surprise_rate = realized - expected<br>• Decomposes into predictable/unpredictable components<br>• **Key calibration**: 40% prior → score ∼0.2, 25% → ∼0.3, 5% → ∼0.8<br>• Structured assessment guided by explicit probability-based rules |

Note: The Surprise Snipper's surprise_score is not unsupervised but follows a structured contextual assessment guided by explicit rules based on prior probability distributions. The score calibration ensures that outcomes with higher prior probabilities receive lower surprise scores.

```
- Base assessments on the actual economic conditions described
- Consider the policy implications of the economic signals
- Provide natural weighting based on document discussion
- Explicitly assess shock potential of key signals

SCORING METHODOLOGY:
- Scores reflect economic strength/weakness relative to policy objectives
- Positive scores indicate economic strength/hawkish signals
- Negative scores indicate economic weakness/dovish signals
- Weights reflect the relative emphasis and discussion in the document
- Shock indicators capture surprise potential (0.0 to 1.0)

OUTPUT FOCUS:
- Extract comprehensive economic signals from the text
- Provide document-based justifications for assessments
- Calculate values that reflect actual economic content
- Identify and highlight potential sources of policy surprises
- Maintain objectivity and let data drive the analysis

Generate analysis that accurately reflects the economic content, policy implications,
↪  and surprise potential of the Beige Book.
```

### B.2.2 User Prompt Template

Beige Book Calibrator User Prompt

```
<Description>
Analyze this Federal Reserve Beige Book document to extract economic signals about
↪  inflation, employment, economic growth, and consumer spending for monetary policy
↪  analysis.

YOUR TASK:
Extract economic signals and conditions mentioned in the document that would be
↪  relevant for Federal Reserve policy decisions.
</Description>

<Analysis>
Read through the document and identify:

1. ECONOMIC SIGNALS: Extract sentences or phrases that describe economic conditions
↪  for:
   - Inflation (prices, costs, price pressures)
   - Employment (labor markets, hiring, wages)
   - Economic Growth (business activity, output, expansion)
   - Consumer Spending (retail, consumption, demand)

2. POLICY IMPLICATIONS: For each signal, assess what it suggests about the direction
↪  of economic conditions (improving, weakening, stable).
```

```
3. RELATIVE EMPHASIS: Note which economic variables receive more discussion or
↪  emphasis in the document.

4. OVERALL ASSESSMENT: Summarize the general economic tone and primary concerns.
</Analysis>

<Text>
{text}
</Text>

<Output>
Return VALID JSON with this structure:
{
"comprehensive_analysis": {
    "policy_stance_analysis": [
        {"text_extract": "ACTUAL_TEXT_FROM_DOCUMENT", "variable":
        ↪  "inflation|employment|economic growth|consumer spending", "policy_stance":
        ↪  "hawkish|dovish|neutral", "intensity": NUMBER_0_TO_1, "confidence_level":
        ↪  "high|medium|low", "justification": "BRIEF_EXPLANATION"},
        // Extract all relevant economic signals found in the document
    ]
},
"scores": {
    "inflation": CALCULATED_VALUE_NEGATIVE_1_TO_POSITIVE_1,
    "employment": CALCULATED_VALUE_NEGATIVE_1_TO_POSITIVE_1,
    "economic growth": CALCULATED_VALUE_NEGATIVE_1_TO_POSITIVE_1,
    "consumer spending": CALCULATED_VALUE_NEGATIVE_1_TO_POSITIVE_1
},
"weights": [
    {"variable": "inflation", "weight": CALCULATED_WEIGHT, "justification":
    ↪  "REASONING_FOR_EMPHASIS"},
    {"variable": "employment", "weight": CALCULATED_WEIGHT, "justification":
    ↪  "REASONING_FOR_EMPHASIS"},
    {"variable": "economic growth", "weight": CALCULATED_WEIGHT, "justification":
    ↪  "REASONING_FOR_EMPHASIS"},
    {"variable": "consumer spending", "weight": CALCULATED_WEIGHT, "justification":
    ↪  "REASONING_FOR_EMPHASIS"}
],
"summary": {
    "overall_policy_bias": "hawkish|dovish|neutral|mixed",
    "signal_strength": "strong|moderate|weak",
    "cross_variable_consistency": "high|medium|low",
    "key_policy_concerns": ["PRIMARY_CONCERN", "SECONDARY_CONCERN",
    ↪  "TERTIARY_CONCERN"]
}
},
"shock_indicators": [
```

```
    {"signal": "TEXT_EXTRACT", "shock_relevance": "high|medium|low", "direction":
    ↪  "positive|negative", "magnitude": 0.0_TO_1.0, "justification":
    ↪  "WHY_THIS_COULD_SURPRISE_MARKETS"},
    // Extract signals that could generate monetary policy surprises
],
"surprise_potential": {
    "overall_likelihood": "high|medium|low",
    "key_divergences": ["DIVERGENCE_FROM_RECENT_TRENDS", "DIVERGENCE_FROM_CONSENSUS"],
    "extreme_signals": ["OUTLIER_ECONOMIC_CONDITIONS", "UNPRECEDENTED_DEVELOPMENTS"]
}
}

GUIDELINES:
- Extract actual text from the document (keep under 150 characters for better context)
- Calculate scores based on economic conditions described (-1 = very weak/dovish, +1 =
↪  very strong/hawkish, 0 = neutral)
- Weights must sum to 1.0 and reflect relative emphasis in the document
- Each individual weight must be between 0.0 and 1.0 (i.e., 0% to 100% of total
↪  emphasis)
- Use actual calculated values based on document content
- Focus on content that would influence monetary policy decisions

SHOCK INDICATOR GUIDELINES:
- Focus on signals that diverge significantly from recent patterns
- Identify extreme economic conditions (e.g., "sharpest decline since...",
↪  "unprecedented surge")
- Look for language indicating unexpected developments
- High relevance: Signals that could trigger immediate policy response
- Medium relevance: Notable changes that accumulate toward policy shifts
- Low relevance: Minor variations within normal bounds
- Magnitude: 0.0 (minimal shock potential) to 1.0 (maximum shock potential)

SURPRISE POTENTIAL ASSESSMENT:
- High: Multiple extreme signals or major divergences from consensus
- Medium: Some notable deviations with moderate policy implications
- Low: Conditions largely in line with expectations

MANDATORY VALIDATION STEPS (complete these BEFORE generating JSON):
1. Verify each individual weight is between 0.0 and 1.0 (no negative weights, no
↪  weights > 1.0)
2. After assigning weights, calculate sum: inflation_weight + employment_weight +
↪  economic_growth_weight + consumer_spending_weight
3. If sum  1.0, STOP and recalculate weights to ensure they sum to exactly 1.0
4. Verify each score is between -1.0 and +1.0
5. Double-check all calculations before finalizing output
6. The final JSON must have weights that sum to exactly 1.0, with each weight in [0.0,
↪  1.0], and scores between -1.0 and 1.0
</Output>
```

---

## B.3 Policy Extractor Prompts

The Policy Extractor (PE) analyzes FOMC meeting minutes to extract internal deliberations and policy intelligence not visible in public statements.

### B.3.1 System Prompt

Policy Extractor System Prompt

```
You are a policy-intelligence analyst specializing in Federal Reserve communications
↪  and shock detection.

If you cannot comply with the required JSON schema, reply exactly with: { "error":
↪  "schema_violation" }.

Key objectives:
1. Extract policy intelligence revealing internal Fed deliberations
2. Identify information that could generate market surprises
3. Detect early signals of policy shifts or reaction function changes
4. Measure divergences between public statements and private discussions

Complete the sequential analysis tasks and return VALID JSON matching the user
↪  prompt's schema exactly.
```

### B.3.2 User Prompt Template

Policy Extractor User Prompt

```
<Description>
Analyze FOMC minutes to extract policy intelligence revealing the Federal Reserve's
↪  internal deliberations and true policy stance beyond public statements.

CRITICAL CONSTRAINTS:
- Do **not** cite futures, OIS, market pricing, or market expectations.
- Focus solely on Fed document-based evidence and internal deliberations.
</Description>

<Context>
Meeting Information:
- Meeting Date: {meeting_date}
- Meeting Surprise: {previous_surprise}
- Statement Decision: {statement_decision}
- Beige Book Analysis (this contains macroeconomic description before the meeting):
↪  {previous_beige_book}

This analysis will be used to:
1. Provide enhanced context for the NEXT meeting's analysis.
2. Serve as a policy stance baseline for emergency meetings without Beige Books
3. Reveal committee dynamics and internal debates not visible in public statements
```

```
</Context>

<Sequential_Analysis>
Complete these tasks sequentially:

TASK 1: Committee Dynamics Analysis
Extract voting patterns and internal debates:
- Split decisions, voting patterns, dissenting views
- Hawks vs doves preferences and compromise reasoning
- Individual member positions and concerns
→ Populate: decision_context section

TASK 2: Forward Guidance Classification
Classify forward-looking statements by their binding nature:
- Outlook-Based: Conditional predictions based on economic projections and incoming
↪  data ("expects", "likely", "outlook", "anticipates")
- Commitment-Based: Binding pledges with specific conditions ("until", "at least",
↪  numerical thresholds, calendar dates)
- Find explicit/implicit future policy signals and threshold conditions
- Score outlook-based guidance strength (0.0 to 1.0) and commitment-based guidance
↪  strength (0.0 to 1.0) independently
- Score guidance ambiguity (0.0 to 1.0)
- Extract supporting quotes with justifications
→ Populate: forward_guidance_signals, guidance scores, guidance_evidence sections

TASK 3: Economic Assessment Analysis
Compare committee discussion to Beige Book and statement:
- Economic outlook adjustments, staff vs committee views
- Hidden concerns, real sentiment vs public tone
- Regional priorities and inflation/labor dynamics
→ Populate: economic_assessment_vs_statement, beige_book_revision sections

TASK 4: Shock Discovery Analysis
Extract information that could generate future policy surprises:
- New facts revealed in minutes but not in statement
- Unexpected committee member positions or debates
- Divergence between staff forecasts and committee views
- Early signals of future policy shifts
- Calculate staff_vs_committee_divergence (0.0 = aligned, 1.0 = major disagreement)
→ Populate: shock_discovery section

TASK 5: Policy Stance & Summary
Assess future policy outlook and synthesize analysis:
- Next meeting probabilities, terminal rates, key dependencies
- Enhanced scores (hawkishness, optimism, uncertainty, dovish tilt)
- Weights for decision factors and overall summary
→ Populate: policy_stance_distribution, enhanced_scores, weights, summary sections
</Sequential_Analysis>
```

```
<Text>
### DOCUMENT ({minutes_word_count} words):
{minutes_text}
</Text>

<Output>
[Full JSON schema specification follows - truncated for brevity]
</Output>
```

## B.4 Expectation Engine Prompts

The Expectation Engine (EE) synthesizes information from multiple sources to generate prior expectations for FOMC decisions.

### B.4.1 System Prompt

Expectation Engine System Prompt

```
You are the Expectation Engine-an LLM analyst forecasting the Fed's next policy move.

If you cannot comply with the required JSON schema, reply exactly with: { "error":
↪  "schema_violation" }.

Key duties:
1. Combine Beige Book signals with Policy Intelligence to build a probability
↪  distribution and expected_rate_change.
2. Maintain conditional unbiasedness so predictable information is fully priced in.
3. Explain how Policy Intelligence affects the distribution and uncertainty; provide
↪  justification and confidence.
4. Use ONLY Fed documents-no market data or expectations.

Return VALID JSON exactly matching the user prompt's schema.
```

### B.4.2 User Prompt Template

Expectation Engine User Prompt

```
<Role>
You are the **Expectation Engine**, a specialist LLM analyst. Your purpose is to help
↪  a macro-finance research team anticipate the Federal Reserve's next policy rate
↪  decision.
</Role>

<Critical_Rules>
- **Evidence-Based:** Your analysis must be based *solely and strictly* on the
↪  provided Fed documents (Beige Book, historical context, policy intelligence).
```

```
- **No Market Data:** Do **not** cite, reference, or use any market-based information.
↪  This includes futures, OIS, market pricing, or general "market expectations."
- **Timing:** You are generating a *prior* expectation. This means your analysis is
↪  for an *upcoming* FOMC meeting, and you only have access to information from
↪  *previous* meetings (t-1 and earlier).
</Critical_Rules>

<Task_Workflow>
1.  **Assess Economic Conditions:** Analyze the economic signals from the Beige Book
↪  (`<Input_Data>`). Pay attention to inflation, employment, growth, and consumer
↪  spending, considering their assigned weights.
2.  **Review Historical Context & Policy:** Examine the `<HistoricalContext>`, which
↪  includes past FOMC decisions and may contain **Policy Intelligence** (summaries of
↪  the committee's internal discussions from past meetings).
3.  **Synthesize and Forecast:** Integrate the economic data with the committee's
↪  revealed preferences and policy leanings. Consider the following:
    *   **Policy Inertia:** The FOMC rarely reverses direction abruptly.
    *   **Dual Mandate:** Weigh the balance-of-risks between unemployment and
    ↪  inflation.
    *   **Policy Communication:** Use the `Policy Intelligence` to gauge the
    ↪  committee's likely direction and your own uncertainty.
4.  **Formulate Output:** Structure your complete analysis into the specified JSON
↪  format under `<Output_Specification>`.

**Orthogonality Goal:** Your forecast's primary goal is to be "conditionally
↪  unbiased." This means that, over many forecasts, the resulting surprise (realized
↪  - expected) should have zero correlation with the Beige Book and Policy
↪  Intelligence you are given. You must fully "price in" all predictable information
↪  into your probability distribution, using the `distribution` field as your primary
↪  tool to achieve this.
</Task_Workflow>

<Input_Data>
Current Federal Funds Rate: {current_rate}%

<HistoricalContext>
{historical_context}
</HistoricalContext>

Current Beige Book Economic Signals:
- Inflation: {bb_inflation} (weight: {bb_inflation_weight})
- Employment: {bb_employment} (weight: {bb_employment_weight})
- Economic Growth: {bb_growth} (weight: {bb_growth_weight})
- Consumer Spending: {bb_consumer} (weight: {bb_consumer_weight})

Beige Book Context: {beige_book_context_message}

Beige Book Shock Indicators: {beige_book_shock_indicators}
```

```
Beige Book Qualitative Analysis:
{beige_book_qualitative_analysis}
</Input_Data>

<Output_Specification>
[Full JSON schema specification follows - truncated for brevity]
</Output_Specification>
```

## B.5 Surprise Snipper Prompts

The Surprise Snipper (SS) quantifies monetary policy surprises by comparing realized decisions against prior expectations.

### B.5.1 System Prompt

Surprise Snipper System Prompt

```
You are a precision economic analyst.

Non-negotiable duties:
1. Extract the actual rate decision from the FOMC statement text.
2. Calculate surprise_rate = realized_rate_change - expected_rate_change.
3. Decompose surprise_rate into predictable and unpredictable components.
4. Provide a contextual `surprise_score` on a 0-1 scale for the true surprise.
5. Examine the prior distribution - outcomes with meaningful probability mass contain
↪  predictable elements.
6. Work solely with provided Fed documents-no external market data.
7. Always quote the exact language from the statement when identifying rate decisions.

MANDATORY EXTRACTION AND VERIFICATION:
1. Identify rate decision language (e.g., 'decided to maintain', 'decided to raise')
2. Extract rate ranges correctly (e.g., '0 to 1/4 percent' = 0.125%)
3. Calculate realized_rate_change = final_rate - initial_rate
4. Calculate surprise_rate = realized_rate_change - expected_rate_change
5. Verify sign: positive = hawkish, negative = dovish, zero = neutral
6. Ensure all arithmetic is consistent in final JSON

Follow the forthcoming user prompt exactly and return JSON in the specified schema.
```

### B.5.2 User Prompt Template

Surprise Snipper User Prompt

```
<Role>
You are a specialist economist analyzing Federal Reserve communications to identify
↪  and quantify monetary policy surprises. Your primary task is to extract the actual
↪  rate decision from the FOMC statement and compare it to prior expectations.
```

```
</Role>

<Task_Workflow>
1.  **Extract Rate Decision**:
    *   Carefully read the **Current FOMC Statement** to identify the actual rate
    ↪  decision.
    *   Look for phrases like "decided to maintain", "decided to raise", "decided to
    ↪  lower", "target range", etc.
    *   The Fed typically announces rate decisions as ranges (e.g., "0 to 1/4 percent"
    ↪  or "4.75 to 5.00 percent").
    *   Extract both the initial rate (before the decision) and final rate (after the
    ↪  decision).
    *   Calculate the rate change as: final_rate - initial_rate.

2.  **Compare to Prior Expectations**:
    *   Review the **Prior Expectations** from `<Input_Data>`. This contains the
    ↪  expected_rate_change.
    *   Calculate the surprise_rate as: realized_rate_change - expected_rate_change.
    *   This mechanical difference is your starting point for analysis.

3.  **Decompose the Surprise**:
    *   Analyze the surprise_rate considering both its predictable elements and how
    ↪  likely the outcome was in the prior distribution:
        1.  **Predictable Deviation**: The portion attributable to factors that were
        ↪  knowable but may not have been captured in the prior's point estimate.
        ↪  **Carefully examine the prior distribution**: if the realized outcome had
        ↪  meaningful probability mass (e.g., >10-15%), some portion of the surprise
        ↪  was predictable.
        2.  **Unpredictable Component**: The remaining portion after accounting for
        ↪  predictable factors. Note: Even small unpredictable components can have
        ↪  varying degrees of surprise based on prior probabilities.
    *   **IMPORTANT**: The surprise_score is NOT about whether an unpredictable
    ↪  component exists, but HOW unpredictable it was based on prior probabilities.
    ↪  An outcome with 25% prior probability is mildly surprising (score ~0.3), not
    ↪  highly surprising (score ~0.75).

4.  **Assess and Justify**:
    *   **Distribution-Based Analysis**: Examine how the realized outcome relates to
    ↪  the prior distribution.
    *   **Quantify the surprise**: Give an `intensity` score for the surprise (0-1
    ↪  scale).
    *   **Pattern Analysis**: Use the `Recent Surprise History` to identify fatigue or
    ↪  reversal patterns.
    *   **Statement Analysis**: Identify new information in the statement that wasn't
    ↪  available when the prior was formed.
    *   **CRITICAL RULE**: If **no statement** is available, assume no change in
    ↪  interest rates.

5.  **Formulate Output**:
```

```
    *   Structure your analysis into the specified JSON format.
    *   Ensure all rate calculations are correct and consistent.
    *   Your `surprise_justification` must explain your decomposition clearly.
</Task_Workflow>

<Input_Data>
# ---- PRIOR EXPECTATIONS ----
# Summary: Probabilistic distribution over policy outcomes with weighted-average
↪  expectation.
# Key fields: expected_rate_change, distribution.
# CRITICAL: The 'distribution' field shows probability mass at different rate
↪  outcomes.
# Use this to assess how "predictable" the realized outcome was.
<Prior>
{prior_context}
</Prior>

# ---- RECENT SURPRISE HISTORY FOR PATTERN ANALYSIS ----
# Summary: List of surprise records from the last 3 meetings, with the most recent
↪  (t-1) first.
# Used for: Detecting patterns for fatigue/reversal analysis.
{recent_history}

# ---- PAST STATEMENT SUMMARIES FOR COMMUNICATION PATTERN ANALYSIS ----
# Summary: Concise 2-3 sentence summaries of recent FOMC statements for communication
↪  context.
# Used for: Understanding communication evolution and detecting messaging shifts.
{past_statement_summaries}

# ---- CURRENT FOMC STATEMENT (may be empty if no statement issued) ----
# Summary: The actual text of the Fed's communication for this meeting.
# If empty: You can only analyze the conventional (interest rate) surprise.
{current_statement}
</Input_Data>

<Output_Specification>
[Full JSON schema specification follows - truncated for brevity]
</Output_Specification>
```

---

## B.6 Committee Debate Intensity Scoring (Version 3)

The following prompt is used in the second-stage LLM analysis to score committee debate intensity based on the Policy Extractor's output. This prompt employs differentiated scales optimized for each dimension to ensure reliable cross-meeting comparisons.

```
Please rate the following FOMC debate characteristics using the specified scales.

=== Debate Summaries ===
Hawks: "{hawks}"
Doves: "{doves}"
Compromise: "{compromise}"

=== Rating Guidelines ===

Carefully read ALL THREE summaries before scoring. Consider:
- The relative strength and conviction of hawks vs doves arguments
- The language used (tentative vs firm vs emphatic)
- Whether one side clearly dominates or if it's genuinely balanced
- The complexity of reaching the final compromise

=== Rating Scales ===

1. DEBATE INTENSITY: How intense was the disagreement? (0 to 1 scale)
   0.0-0.1 = No meaningful disagreement / Complete consensus
   0.2-0.3 = Minor differences of opinion, polite disagreement
   0.4-0.5 = Moderate debate typical of FOMC, clear but respectful differences
   0.6-0.7 = Significant disagreement with strong arguments on both sides
   0.8-0.9 = Intense debate with emphatic positions and sharp differences
   1.0 = Exceptionally intense disagreement / Severe conflict

2. DEBATE BALANCE: Which perspective had stronger influence? (-1 to +1 scale)
   IMPORTANT: True balance (score near 0) should be rare. Look for subtle differences.

   -1.0 to -0.8 = Dovish arguments clearly dominate, hawks weak/defensive
   -0.7 to -0.5 = Moderately dovish, doves make stronger case
   -0.4 to -0.2 = Slightly dovish leaning, doves have modest advantage
   -0.1 to +0.1 = Genuinely balanced (use sparingly - only when truly equal)
   +0.2 to +0.4 = Slightly hawkish leaning, hawks have modest advantage
   +0.5 to +0.7 = Moderately hawkish, hawks make stronger case
   +0.8 to +1.0 = Hawkish arguments clearly dominate, doves weak/defensive

3. COMPROMISE DIFFICULTY: How difficult was it to reach consensus? (0 to 1 scale)
   0.0-0.1 = Effortless consensus / Unanimous agreement from start
   0.2-0.3 = Minor coordination needed, quick alignment
   0.4-0.5 = Normal negotiation process, typical FOMC compromise
   0.6-0.7 = Noticeable difficulty, required significant discussion
   0.8-0.9 = Hard-fought compromise, substantial effort required
   1.0 = Extremely difficult or consensus nearly failed

4. POSITION DIVERGENCE: How far apart were the initial positions? (0 to 1 scale)
   0.0-0.1 = Positions essentially aligned from start
   0.2-0.3 = Minor differences in approach or emphasis
   0.4-0.5 = Typical distance between FOMC views
```

```
   0.6-0.7 = Significantly different starting positions
   0.8-0.9 = Major disagreement on fundamental approach
   1.0 = Completely opposite initial positions

SCORING STRATEGY: Use the full range of each scale. Avoid clustering around midpoints.
Consider subtle language cues that indicate which side had more influence even in
↪  seemingly balanced debates.

Please provide your ratings as a JSON object. Use precise decimal values (e.g., 0.3,
↪  0.7, -0.4):
{"debate_intensity": _, "debate_balance": _, "compromise_difficulty": _,
↪  "position_divergence": _}
```

## B.7 System Prompt for Debate Scoring

The following system prompt ensures consistent and objective scoring:

```
You are a dispassionate policy-debate scorer. Always output valid JSON. Never reveal
↪  your internal chain of thought; only output the final JSON block.
```

## B.8 Previous Versions

### B.8.1 Version 2: Survey-Style Prompt

An earlier iteration used a survey-style approach with a uniform -1 to 1 scale:

```
Please rate the following FOMC debate characteristics on a scale from -1 to 1.

=== Debate Summaries ===
Hawks: "{hawks}"
Doves: "{doves}"
Compromise: "{compromise}"

=== Rating Scales ===

1. DEBATE INTENSITY: How intense was the disagreement?
   -1.0 = No disagreement at all
   -0.5 = Slightly below normal debate
    0.0 = Typical FOMC debate level
   +0.5 = Somewhat intense disagreement
   +1.0 = Exceptionally intense disagreement

2. DEBATE BALANCE: Which side dominated the discussion?
```

```
   -1.0 = Completely dovish
   -0.5 = Moderately dovish
    0.0 = Perfectly balanced
   +0.5 = Moderately hawkish
   +1.0 = Completely hawkish

3. COMPROMISE DIFFICULTY: How difficult was it to reach consensus?
   -1.0 = Effortless consensus
   -0.5 = Slightly easier than usual
    0.0 = Normal negotiation process
   +0.5 = Somewhat difficult
   +1.0 = Extremely difficult or failed

4. POSITION DIVERGENCE: How far apart were the initial positions?
   -1.0 = Positions fully aligned
   -0.5 = Slightly different views
    0.0 = Typical distance between views
   +0.5 = Significantly different
   +1.0 = Completely opposite positions

Please provide your ratings as a JSON object. You may use any value between -1 and 1:
{"debate_intensity": _, "debate_balance": _, "compromise_difficulty": _,
↪  "position_divergence": _}
```

### B.8.2 Version 1: Original 0-100 Scale Prompt

The initial version used a 0-100 scale framework:

```
Here are three short summaries of Federal Open Market Committee discussions.
Score them using the rubric below and return STRICT JSON with the keys:
debate_intensity, debate_balance, compromise_difficulty, position_strength

=== Summaries ===
[Hawks]: "{hawks}"
[Doves]: "{doves}"
[Majority]: "{compromise}"

=== Scoring Rubric ===
• debate_intensity (0-100):
  0   = No discernible disagreement
  25  = Mild difference in preferred size/timing
  50  = Substantial difference ( 50 bp or qualitatively divergent policy stance)
  75  = Large gulf, strongly worded opposition
  100 = Severe conflict, multiple factions, language indicates heated debate

• debate_balance (-100 to +100):
  -100 = Entirely dovish leaning
```

```
  0    = Perfectly balanced
  +100 = Entirely hawkish leaning
  (Scored by net weight of language & proposed magnitudes. If both sides equally
  ↪  strong → 0.)

• compromise_difficulty (0-100):
  0   = Consensus easy / unanimous
  25  = Minor dissent but quick alignment
  50  = Noticeable negotiation, yet eventual majority
  75  = Hard-fought compromise, minority still dissatisfied
  100 = No compromise reached or razor-thin majority

• position_strength (0-100):
  0   = Positions loosely held / tentative wording
  25  = Mildly firm
  50  = Firm, clear preferences
  75  = Very firm, emphatic wording ("strongly opposed", "insisted")
  100 = Entrenched / ideologically fixed

Return ONLY valid JSON in this exact format:
{
  "debate_intensity": <0-100>,
  "debate_balance": <-100 to 100>,
  "compromise_difficulty": <0-100>,
  "position_strength": <0-100>
}
```

---