Monetary Policy Shocks: A New Hope

Large Language Models and Central Bank Communication*

Rubén Fernández-Fuertes†

Job Market Paper

First Version: 12th October 2025 – This Version: October 29, 2025

Please check here for the latest version

Abstract

I develop a multi-agent LLM framework that processes Federal Reserve communications to construct narrative monetary policy surprises. By analyzing Beige Books and Minutes released before each FOMC meeting, the system generates conditional expectations that yield less noisy surprises than market-based measures. These surprises produce theoretically consistent impulse responses where contractionary shocks generate persistent disinflationary effects, and enable profitable yield curve trading strategies that outperform alternatives. By directly extracting expectations rather than cleaning surprises ex post, this approach demonstrates how multi-agent LLMs can implement narrative identification at scale without contamination in high-frequency measures.

JEL Classification: E52, E58, E43, G14, C45, C55

Keywords: Monetary Policy Shocks, Central Bank Communication, Large Language Models, FOMC, Federal Reserve, Natural Language Processing, High-Frequency Identification, Term Structure

*I am deeply grateful to Max Croce, Carlo A. Favero, and Claudio Tebaldi for their invaluable supervision and guidance throughout my Ph.D. journey. I acknowledge the financial support of the Baffi Centre. I also thank Josefina Cenzon, Fiorella de Fiore, Mohammad R. Jahan-Parvar, Alejandra Inzunza, Nicola Gennaioli, Martin Fankhauser, Tommaso Monacelli, David Murakami, Fernando Pérez-Cruz, Angelo Ranaldo, Kevin Schneider, Ivan Shchapov, Damiano Sandri, Jakob Ahm Sørensen, Isabella M. Wolfskeil for their insightful comments and suggestions. Special thanks go to participants of the following conferences: AFA Ph.D. Student Poster Session, Bocconi Alumni Conference, Lausanne PhD Macroeconomics Conference; and the staff at the Bank of International Settlements, where I had the pleasure of presenting my research. Parts of this paper were written whilst I was visiting the Bank for International Settlements to whom I am grateful for their hospitality.

†Department of Finance, Bocconi University, Milan, Italy. Email: ruben.fernandez@phd.unibocconi.it. Website: rubenfernandezfuertes.com

1 Introduction

Measuring monetary policy *shocks* requires isolating the component of Federal Reserve decisions that is genuinely unpredictable from available information. Identification strategies have evolved through three generations—narrative measures, structural VARs, and high-frequency identification—each addressing limitations of its predecessor while introducing new challenges. High-frequency identification, pioneered by Kuttner (2001), constructs market-based *surprises* from federal funds futures to separate anticipated from unanticipated policy changes. However, these *surprises*, measured from high-frequency interest rate movements around FOMC announcements, suffer from contamination by non-policy information (Bauer & Swanson, 2023a; Jarociński & Karadi, 2020; Miranda-Agrippino & Ricco, 2021; Nakamura & Steinsson, 2018). Narrative measures, such as those in Romer and Romer (1989) (henceforth, R&R), are thought to avoid this contamination but are impossible to implement in real time and are limited to binary shock indicators (shock or no shock). The common practice has been to clean these surprises *ex post* of information contamination to obtain stronger instruments for monetary policy shocks identification.

In this paper, I develop an *ex ante* method to construct monetary policy *surprises* by eliciting conditional expectations from the Federal Reserve's public communications released weeks before each FOMC decision. Using a multi-agent LLM system that processes Beige Books, Minutes, and Statements in a survey-like manner (querying the model to extract probabilistic beliefs from narrative content), I form complete probability distributions over potential Fed actions based solely on the documentary record available to all market participants. This nonparametric, nonlinear extraction of conditional expectations from narrative text yields less noisy surprises that explain 52% of policy rate variation, compared with 15–17% from market-based measures. Such 52% reflects the unpredictable component of policy decisions, whereas the remainder represents systematic Fed responses that markets should anticipate from official communications. By measuring surprises against the information set frozen at Beige Book release (two weeks before meetings), I isolate what was genuinely unexpected from the Fed's public messaging. The difference from market-based measures reflects distinct information sets: my surprises capture what was unpredictable from Fed documents released 2–3 weeks before meetings, while market measures incorporate all information up to announcement moments and may include flows that may contaminate identification. By measuring against the Fed's

communications before each FOMC meeting, I better isolate policy variation that is genuinely surprising with respect to the Fed's information set.

The methodology revives R&R's identification approach while addressing its fundamental limitations through three advances. First, I employ survey-elicitation from narrative text using a multi-agent LLM system, processing the Fed's entire documentary corpus at scale to extract conditional expectations nonparametrically. This solves both the practical constraint that makes R&R's approach infeasible in real time and the functional form restrictions inherent in parametric linear residual methods. Second, I generate complete probability distributions over potential Fed actions rather than binary shock indicators, capturing information about magnitude and uncertainty that discrete classifications discard. Third, by forming expectations exclusively from documents released before each FOMC meeting decision's blackout period, my surprises are predetermined relative to announcement-day information flows. Specifically, I use the Beige Book released two weeks before meeting and the Minutes from the meeting before (released three weeks after that prior decision), mitigating contamination issues that may affect high-frequency measures.

By doing so, I promote an important methodological contribution: direct extraction of conditional expectations offers a path forward beyond *ex post* cleaning. While I implement this approach using only public Fed documents, my framework naturally extends to richer information sets—central banks incorporating internal forecasts, market participants integrating real-time news flows, or researchers combining multiple communication channels. This paper establishes the following baseline: even the simplest version, processing only Beige Books and Minutes, achieves minimal measurement error and properly calibrated expectations. The *New Hope* is that identification quality improves as extraction methods become more sophisticated, not through ever-more-elaborate *ex post* econometric adjustments.

I show that my narrative surprises are less noisy and better capture monetary policy shocks with cleaner identification through three sets of findings. First, they exhibit minimal measurement error, compared with attenuation bias in R&R and contamination in market measure. These classical measurement-error diagnostics reveal that narrative surprises directly extracted from Fed communications provide less noisy signals of true policy shocks.

Second, impulse response analysis (Ramey, 2016) reveals transmission patterns that are consistent with benchmark DSGE models. These results distinguish my measure from alternatives. Comparing against established narrative methods (Aruoba & Drechsel, 2024; Romer & Romer,

2004) and *ex post* cleaned market measures (Bauer & Swanson, 2023b; Miranda-Agrippino & Ricco, 2021), my measure generates more pronounced and synchronized contractionary effects, particularly for unemployment, industrial production, and equity markets, while maintaining tighter confidence bands than *ex post* cleaned approaches.

Following a 25 basis point contractionary surprise, my measure generates theoretically coherent transmission dynamics across macroeconomic and financial variables. Price levels decline immediately and persistently (avoiding the "price puzzle"), real GDP and industrial production show sustained contractionary effects, and unemployment rises with delayed response. While confidence bands are wide (reflecting the well-known difficulty in estimating monetary policy effects), the consistent contractionary signs across all variables distinguish my measure from market-based alternatives that often produce sign anomalies requiring instrumental variable corrections. Simultaneously, the yield curve decomposes into two distinct transmission channels: initial spread compression driven by rising expected short rates, followed by delayed spread steepening driven by rising term premia as policy uncertainty compounds. Equivalently, these two channels reveal that monetary policy shocks affect both expectations and risk compensation.

Third, my measure advantages translate into economic profitability. An implementable duration-hedged yield curve strategy trading on the surprise as a signal generates 43.7% cumulative returns over the entire sample, more than doubling market-based alternatives. This out-of-sample validation across 265 FOMC meetings demonstrates that my surprises contain genuine policy information that markets have not yet priced.

The multi-agent LLM framework processes the Federal Reserve's public communication timeline to form probabilistic expectations before each FOMC decision. For each meeting t, the Fed releases: (1) the Beige Book approximately two weeks before the meeting, containing qualitative assessments of regional economic conditions; and (2) the Minutes from the *previous* meeting $t-1$, released approximately three weeks after that decision (hence available before meeting t's blackout period), revealing the Committee's deliberations, forward guidance intentions, and distribution of views from the prior decision. This structured communication system motivates a four-agent pipeline: one agent quantifies economic conditions from the Beige Book across dual-mandate variables, a second agent extracts policy stance and forward guidance from Minutes, a third synthesizes these inputs with historical FOMC statements into probability distributions over potential Fed actions, and a fourth computes surprises by comparing prior distributions to actual decisions. This architecture mirrors the Fed's actual communication

timeline rather than treating documents in isolation.

Following Li et al. (2024) and Tillmann (2025), I employ multiple agents to mitigate inherent variability in individual LLM responses. The resulting continuous probability distributions maintain the conceptual clarity that distinguished R&R's approach from reduced-form VAR identification while overcoming a critical limitation: early narrative methods relied on binary shock indicators that discarded information about magnitude and uncertainty. By processing only public documents released before the blackout period (when Committee members cease public commentary), my framework produces surprises that are predetermined relative to announcement-day information flows. This timing structure provides three identification advantages: (1) surprises are orthogonal to announcement-day asset price movements and data releases, eliminating simultaneity bias; (2) institutional alignment with the Committee's deliberative timeline reduces measurement error; and (3) reduced information-effect contamination, since surprises do not conflate policy stance shifts with news about fundamentals revealed during announcements. Validation tests confirm measurement reliability: repeated pipeline runs show economically negligible variability, and out-of-sample comparisons rule out look-ahead bias as a driver of results.

Related literature. My approach addresses an identification impasse that has fragmented monetary policy research since the 1990s. The literature evolved through three waves, each resolving problems from its predecessor while introducing new limitations. First, narrative approaches (Romer & Romer, 1989, 2004) measured policy shocks through direct reading of FOMC records, but Leeper (1997) showed these measures were predictable from past macroeconomic variables and generated price puzzles. Second, structural VARs (Christiano et al., 1999; Sims, 1980) imposed identifying restrictions on dynamic systems, but Sims (1992) documented persistent price puzzles and Stock and Watson (2001) questioned the strong, untestable assumptions about contemporaneous relationships. Third, high-frequency identification (Gertler & Karadi, 2015; Kuttner, 2001) exploited narrow event windows around announcements, but Ramey (2016) demonstrated these instruments remained predictable from Greenbook forecasts and produced different results across estimation methods (VAR versus local projections).

Current research debates whether high-frequency measures suffer from information effect contamination (Miranda-Agrippino & Ricco, 2021; Nakamura & Steinsson, 2018) (henceforth, M-A&R for Miranda-Agrippino and Ricco) or misspecified reaction functions (Bauer & Swan-

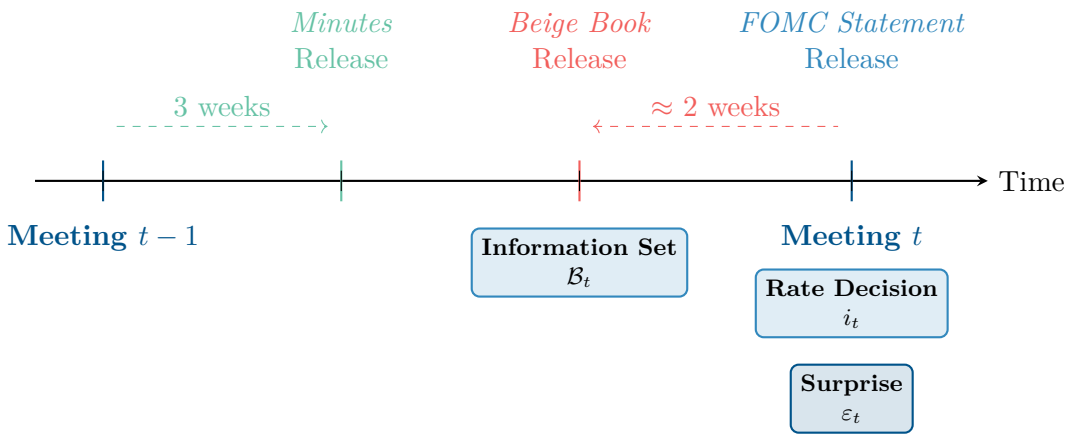
son, 2023a, 2023b) (henceforth, B&S), with recent evidence from Ricco and Savini (2025) favoring the information channel interpretation. Regardless of which mechanism prevails, both camps acknowledge that high-frequency measures conflate policy stance shifts with information revelation, contaminating shock identifithe cation. My narrative approach sidesteps this contamination by constructing expectations exclusively from public Fed documents released before the blackout period, ensuring surprises are predetermined relative to announcement-day information flows.

This distinguishes my work from two related strands of textual analysis. First, research extracting sentiment from Fed communications (Armesto et al., 2009; Balke et al., 2017; Doh et al., 2020; Filippou et al., 2024) focuses on characterizing tone rather than constructing counterfactual expectations for surprise measurement. Second, Aruoba and Drechsel (2024) use natural language processing on internal Greenbook documents to control for the Fed's private information when identifying exogenous policy shocks. Crucially, they analyze *private* documents (Greenbook forecasts unavailable to markets) to address the exogeneity problem—ensuring shocks are orthogonal to the Fed's internal information set. I instead construct expectations from *public* documents (Beige Books and Minutes released before blackout periods), addressing the surprise measurement problem—what markets should have anticipated from observable Fed communications. This public-private distinction is fundamental: they construct the Fed's information set to purge endogenous responses; I construct the market's information set to identify genuinely unpredictable policy shifts from the perspective of real-time observers.

Following Li et al. (2024) and Tillmann (2025), I employ multi-agent LLM systems to mitigate individual response variability while implementing temporal constraints preventing look-ahead bias. The framework systematically processes 256 FOMC meetings' worth of Beige Books, Minutes, and Statements (1996–2025), solving the scale constraint that limited Romer and Romer (1989) to small samples or binary measures. This methodological advance makes the narrative approach implementable at scale while maintaining its conceptual advantage of measuring surprises against the Fed's actual communication timeline.

The remainder of the paper proceeds as follows. Section 2 describes the multi-agent system architecture. Section 3 presents the data sources. Section 4 validates measurement properties of narrative surprises, impulse response analysis, and yield curve trading strategies. Section 5 discusses implications for monetary economics research and the application of Large Language Models to systematic document processing.

Figure 1: FOMC Communication Timeline and Information Structure



Note: The figure illustrates the temporal sequence of Federal Reserve communications surrounding FOMC meetings. Starting from meeting $t-1$, the Minutes are released after 3 weeks, followed by the Beige Book approximately 2 weeks before meeting t. These documents constitute the information set $\mathcal{B}_t$ available to form expectations. At meeting t, the FOMC Statement announces the rate decision i_t, from which the monetary policy surprise ε_t can be computed as the deviation from expectations formed using $\mathcal{B}_t$.

2 Methodology

2.1 General Framework

The Federal Reserve's staggered document release schedule creates a natural structure for expectation formation. Forming coherent expectations requires synthesizing information scattered across documents released weeks apart, each serving different institutional purposes. To address this complexity, I develop a multi-agent system architecture that mirrors the Fed's communication structure.

The narrative approach 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 three key documents:

(1) *Beige Book* (≈2 weeks before): qualitative assessments from the 12 Districts.

(2) *FOMC Statement* (meeting day, 2:00 p.m. ET): the policy decision.

(3) *Minutes* (≈3 weeks after meeting $t-1$): deliberations and distribution of views.

I report a schematic of the FOMC's communication releases in Figure 1.

Rather than treating multi-agent collaboration as a general-purpose improvement over single models, I exploit the institutional structure that enables ontologically grounded characterization of surprises: the Federal Reserve's staggered document release schedule naturally decomposes the expectation-formation problem into distinct sub-tasks. This task decomposition princi-

ple—assigning focused, domain-specific objectives to individual agents rather than overloading a single model with competing instructions—has emerged as a core design pattern in recent LLM system architecture (Feng et al., 2025; Wu et al., 2024).

The rationale for this decomposition is economic, not merely computational. Each document type exhibits distinct linguistic properties requiring specialized analytical focus: Beige Book narratives demand sentiment extraction from qualitative regional reports, Minutes require identification of internal Committee dynamics and forward guidance signals, and expectation formation demands probabilistic reasoning that synthesizes these inputs with historical policy patterns. Moreover, the temporal sequencing creates natural information dependencies: one cannot form coherent expectations without first processing both the prior meeting's Minutes and the current Beige Book, nor can one compute surprises without a well-formed prior. Formally, this constitutes a filtration $\{\mathcal{F}_t\}_{t\geq 0}$ where each agent sequentially expands the information set, ultimately computing the conditional expectation $E[\Delta i_t|\mathcal{F}_t]$ and ensuring surprises are true innovations with $E[s_t|\mathcal{F}_t] = 0$. A sequential architecture makes these dependencies explicit in the system design, enabling agent-specific validation of logical constraints.

Given this filtration structure, I design a system of four sequential agents, each mapping directly to a stage in the FOMC communication cycle (Figure 1). Agent IM extracts policy intelligence from the previous meeting's Minutes $(t-1)$—internal deliberations, forward guidance signals, and the distribution of Committee views. Agent IB quantifies economic conditions from the current Beige Book, producing sentiment scores for inflation, employment, economic growth, and consumer spending on a $[-1, 1]$ scale. Agent II synthesizes these inputs with historical FOMC statements to generate a prior probability distribution over possible policy outcomes for meeting t, reconciling potentially conflicting signals while maintaining strict temporal cutoffs. Agent III computes the monetary policy surprise by comparing this prior distribution with the realized FOMC decision, producing both mechanical deviations and contextual salience measures that account for ex-ante probability and historical patterns. Figure 2 illustrates the information flow, with Agent II serving as the integration point for pre-meeting intelligence and Agent III as the surprise calculator. Detailed agent-specific methodology appears in Sections 2.2–2.5.

Existing textual analysis approaches navigate different trade-offs when processing this information architecture. Traditional dictionary-based methods (Ahrens & McMahon, 2021; Ahrens et al., 2024) can process large document corpora but sacrifice contextual understanding by re-

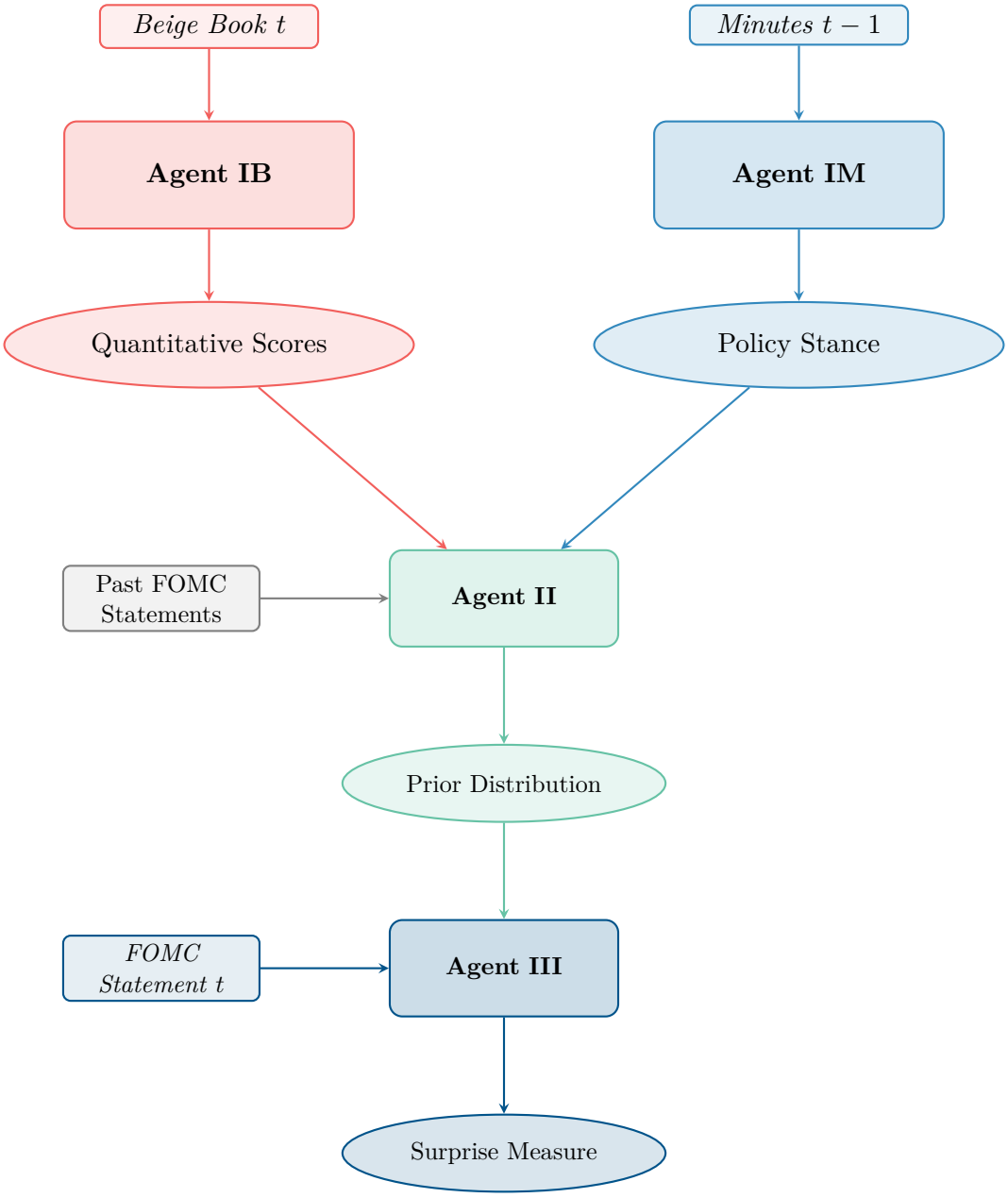
ducing text to keyword frequencies. More sophisticated natural language processing (Aruoba & Drechsel, 2024) or single-model LLM applications (De Fiore et al., 2024; Gambacorta et al., 2024; Hansen & Kazinnik, 2023) improve semantic comprehension but struggle to maintain coherent reasoning across the full documentary timeline while respecting the distinct analytical requirements each document type demands. A single prompt processing all documents simultaneously must juggle multiple conflicting objectives: extract economic sentiment from Beige Book narratives, decode Committee deliberation dynamics from Minutes, synthesize historical policy patterns, and compare expected versus realized outcomes, all while maintaining temporal consistency and minimizing look-ahead bias. I aim to address this challenge through the multi-agent system architecture designed specifically to mirror the FOMC's communication structure.

A critical methodological challenge is preventing look-ahead bias—ensuring that the system forms expectations using only information publicly available before each FOMC decision, not ex-post knowledge that could contaminate ex-ante forecasts. I address this through architectural constraints—strict temporal cutoffs for document processing and agent prompts with explicit time anchors—and empirical validation using out-of-knowledge-cutoff testing with data beyond the LLM's training window. Section 2.6 details these safeguards and presents multi-run stability analysis demonstrating that narrative measures remain consistent across independent executions.

Having outlined the framework's structure and bias-prevention mechanisms, I now detail each agent's specific methodology, followed by validation tests that establish the system's technical soundness.

The multi-agent system is implemented using OpenAI's GPT-4.1-mini model with a knowledge cutoff of June 2024. This model provides the necessary reasoning capabilities for complex document analysis while maintaining cost efficiency for processing the full FOMC corpus spanning 2000–2023. The June 2024 knowledge cutoff enables natural out-of-knowledge-cutoff validation: meetings through mid-2024 fall within the model's training window, while later meetings test the system's ability to analyze unseen Fed communications using only the provided document context rather than memorized training patterns.

Figure 2: Multi-Agent Architecture for FOMC Communication



Note: This figure shows how my four agents process documents sequentially in a top-down flow. Agent IB analyzes the Beige Book for economic conditions, Agent IM extracts policy intelligence from Minutes, Agent II synthesizes these inputs with historical context to form prior expectations, and Agent III computes surprises by comparing the prior distribution with realized FOMC decisions. This architecture ensures temporal consistency and prevents look-ahead bias.

2.2 Agent IM

Agent IM operates exclusively on the publicly released *Minutes* published 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 window into the Committee's closed-

door deliberations. Agent IM extracts *policy intelligence*—internal debates, forward guidance, and risk assessments—absent from the same-day Statement. Since Minutes are released three weeks post-decision, they become available 3-5 weeks before the incoming meeting, and they constitute a conscious and planned disclosure of information by the central bank, which may decide to distort the actual content discussed and tone used in the meeting to shape market reactions.

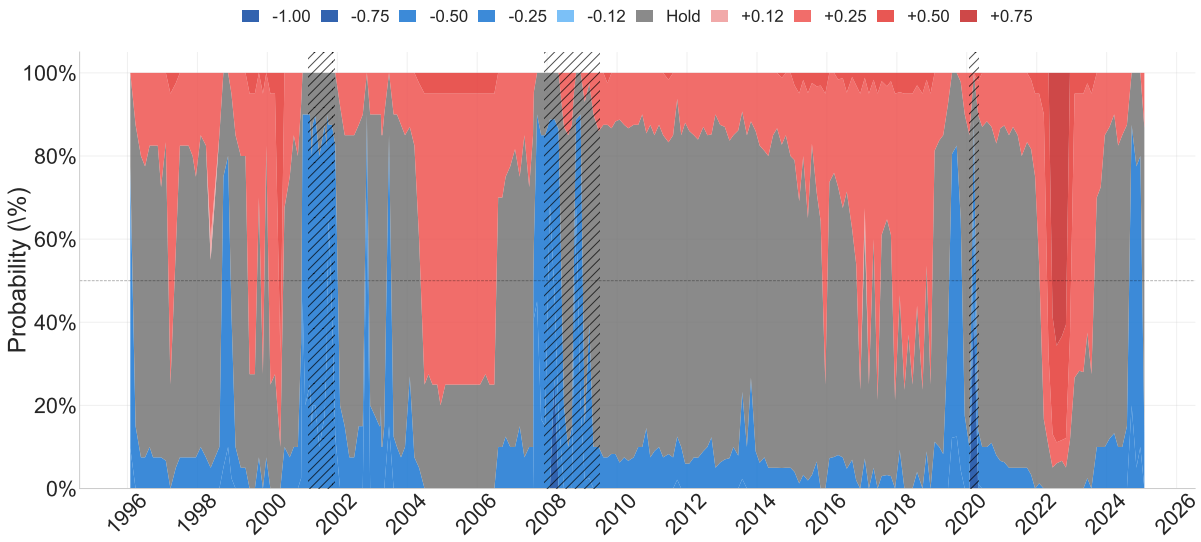
The agent produces two output types. First, *quantitative outputs*—probability distributions, continuous scores on bounded scales, and normalized weights—flow directly into Agent II's probabilistic calculations. Second, *textual outputs*—debate summaries, forward guidance language, and qualitative assessments—preserve the semantic richness of Committee deliberations for subsequent LLM interpretation.

The core quantitative output is the *updated policy distribution* for the *next* meeting, formatted as a probability mass function over discretized policy scenario. Each scenario takes the form (m, d) where $m \in \mathbb{R}_+$ represents the magnitude in percentage points and $d \in \{-1, 0, 1\}$ indicates direction (cut, hold, hike). The agent assigns probabilities $p_i \geq 0$ to each scenario i such that $\sum_i p_i = 1$. This distribution incorporates the Committee's forward-looking deliberations as revealed in Minutes, including explicit probability language (e.g., "some members judge further tightening likely"), the balance of hawks versus doves preferences, and any threshold conditions mentioned for future policy actions. Agent IM also computes scores for policy hawkishness, and level of uncertainty. The hawkishness score captures the overall tightening bias in deliberations, while the uncertainty measure quantifies the dispersion of Committee views.

To classify forward guidance strategies, Agent IM implements the theoretical framework of Campbell et al. (2012), who distinguished between Delphic guidance (conveying the central bank's economic outlook) and Odyssean guidance (constituting binding policy commitments). The agent operationalizes these concepts through linguistic proxies: outlook-based guidance score, captures conditional predictions identified through phrases like "expects," "likely," and "anticipates," while commitment-based guidance score identifies binding pledges marked by "until," "at least," and numerical thresholds. A third measure, guidance ambiguity, quantifies vagueness or internal contradictions in the Committee's communication. The agent prompt avoids the terms "Delphic" and "Odyssean" to prevent look-ahead bias. All scores are between zero and one.

The textual outputs comprise three types of extractions. *Internal debate narratives* cap-

Figure 3: Policy Stance Probabilities From Agent IM

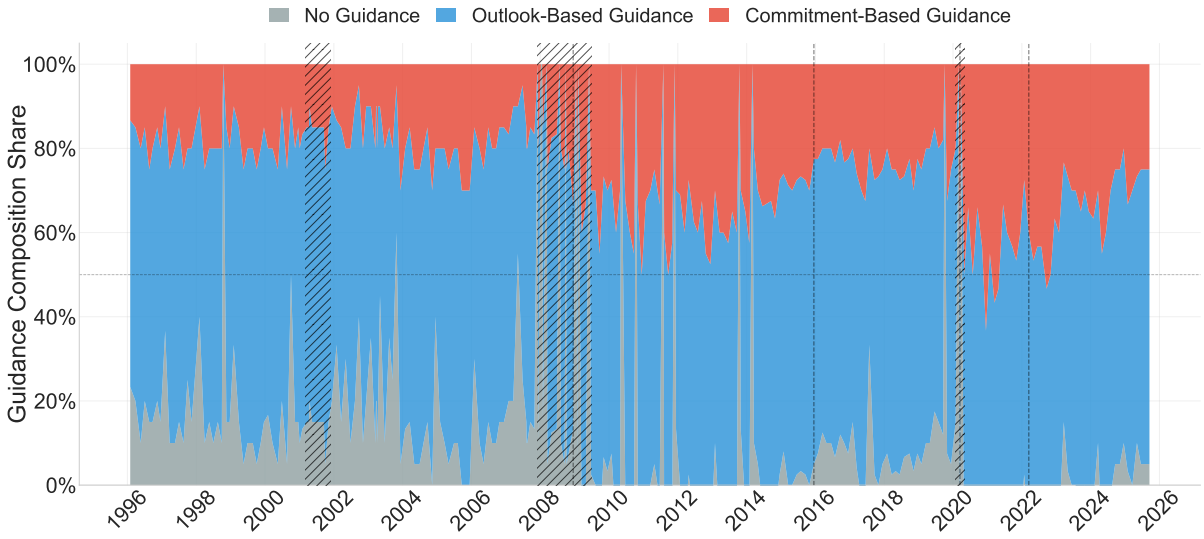


Note: The stacked bar chart shows the probability distribution for the next meeting's policy decision, decomposed into three set of outcomes: hike (red), hold (gray), and cut (blue). Key monetary policy regimes are evident, including the extended zero lower bound period (2008-2015) dominated by hold probabilities, the December 2015 liftoff marking the transition to normalization, and the aggressive tightening cycle of 2022-2023 characterized by high hike probabilities. Probabilities are extracted from FOMC Minutes released three weeks after each meeting.

ture hawks' preferences, doves' preferences, and compromise reasoning—preserving qualitative context about disagreement intensity and the fragility of consensus. *Forward guidance extractions* yield textual descriptions of explicit guidance, implicit signals, and threshold conditions for taking whatever action. For instance, vague language like "the Committee will monitor developments" carries different informational content than precise thresholds like "rate increases will continue until unemployment reaches 4.5%," even if both receive similar guidance scores. *Shock discovery* identifies textual passages revealing information that shocked the committee and forced them to take a decision that they might not have taken had this event not happened. This information has to be included in the filtration to compute a realistic and sensitive probability distribution.

I explicitly prohibit this agent from making any reference to federal funds futures, overnight index swap rates, or market-implied probabilities, ensuring that all extracted intelligence reflects the Committee's information and thereby reducing the probability of hallucination by reminding the agent that it only receives Fed information. Appendix A.1 presents complete output specifications with detailed examples from both the financial crisis period (December 2008) and tightening cycle initiation (March 2022), demonstrating the agent's dual-output structure across different monetary regimes.

Figure 4: Forward guidance composition over time



Note: The normalized 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 Agent IM'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. The classification follows the Campbell et al. (2012) framework, with outlook-based guidance as a proxy for Delphic communication and commitment-based guidance as a proxy for Odyssean communication.

Figure 3 presents the time series properties of Agent IM's policy distributions. The stacked bar chart decomposes next-meeting probabilities into three outcomes: hike 25bp (red), hold (gray), and cut 25bp (blue). During the zero lower bound period (2008-2015), overwhelming hold probabilities reflect the Committee's constrained policy space and reliance on unconventional tools rather than rate adjustments. As an illustration, the dovish spikes in 2013 are consistent with the Committee's September 2013 decision not to begin tapering asset purchases, accompanied by guidance that maintained an extended low-rate stance. The December 2015 liftoff marks a clear regime shift, with increasing hike probabilities signaling tightening bias in the Minutes language. The 2022-2023 period exhibits sustained high hike probabilities during the aggressive tightening cycle, while the March 2020 spike captures emergency pandemic cuts.

Figure 4 presents the temporal evolution of forward guidance types across the full sample. The normalized stacked area chart decomposes each meeting into three components that sum to 100%: no guidance (gray), outlook-based guidance (blue), and commitment-based guidance (red). The 2008-2015 period shows a clear shift in the communication strategy, as the Fed relied increasingly on forward guidance when conventional policy reached its limits, with the

"no guidance" share dropping significantly. During this period, commitment-based guidance share increased to around a 50%. The post-2015 period shows a return to more traditional communication patterns as policy normalized. The post-COVID-19 period shows, for the first time in my sample, greater share of commitment-based guidance than outlok-based guidance, at least until 2024, where inflation started to approach the target.

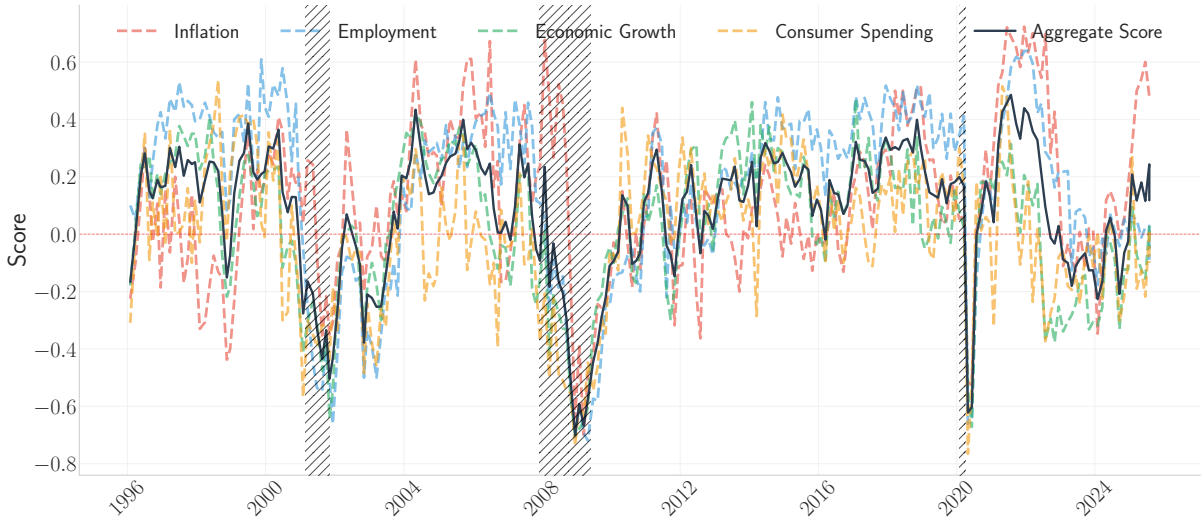
2.3 Agent IB

Agent IB operates on the Beige Book released approximately two weeks before each FOMC meeting, extracting quantitative assessments of economic conditions from the qualitative narratives across the twelve Federal Reserve districts. Like Agent IM, Agent IB produces both numerical outputs and textual extractions.

The core quantitative outputs comprise four economic condition scores $s_v \in [-1, 1]$ for variables $v \in \{\text{inflation, employment, economic growth, consumer spending}\}$, where negative values indicate weak or dovish-leaning conditions and positive values indicate strong or hawkish-leaning conditions. The agent identifies sentences describing economic conditions for each variable, assigns both a policy stance classification (hawkish, dovish, or neutral) and an intensity measure on a $[0, 1]$ scale, then aggregates across sentences to produce the four summary scores. The agent also computes a normalized weight vector $\mathbf{w} = (w_{\text{inf}}, w_{\text{emp}}, w_{\text{growth}}, w_{\text{cons}})$ with $\sum_v w_v = 1$ reflecting the relative emphasis each variable receives in the Beige Book discussion. These weights construct the weighted aggregate score $s_{\text{agg}} = \sum_v w_v s_v$ summarizing overall economic conditions.

The textual outputs comprise two types. *Policy stance analysis* provides sentence-level extractions showing the actual Beige Book text underlying each score, along with justifications for the policy stance classification and intensity assessment. For instance, an inflation score of $+0.6$ might reflect "widespread price increases across manufacturing and services" or "isolated cost pressures in specific sectors" yet represent fundamentally different economic conditions despite identical numerical scores. *Shock indicators* identify textual passages describing extreme economic conditions, sharp divergences from recent patterns, or unprecedented developments, extracted with relevance scores (high, medium, low) and magnitude assessments. Appendix A.2 provides complete output specifications with a detailed example from the March 2022 tightening cycle liftoff period, showing how the agent captures mixed economic signals through both quantitative scores and preserved textual context.

Figure 5: Beige Book Scores and Aggregate Index



Note: Individual variable scores (dashed lines) and weighted aggregate score (solid line) extracted from Beige Book text. Scores range from -1 (dovish/weak conditions) to +1 (hawkish/strong conditions). The aggregate score $s_{\text{agg}} = \sum_v w_v s_v$ uses time-varying weights reflecting each variable's emphasis in the Beige Book. Variables include inflation, employment, economic growth, and consumer spending. Recession periods are shaded in gray.

Agent IB implements automatic chunking to handle Beige Books exceeding standard LLM context windows: large documents are divided into overlapping segments at logical boundaries, processed independently, and merged into comprehensive sentence-level analysis. Figure 5 shows both individual variable scores and the weighted aggregate score over time.

The time series shows that scores track business cycle dynamics: the aggregate score declines sharply during recession periods (shaded), remains subdued during the zero lower bound era, and rises during expansions with particularly pronounced increases during 2021-2023 when both inflation and employment scores turned strongly positive. Individual variable scores exhibit differential behavior—during 2022, inflation scores reached extreme hawkish levels (+0.8 to +1.0) while growth scores remained moderate, reflecting the stagflationary environment. Correlations with conventional macroeconomic indicators confirm economic content: the aggregate score correlates negatively with unemployment ($\rho = -0.363$) and positively with GDP growth ($\rho = 0.642$), CPI inflation ($\rho = 0.242$), and PCE growth ($\rho = 0.192$).

2.4 Agent II

Agent II serves as the synthesis agent in the pipeline, integrating the dual-channel outputs from both Agent IM (Minutes-based policy intelligence) and Agent IB (Beige Book economic conditions) to form a coherent probability distribution for the upcoming FOMC decision. The agent's

critical function is to perform Bayesian updating: starting from Agent IM's policy distribution extracted from the previous meeting's Minutes, Agent II adjusts this baseline using Agent IB's current economic condition scores from the Beige Book released before meeting t, producing an updated prior that incorporates all Fed communications available at least two weeks before the decision. This synthesis task requires reconciling potentially conflicting signals—for instance, hawkish Committee deliberations in Minutes versus dovish economic conditions in the Beige Book—while maintaining the market-agnostic design principle that ensures priors reflect only Fed documents rather than market pricing that might be contaminated by non-monetary information flows (Bauer & Swanson, 2023a).

Agent II receives three distinct input channels, each contributing different information types to the synthesis process. From Agent IB, Agent II receives four economic condition scores $s_v \in [-1, 1]$ as numerical inputs, alongside textual reasoning that explains the economic developments underlying each score and identifies shock indicator passages flagging potential surprise sources. From Agent IM, Agent II receives the updated policy distribution from the previous meeting as a numerical baseline probability mass function, plus hawkishness and uncertainty scores as quantitative calibration signals, alongside textual extractions that include debate narratives, forward guidance language, and shock discovery insights. The third input comprises historical context—formatted text summarizing recent FOMC decisions and statements that establishes policy inertia patterns and the prevailing monetary stance trajectory.

The synthesis process operates through structured probabilistic reasoning that the agent's prompt explicitly formalizes. Agent II begins with Agent IM's baseline distribution, then sequentially adjusts probabilities based on directional signals from economic conditions (Beige Book scores) and contextual intelligence (Minutes narratives). The prompt instructs the agent to maintain conditional unbiasedness: given the framework $\Delta i_t = E[\Delta i_t|\mathcal{B}_t] + s_t$ where Δi_t represents the actual decision, $\mathcal{B}_t$ denotes the information set (Beige Book + Minutes + historical context), and s_t is the surprise, Agent II's objective is producing $E[\Delta i_t|\mathcal{B}_t]$ such that $E[s_t|\mathcal{B}_t] = 0$. This requires fully incorporating all directional content from input signals into the distribution rather than neutralizing information: If Beige Book scores are dovish and Minutes intelligence is hawkish, the distribution should reflect this mixed signal through appropriate probability spreads, not revert to a symmetric prior that ignores the information.

Agent II's output architecture mirrors the dual structure of Agents IM and IB: numerical probability distributions and scores alongside textual justifications. The core quantitative

output is a probability distribution $\{p_i\}$ over policy scenarios (m_i, d_i) with $\sum_i p_i = 1$, from which a probability-weighted expected rate change $\bar{r} = \sum_i m_i d_i p_i$ is calculated. The agent also produces a numerical confidence score $c \in [0, 1]$ quantifying certainty in the forecast and a categorical signal strength assessment (strong, moderate, weak) indicating the clarity of directional signals. The textual outputs comprise four components that provide transparency into Agent II's probabilistic reasoning. First, detailed natural language justifications explain how Agent II reconciled potentially conflicting information sources. Second, separate influence assessments quantify how much the Beige Book versus Minutes intelligence shifted probabilities, while uncertainty driver descriptions connect information characteristics to the distribution's spread. Third, an orthogonality check documents the chain-of-thought reasoning that ensures conditional unbiasedness. These textual outputs enable validation that the synthesis process follows economically sensible logic rather than opaque pattern matching.

Appendix A.3 presents complete output specifications with contrasting examples from the December 2008 financial crisis (aligned dovish signals producing concentrated distributions) and March 2022 tightening initiation (mixed signals generating dispersed distributions), demonstrating how the synthesis process adapts across monetary regimes while maintaining economically coherent reasoning grounded in specific Fed document content.

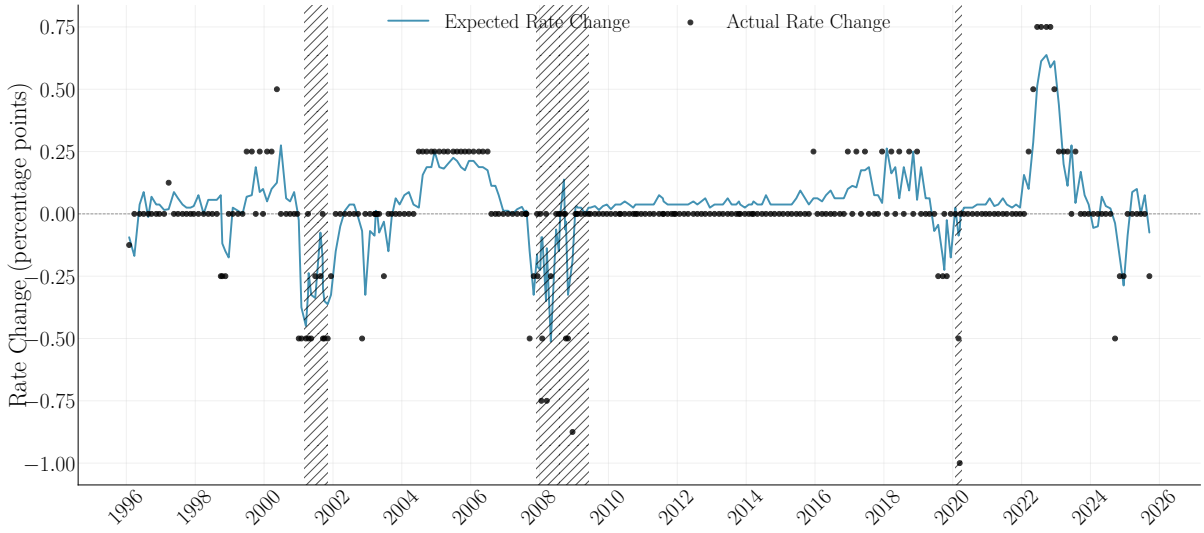
Figure 6 validates Agent II's synthesis performance. Panel (a) shows strong positive correlation between the probability-weighted expected rate change and actual FOMC decisions. Panel (b) reveals how the probability distribution evolves with monetary policy cycles: hike probabilities dominate during tightening periods, cut probabilities during easing periods, and hold probabilities increase during regime transitions as Agent II captures Committee uncertainty.

To sum up, Agent II synthesizes heterogeneous information from Fed documents into coherent probability distributions that track realized policy closely enough to serve as meaningful benchmarks for surprise identification, yet maintain sufficient uncertainty to avoid overfitting. The agent's outputs—numerical probability distribution, expected rate change, and textual justifications—complete the expectation formation stage. Agent III then compares this prior distribution against the actual FOMC announcement to quantify the surprise component.

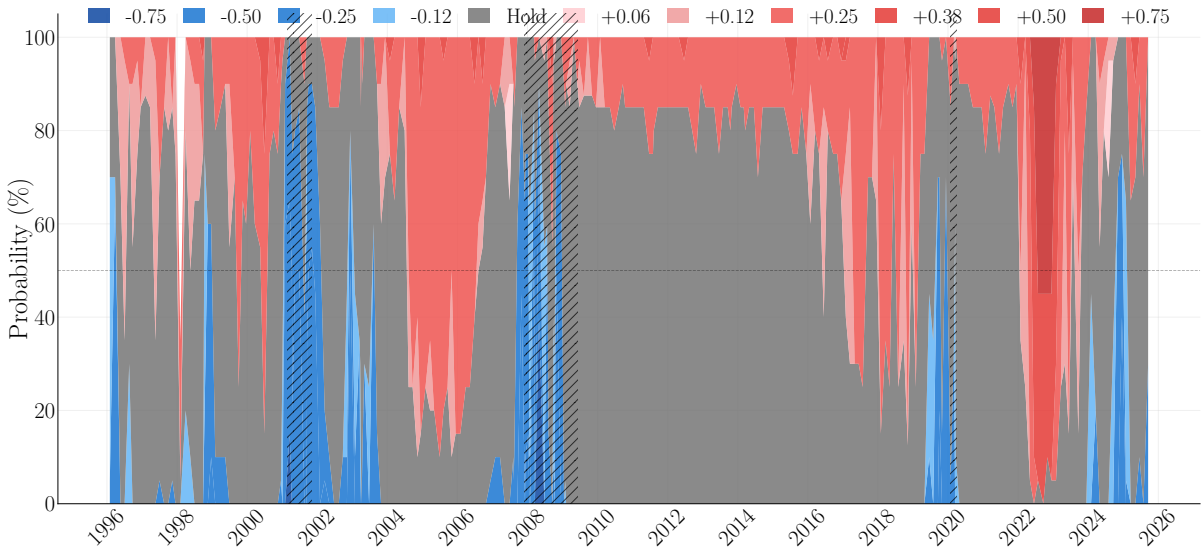
2.5 Agent III

Agent III quantifies the monetary policy surprise on FOMC announcement day. The agent receives Agent II's prior probability distribution and expected rate change, extracts the realized

Figure 6: Agent II Output Analysis



(a) Expected vs. Actual Rate Changes



(b) Evolution of Prior Probability Distribution

Note: Panel (a) compares the agent's probability-weighted expected rate change (solid line) with the actual FOMC decision (dots). Panel (b) shows the time series of the probability distribution for a rate hike (orange), hold (gray), and cut (blue). Recession periods are shaded in gray. Agent II synthesizes information from Agent IM (Minutes-based policy intelligence) and Agent IB (Beige Book economic conditions) to form these prior expectations approximately two weeks before each FOMC meeting.

policy decision from the FOMC Statement text, and computes the deviation between realized and expected outcomes. The agent produces three quantitative measures. The *surprise rate* r_s captures the mechanical baseline:

$$s = r_{\text{realized}} - r_{\text{expected}} = r_{\text{realized}} - \bar{r}$$

where r_{realized} is the announced rate change and $\bar{r} = \sum_i m_i d_i p_i$ is Agent II's probability-weighted expectation. The *surprise score* $\sigma_s \in [0, 1]$ provides contextual assessment by integrating two factors. First, the agent evaluates the prior probability mass assigned to the realized outcome: outcomes with higher prior probabilities receive lower surprise scores (e.g., 40% prior probability maps to scores around 0.2, while 5% maps to 0.8). Second, the agent examines recent surprise history to identify pattern effects: consecutive surprises in the same direction may warrant downward adjustments reflecting "surprise fatigue," while direction reversals may warrant upward adjustments as pattern-breaking moves. The *contextual salience* $\xi = s \times \sigma_s$ combines both measures. Large mechanical surprises with high contextual scores (rare outcomes) produce elevated salience, while large mechanical surprises with low contextual scores (high-probability outcomes) generate moderate salience.

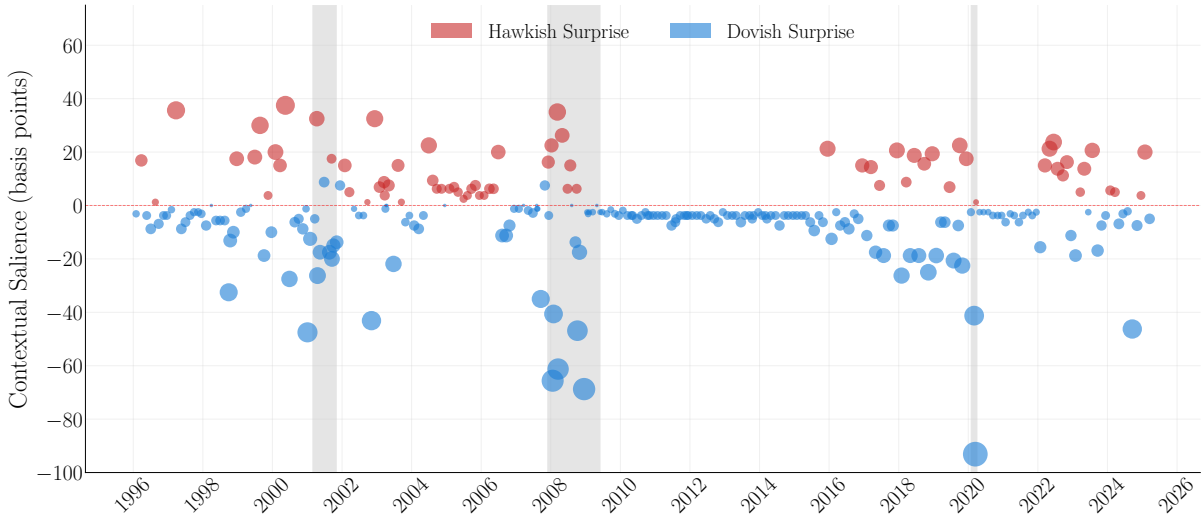
For comparative analysis against market-based measures (FF1–FF4, ED1–ED4, MP1) and the Romer & Romer (2004) series, I use the mechanical surprise rate s as the primary measure. This ensures methodological comparability: market-based surprises represent price changes over 30-minute announcement windows without contextual adjustments, while R&R constructs surprises as residuals from linear regressions. The contextual salience ξ and surprise score σ_s serve as supplementary measures that incorporate institutional knowledge about FOMC communication patterns, but would introduce heterogeneity in cross-measure comparisons.

The textual outputs document the decomposition logic. The agent extracts the actual FOMC Statement language announcing the rate decision, performs pattern analysis by comparing the current surprise against recent surprise history, assesses adaptation effects, and provides detailed justifications explaining how the prior probability distribution informed the surprise calculation. For unconventional policy dimensions—forward guidance shifts, balance sheet announcements—the agent identifies the tool type, assesses directional impact, and generates tool-specific justifications.

Appendix A.4 presents complete output specifications with examples from both the December 2008 crisis period (large dovish surprise with high contextual salience) and March 2022 tightening initiation (modest hawkish surprise with moderate salience), illustrating how the agent distinguishes between mechanical deviations and their economic significance.

Figure 7 plots the contextual salience metric over time. The highest-salience events cluster during policy regime transitions (2008 zero-lower-bound adoption, 2015 liftoff, 2022 tightening initiation) and pattern reversals (2019). Some of the largest mechanical surprises (2008

Figure 7: Contextual Salience of Fed Policy Surprises



Note: Bubble size represents the surprise score σ_s, which captures the contextual assessment of each surprise. The contextual salience metric $\xi = s \times \sigma_s$ combines mechanical magnitude with contextual importance and confidence. The largest salience values occur during regime transitions (2008 zero-lower-bound adoption, 2015 liftoff, 2022 tightening initiation) and pattern reversals (2019) rather than simply the largest mechanical deviations. Agent III computes these measures by comparing Agent II's prior distribution with realized FOMC decisions.

crisis cuts) generated only moderate salience due to adaptation effects, while smaller mechanical deviations during regime shifts produced elevated salience due to their pattern-breaking significance.

2.6 Validation and Robustness

The multi-agent architecture requires rigorous validation to address two critical challenges: preventing look-ahead bias where the system might incorporate information unavailable at the time of analysis, and ensuring output stability despite stochastic LLM inference. Through a combination of architectural constraints and empirical validation, I demonstrate that the narrative measures are both temporally valid and statistically stable.

The system implements multiple layers of look-ahead bias prevention, including strict document-level temporal cutoffs, explicit temporal anchoring in prompts, automated validation checks for forbidden future references, and out-of-knowledge-cutoff testing using meetings beyond the model's training window. Multi-run stability validation across 17 independent pipeline executions shows that the system produces consistent outputs with median cross-run standard deviations of only 4.2 basis points for expected rate changes and 5.4 basis points for surprise measurements. These magnitudes are economically negligible relative to typical Fed surprise

magnitudes of 25–100 basis points.

I find that meetings with elevated cross-run variability occur primarily during major regime transitions (2008 crisis onset, 2020 pandemic response) and out-of-sample periods, suggesting the system appropriately reflects genuine Fed communication ambiguity rather than introducing artificial measurement noise. The strongest validation comes from comparing narrative-based results against market-based benchmarks in Section 4, where systematic measurement issues would become apparent. Detailed validation procedures, including look-ahead bias tests, stability metrics across different monetary regimes, and complete multi-run analysis results, are provided in Appendix C.

Having established the architecture for constructing narrative surprises and validated its technical properties, the critical question becomes whether these measures improve upon existing identification strategies. If the multi-agent framework merely replicates the properties of market-based or narrative measures, it offers no methodological advance. I now examine the framework's empirical performance in two stages: first validating individual agent outputs to confirm the framework extracts policy-relevant information, then assessing the resulting surprise measures through measurement quality diagnostics, impulse response analysis, and out-of-sample trading performance.

3 Data

The analysis incorporates several external data sources for validation and comparison. The updated Romer and Romer (2004) series of monetary policy shocks is obtained from Acosta (2023).[1] Market-based surprise measures are obtained from Jarociński and Karadi (2020).[2] Fed Funds futures (FF1–FF4) are sourced from LSEG DataScope Tick History from 1996 onwards; pre-1996 data come from Gürkaynak et al. (2005a). The MP1 shock, constructed from FF1 and FF2 following Gürkaynak et al. (2005a), represents the policy surprise component orthogonal to the Fed's information effect as identified by Jarociński and Karadi (2020). Eurodollar futures (ED1–ED4) are sourced from TickData (until 2019) and LSEG DataScope Tick History (2019–2022); from January 2023 onwards, SOFR futures from LSEG DataScope Tick History are used. High-frequency data are aggregated to one-minute frequency. Surprises are computed

[1] Available at https://www.acostamiguel.com/data.html. Last accessed: October 2025.

[2] Available at https://github.com/marekjarocinski/jkshocks_update_fed_202401. Last accessed: October 2025. All high-frequency financial variables described below were collected by Jarocinski.

over a 30-minute window as the difference between the post-announcement value (median over [t+15min, t+25min), where t is the announcement time) and the pre-announcement value (median over (t-15min, t-5min]). When these windows contain fewer than three observations, they are extended up to 24 hours to ensure robustness; missing values are recorded if insufficient observations remain.

Cleaned monetary policy shocks are obtained from two sources. Bauer and Swanson (2023a) provide ex-post cleaned surprises constructed by regressing raw market-based surprises on contemporaneous macroeconomic data releases to remove the Fed information component.[3] Miranda-Agrippino and Ricco (2021)provide Greenbook-based cleaned instruments by regressing raw surprises on internal Fed staff forecasts to isolate pure monetary policy shocks.[4] Both cleaned measures are provided as monthly series and are matched to FOMC meeting months for clean identification in the impulse response analysis. Additional macroeconomic and financial data are sourced from FRED, the Federal Reserve Board, and standard databases as detailed in the relevant subsections.

4 Results

4.1 Stage 1: Agent IB Predictive Power

Having validated the multi-agent framework, I examine the predictive power of Agent IB's Beige Book scores for monetary policy decisions. Table 1 presents specifications estimating

$$\Delta i_t = \alpha + \rho i_{t-1} + \sum_j \beta_j \text{BB}_t^j + \varepsilon_t \tag{1}$$

where Δi_t denotes the rate change and BB_t^j represents Beige Book scores for variable j.[5] Beige Book scores are available approximately two weeks before each meeting, making this a predictive regression.

Table 1 analyzes 265 FOMC meetings (1996-2025). Policy inertia alone explains virtually nothing ($R^2 = 0.004$), with an insignificant lagged rate coefficient near zero. The weighted Beige

[3] Available at https://www.michaeldbauer.com/files/monetary_policy_surprises_data.xlsx. The monthly series covers 431 observations from February 1988 to December 2023, providing 298 FOMC meeting months in the analysis sample.

[4] Available in the replication materials for their AER publication. The monthly series covers 228 observations from January 1991 to December 2009, providing 159 FOMC meeting months. Data availability is limited to the Greenbook confidentiality period; Greenbooks after 2009 were not yet publicly released at the time of their analysis.

[5] When the FOMC sets a target range, I use the midpoint for calculations.

Table 1: From Inertia to Beige Book News: Sequential Addition of Beige Book Variables

	(1)	(2)	(3)	(4)	(5)
i_{t-1}	-0.006	-0.008	-0.014**	-0.011*	-0.011*
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
$BB^{agg.}$	—	0.327***	—	—	—
		(0.054)			
$BB^{empl.}$	—	—	0.252***	0.468***	0.666**
			(0.040)	(0.133)	(0.267)
BB^{growth}	—	—	—	0.122	0.255
				(0.123)	(0.258)
$BB^{infl.}$	—	—	—	—	0.311
					(0.221)
$BB^{cons.\ spend.}$	—	—	—	—	-0.035
					(0.319)
R^2	0.004	0.126	0.135	0.133	0.135
Adj. R^2	-0.000	0.119	0.129	0.123	0.118
% of Full Model R^2	2.6%	93.6%	100.4%	99.0%	100.0%
Obs.	265.0	265.0	265.0	265.0	265.0

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

Book aggregate achieves $R^2 = 0.125$, capturing 94% of the full model's explanatory power. Employment emerges as the dominant predictor ($R^2 = 0.134$), with a coefficient of 0.252 (significant at 1%), indicating that a one-unit increase predicts 25 basis points of tightening. Adding economic growth barely improves fit ($R^2 = 0.132$); employment remains highly significant (0.462) while growth (0.127) does not. The full specification reaches $R^2 = 0.134$, with inflation (0.316) and consumer spending (-0.035) both insignificant. This pattern suggests that while the Fed monitors inflation continuously, employment discussions signal marginal information that drives policy regime changes.

Agent IB's text-based extraction successfully identifies the Fed's revealed preferences: employment dominates policy responses, consistent with the dual mandate. Variance inflation factors remain below standard thresholds (Appendix D), confirming the results are not driven by multicollinearity.

4.2 Stage 2: Agent II Synthesis Performance

Having established Agent IB's contribution, I examine whether Agent II's synthesis adds value. While Beige Book scores alone explain 13.4% of policy variation (Table 1, column 3), Agent

Table 2: Agent II Statistical Moments and Monetary Policy

	(1)	(2)	(3)	(4)	(5)	(6)
i_{t-1}	-0.006	-0.008	0.000	-0.008	-0.013**	-0.001
	(0.006)	(0.006)	(0.005)	(0.006)	(0.006)	(0.005)
BB^{agg}	—	0.326***	0.072	0.326***	0.317***	0.078*
		(0.054)	(0.044)	(0.055)	(0.054)	(0.045)
$E[\Delta i_t]$	—	—	0.988***	—	—	1.008***
			(0.070)			(0.072)
$\sigma[\Delta i_t]$	—	—	—	0.009	—	0.497*
				(0.375)		(0.294)
$Skew[\Delta i_t]$	—	—	—	—	-0.028**	0.004
					(0.012)	(0.010)
R^2	0.003	0.125	0.507	0.125	0.143	0.513
Obs.	264	264	264	264	264	264

Note: This table shows how different statistical moments of Agent II's probability distribution contribute to predicting Federal Funds Rate changes. Column (1) includes only policy inertia. Column (2) adds the Beige Book aggregate score. Columns (3-5) add the 1st moment (expected rate change), 2nd moment (variance/uncertainty), and 3rd moment (skewness/asymmetric risk) respectively. Standard errors in parentheses. ***, **, and * denote significance at 1%, 5%, and 10% levels.

II synthesizes these scores with Agent IM's Minutes-based policy intelligence to form complete probability distributions. Table 2 presents the results.

Table 2 reveals the value of probabilistic synthesis. Augmenting inertia with the Beige Book aggregate yields $R^2 = 0.124$. Incorporating Agent II's expected rate change (the probability-weighted mean synthesizing Beige Book conditions with Minutes intelligence) dramatically improves R^2 to 0.501, quadrupling explanatory power. Variance and skewness individually achieve $R^2 = 0.124$ and 0.142, though variance adds nothing beyond the Beige Book aggregate. The full specification with all moments reaches $R^2 = 0.506$, confirming that the expected rate change captures most predictive content.

The synthesis integrates heterogeneous Fed communications, quadrupling explanatory power from 12.4% to 50.1%. Agent II's expectations are properly calibrated: the expected rate change coefficient of 0.986 (s.e. = 0.070) is statistically indistinguishable from unity, confirming that 25 basis-point expectations translate to 25 basis-point realizations. This unbiased forecasting, combined with $R^2 = 0.501$, validates both predictive power and rationality.

4.3 Surprise (Un-)predictability

Valid instruments require orthogonality to available information. Table 3 regresses narrative and market-based surprises on six B&S predictors: nonfarm payroll surprises, 12-month employ-

ment growth, 3-month S&P 500 changes, term spread, commodity prices, and Treasury market skewness. The analysis covers 223 FOMC meetings (1996-2023) using meeting-level data for precise temporal alignment.

Table 3: Predictability of FOMC Meeting Surprises

	Narrative		Market-Based			
Variable	R&R	My Surprise	FF1	FF4	ED1	ED4
Comm. Index (3m)	-0.007	0.018*	-0.000	0.003	0.004	0.010**
	(0.020)	(0.011)	(0.002)	(0.003)	(0.004)	(0.005)
Nonf. Payrolls (12m)	0.013	0.016*	0.000	0.002	0.004*	0.010**
	(0.028)	(0.009)	(0.001)	(0.002)	(0.002)	(0.004)
Term Spread (3m)	-0.054***	-0.016*	-0.003	-0.010***	-0.008**	-0.010**
	(0.014)	(0.009)	(0.002)	(0.003)	(0.004)	(0.004)
S&P 500 (3m)	-0.012	0.043***	0.006	0.012***	0.012**	0.016***
	(0.016)	(0.015)	(0.004)	(0.005)	(0.005)	(0.005)
Treasury Skewness	0.053***	0.016	0.005*	0.006***	0.006**	0.010***
	(0.012)	(0.011)	(0.003)	(0.002)	(0.003)	(0.003)
R^2	0.199	0.161	0.053	0.144	0.109	0.189
Observations	184	223	230	230	230	231

Note: This table presents predictability regressions of various surprise measures on a set of macroeconomic and financial predictors from Bauer and Swanson (2023a). The predictors include Nonf. Payrolls (12m), S&P 500 (3m), Term Spread (3m), Comm. 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.

My narrative surprise exhibits moderate predictability ($R^2 = 0.164$), statistically indistinguishable from market-based measures (5.4-19.8%) and R&R (20.3%). A variance decomposition reveals the source: B&S predictors explain 35.8% of expectation variance but only 16.4% of surprise variance (Appendix C.5). The key finding involves S&P 500 returns, which account for 72% of total surprise predictability. These returns do *not* significantly predict expectations (loading: 0.0003, insignificant) but strongly predict surprises (loading: 0.0411, significant).

This pattern identifies the mechanism: equity market predictability arises from information arriving during the 2-3 week blackout period between Beige Book release and FOMC decisions, not from missing Fed documents. The conditional expectation $E[\Delta i_t|\mathcal{B}_t]$, frozen at Beige Book release, cannot incorporate subsequent market movements. This timing mechanism cannot explain market-based surprise predictability, since those surprises are computed in short windows around announcements. The moderate predictability validates the direct extraction methodology—accepting a predetermined information set rather than applying *ex post* econometric cleaning. While a news-reading agent processing blackout-period information could likely re-

Table 4: Incremental Explanatory Power of Surprise Measures

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Surprise	1.020***	0.730***	0.687***	0.692***	0.681***	0.686***	0.672***
	(0.083)	(0.119)	(0.088)	(0.091)	(0.087)	(0.091)	(0.091)
R&R (2004)		0.507***	0.471***	0.473***	0.462***	0.461***	0.470***
		(0.096)	(0.093)	(0.092)	(0.091)	(0.091)	(0.092)
FF4			0.860***	0.939***	1.855***	2.184***	1.709**
			(0.333)	(0.357)	(0.529)	(0.759)	(0.731)
FF1				-0.251	2.122	2.182	2.030
				(0.520)	(1.390)	(1.358)	(1.391)
MP1					-1.995**	-1.941**	-1.580
					(0.953)	(0.980)	(1.002)
ED1						-0.398	-0.608
						(0.598)	(0.483)
ED4							0.395**
							(0.157)
R^2	0.520	0.649	0.676	0.676	0.700	0.702	0.708
Observations	272	178	178	178	178	178	178

Note: This table shows incremental specifications where variables are added sequentially. Column (1) includes only the narrative surprise measure. Each subsequent column adds one additional variable to the regression. The regression form is $\Delta i_t = \alpha + \sum_j \beta_j \hat{s}_t^j + \varepsilon_t$. Newey-West HAC standard errors (4 lags) are in parentheses. Significance stars test H_0: $\beta = 0$ (no explanatory power). ***, **, and * denote significance at the 1%, 5%, and 10% levels. Time window: 1996-01 to 2025-09.

duce this predictability further, I leave this extension for future research.

4.4 Surprise Diagnostics

Having established that my surprises reflect information unavailable from public Fed communications, I test whether they constitute valid instruments in the classical measurement-error framework. Surprises are typically conceived relative to the Fed's private information set $\mathcal{G}_t$ (honoring Greenbooks), yielding

$$\Delta i_t = \mathbb{E}[\Delta i_t \mid \mathcal{G}_t] + s_t \tag{2}$$

where s_t is the true monetary policy shock. Since we observe only the public subset $\mathcal{B}_t \subset \mathcal{G}_t$, the observed surprise is

$$\Delta i_t = \mathbb{E}[\Delta i_t \mid \mathcal{B}_t] + (\mathbb{E}[\Delta i_t \mid \mathcal{G}_t] - \mathbb{E}[\Delta i_t \mid \mathcal{B}_t]) + s_t \tag{3}$$

Table 5: Measurement Validity: Raw and Cleaned Surprise Measures

Panel A: Raw Surprise Measures							
	My Surprise	R&R (2004)	FF4	FF1	MP1	ED1	ED4
Coefficient	1.020	0.783**	1.827**	2.375***	1.410	1.683*	1.321
$H_0 : \beta = 1$	(0.083)	(0.090)	(0.359)	(0.485)	(0.267)	(0.359)	(0.309)
R^2	0.520	0.432	0.168	0.158	0.148	0.162	0.144
First-stage F	19.6	41.3	18.8	35.3	21.2	18.8	17.9
Observations	272	178	221	221	221	221	222
Panel B: Cleaned Market Surprises (LLM Moments)							
	My Surprise	R&R (2004)	FF4	FF1	MP1	ED1	ED4
Coefficient	—	0.494***	0.855	1.565	0.874	0.908	0.506
$H_0 : \beta = 1$	—	(0.155)	(0.453)	(0.458)	(0.297)	(0.375)	(0.317)
R^2	—	0.126	0.033	0.065	0.053	0.043	0.019
First-stage F	—	1.0	9.9	17.1	8.4	13.8	7.8
Observations	—	178	221	221	221	221	222

Note: This table tests measurement validity by regressing FOMC policy decisions (rate changes in basis points) on individual surprise measures. Each column represents a separate regression of the form $\Delta i_t = \alpha + \beta \hat{s}_t + \varepsilon_t$. Panel A shows raw surprise measures. Panel B shows cleaned market surprises (residuals from regressing raw surprises on LLM distribution moments: $E[\Delta i]$, Variance, Skewness). My Surprise is already clean by construction (extracted directly from Fed communications 2-3 weeks before meetings), so Panel B shows "—" for this column. First-stage F-statistic is from regressing VAR(12) residual for 1-year Treasury yield on the instrument (Miranda-Agrippino & Ricco, 2021 methodology). Stock and Yogo (2005) critical values: F > 16.38 (strong instrument), F > 8.96 (marginal). High F validates instrument strength for LP-IV; low F indicates weak instrument problem with wide confidence bands. Newey-West HAC standard errors (4 lags) are in parentheses. Significance stars test H_0: $\beta = 1$ (no measurement error). Under classical measurement error, $\beta = \text{Var}(s_{true})/[\text{Var}(s_{true}) + \text{Var}(\varepsilon)]$: $\beta \approx 1$ indicates minimal noise, $\beta < 1$ suggests attenuation bias (noisy signal), $\beta > 1$ suggests contamination (measure inflates true variation). ***, **, and * denote significance at the 1%, 5%, and 10% levels. Time window: 1996-01 to 2025-09.

Rewriting (3) yields

$$\Delta i_t = \mathbb{E}[\Delta i_t \mid \mathcal{B}_t] + \hat{s}_t \tag{4}$$

where $\hat{s}_t = s_t + \varepsilon_t$ contains the true shock plus measurement error. To assess whether my narrative surprises measure true policy shocks with minimal error, I estimate

$$\Delta i_t = \alpha + \beta \hat{s}_t + \eta_t \tag{5}$$

Under classical measurement error with $\text{Cov}(s_t, \varepsilon_t) = 0$, the coefficient

$$\beta = \frac{\text{Var}(s_t)}{\text{Var}(s_t) + \text{Var}(\varepsilon_t)} \tag{6}$$

provides a validity test: $\beta \approx 1$ indicates minimal measurement error, while deviations suggest attenuation bias ($\beta < 1$) or contamination ($\beta > 1$). For valid measures with $\beta \approx 1$, the R^2 represents the fraction of policy variance attributable to surprises.

Table 5 tests measurement quality. Panel A shows that my narrative surprise passes the $\beta = 1$ test, validating minimal measurement noise. R&R exhibits significant attenuation ($\beta = 0.783$), while market-based measures show coefficients exceeding unity, indicating contamination. Panel B applies *ex post* cleaning procedures analogous to B&S (using public macroeconomic information) and M-A&R (using Greenbook forecasts). I regress each raw surprise on my LLM-extracted distribution moments ($E[\Delta i]$, variance, skewness) and use residuals as cleaned instruments. While cleaning removes contamination (coefficients near unity), first-stage F-statistics from M-A&R's external instrument VAR approach reveal the cost: instrument strength collapses (FF4: F falls from 18.8 to 9.9; R&R: from 41.3 to 1.0). Direct narrative extraction thus dominates *ex post* cleaning by maintaining both minimal measurement error and strong predictive power.

4.4.1 Narrative Surprise Comparison: My LLM Approach vs R&R

Table 4 examines incremental explanatory power. Combining my narrative surprise with R&R yields both measures highly significant, indicating that Greenbook forecasts capture complementary private Fed information. Adding all market-based measures increases R^2 by only 5.9 percentage points beyond this baseline, reflecting distinct information timing: market surprises capture 30-minute pricing errors, R&R uses Greenbook forecasts with five-year lags, while my measure uses Fed communications released weeks before decisions. Most policy information flows through official channels well before announcements, consistent with Lucca and Moench (2015).

I compare my approach with R&R during the Zero Lower Bound period (2009-2015) when conventional rate changes ceased, providing a critical test of measurement properties.

Table 6 reveals stark distributional differences: R&R exhibits persistent hawkish bias (in a period that was characterised more by dovish stance) with near-symmetric distribution (skewness: 0.47), classifying 58.8% of meetings as hawkish without recording a single zero. My LLM measure shows slight dovish tilt with strong left skewness (-2.46) and fat tails (kurtosis: 21.51), identifying only 25.5% as hawkish with 13.7% exact zeros. This left-skewed distribution, characterized by mostly small hawkish surprises with occasional large dovish outliers, aligns with

Table 6: Statistical Moments Comparison During ZLB Period

Measure	N	Mean	Std Dev	Skewness	Kurtosis	Min	Median	Max	Zero%	Pos%
My Surprise	51	-3.7	3.5	5.17	33.67	-9.4	-3.8	18.8	2.0	2.0
R&R (2004)	51	3.9	9.4	0.47	-0.30	-12.7	3.8	28.7	0.0	58.8
B&S (2023)	51	0.5	2.8	0.32	0.08	-5.5	0.1	6.6	0.0	51.0

Note: All statistics computed in basis points. Skewness and kurtosis are sample statistics. Zero% = percentage of zero observations, Pos% = percentage of positive observations.

maintained accommodation during the ZLB era.

Table 7 examines overlap. Full-sample correlation is negligible (0.032). Trimming extreme observations yields moderate positive correlation (0.422), suggesting agreement on typical surprises but divergence on tail events. Excluding key regime transitions (2009-01, 2015-12) produces weak negative correlation (-0.128), indicating fundamental disagreement concentrates in non-transition periods while the measures show relative agreement during major policy regime shifts.

High-frequency market-based measures face analogous difficulties during the ZLB period. When conventional policy tools are suspended, 30-minute announcement windows capture primarily noise in Treasury and futures markets rather than systematic policy signals, as evidenced by dispersed distributions centered near zero. This measurement breakdown during unconventional policy episodes motivates the narrative approach, which maintains informativeness by processing qualitative forward guidance and balance sheet signals that markets price gradually rather than instantaneously.

The distributional differences reflect these distinct extraction methodologies. R&R's linear residuals impose additive separability—appropriate for conventional policy regimes where target rate changes dominate. During the ZLB period, however, policy operates through multiple channels (forward guidance, balance sheet operations...) with complex interactions that linear projections cannot fully capture. Survey-elicitation through LLMs processes these interactions nonparametrically: the system synthesizes economic condition and Fed's communication without requiring *ex ante* functional form assumptions.

The resulting measurement differences manifest in validity tests: my approach yields $\beta = 1.007$ while R&R exhibits attenuation ($\beta = 0.768$). This pattern suggests that Greenbooks contain valuable private information that parametric linear extraction leaves partially unutilized. A natural extension would apply survey-elicitation methods to Greenbook data—potentially

Table 7: Correlation Robustness Analysis During ZLB Period

Sample	N	Correlation	P-Value	Description
Full Sample	51	0.322**	0.021	All ZLB period observations
Trimmed (1%-99%)	47	0.088	0.556	Excluding extreme percentiles
No Key Episodes	49	0.110	0.451	Excluding 2009-01, 2015-12

Note: ** indicates significance at 5% level. The correlation between My Surprise and R&R measures during the Zero Lower Bound period (2009-2015) is shown for different sample specifications.

achieving both the informational richness of internal Fed forecasts and the measurement precision of nonparametric extraction. That R&R's measure adds 12.8pp of explanatory power beyond mine (Table 4) confirms Greenbooks' incremental information content, even when extracted through linear methods.

4.4.2 External Validation: Sign-Restriction Decomposition

I validate my measure using J&K's sign-restriction identified shocks, which distinguish pure monetary policy shocks (MP: yields rise, stocks fall) from central bank information shocks (CBI: both rise). This tests whether my measure isolates policy stance shifts or remains contaminated.

Table 8 regresses three surprise measures on J&K's orthogonal shocks. If *ex post* cleaned shocks (M-A&R, B&S) measured pure monetary policy, they should load only on β_{MP}, not β_{CBI}. Yet both load significantly on both components, indicating contamination by central bank information effects. Their β_{MP} coefficients also differ significantly from unity, suggesting attenuation.

My narrative surprise passes both validation tests at the 5% level (loading on CBI only at 10%). The low R^2 reflects timing differences: my surprise uses Fed communications from 2-3 weeks before decisions, while J&K uses 30-minute announcement windows. When forward guidance shapes expectations effectively, these information sets converge, producing low correlation despite capturing identical policy shifts.

4.4.3 Interpretation and Implications

The measurement validity tests establish a clear hierarchy: my narrative surprise passes the $\beta = 1$ test ($\beta = 1.007$), R&R exhibits significant attenuation ($\beta = 0.768$), and market-based

Table 8: J&K Decomposition: Testing for CB Information Contamination

	M-A&R	B&S	My Surprise
β_{MP}	0.499***	0.740***	0.828**
	(0.064)	(0.088)	(0.381)
β_{CBI}	0.595***	0.584***	1.308*
	(0.181)	(0.177)	(0.669)
R^2	0.508	0.662	0.111
Observations	161	281	221

Note: This table regresses surprise measures on Jarociński and Karadi (2020) sign-restriction identified shocks without a constant: $\text{Surprise}_t = \beta_{\text{MP}} \cdot \text{MP}_t + \beta_{\text{CBI}} \cdot \text{CBI}_t + \varepsilon_t$. MP is the pure monetary policy shock (contractionary policy: Treasury yields rise, stock prices fall). CBI is the central bank information shock, where the Fed reveals positive news about the economy (both Treasury yields and stock prices rise together, reflecting improved growth expectations without policy tightening). M-A&R refers to Miranda-Agrippino and Ricco (2021) ex-post VAR-cleaned instrument. B&S refers to Bauer and Swanson (2023a) orthogonalized surprise. My Surprise refers to the LLM-extracted surprise from Fed communications. Pure MP shock isolation requires $\beta_{\text{MP}} \approx 1$ (captures pure policy shock) and $\beta_{\text{CBI}} \approx 0$ (no Fed information effect contamination). All instruments aggregated to monthly frequency. Newey-West HAC standard errors (Newey & West, 1987), 6 lags in parentheses. ***, **, and * denote significance at the 1%, 5%, and 10% levels.

measures display contamination ($\beta > 1.3$). These differences reflect methodologies, not merely timing or data sources.

My narrative measure achieves minimal measurement error through multi-agent non-linear extraction from Fed communications (Beige Book → Minutes → Statement). R&R's attenuation stems from linear regression residuals that impose additive separability, failing during unconventional policy. Crucially, while R&R extracts information from private Greenbooks through linear OLS projections, my non-linear extraction from public documents achieves superior measurement ($\beta = 1.007$ vs 0.768). The ZLB evidence confirms this methodological advantage: R&R shows persistent hawkish bias while my approach correctly identifies the asymmetric pattern of mostly small hawkish surprises with occasional large dovish signals. That R&R still adds 12.8pp when combined with my measure suggests Greenbooks contain valuable private information poorly extracted by linear methods—information that non-linear extraction could potentially capture more effectively.

Market-based contamination ($\beta > 1.3$) arises from information effects. The J&K validation confirms this: *ex post* cleaned market surprises load significantly on both monetary policy and central bank information shocks, with coefficients differing from unity. My narrative surprise passes both tests at 5%, loading on pure policy shocks while remaining orthogonal to information effects. Market measures add only 6.0pp beyond my narrative-R&R baseline, capturing announcement-day pricing errors rather than weeks of Fed signals.

My measure's moderate predictability ($R^2 = 0.164$) has a clear source. B&S predictors explain 35.8% of expectations but only 16.4% of surprises. S&P 500 returns predict surprises but not expectations, confirming predictability arises from blackout-period information arriving between Beige Book release and decisions—not systematic bias. This timing mechanism cannot explain market-based predictability in short announcement windows.

The implications are straightforward: direct extraction from narrative sources passes the $\beta = 1$ test using only public information, enabling real-time implementation. The sequential revelation pattern (52.1% through communications, 12.8pp through Greenbooks, 6.0pp through announcement-day pricing) validates pre-FOMC drift (Lucca & Moench, 2015) and confirms most policy information flows through official channels well before decisions. For real-time analysis, direct extraction of conditional expectations dominates *ex post* cleaning in both measurement properties and feasibility.

4.5 Impulse Responses to Monetary Policy Surprises

To evaluate whether narrative surprises identify genuine monetary policy shocks, I examine their dynamic effects on macroeconomic and financial variables. I trace these effects using local projections (Jordà, 2005) at each forecast horizon h:

$$y_{t+h} = \alpha_h + \beta_h \cdot \text{Surprise}_t + \sum_{j=1}^{L_y} \gamma_{h,j} y_{t-j} + \sum_{k=1}^{L_x} \boldsymbol{\delta}_{h,k}^{\top} \mathbf{X}_{t-k} + \varepsilon_{t+h} \tag{7}$$

where y_{t+h} is the outcome variable at horizon h, Surprise_t is the monetary policy surprise measure, and lag lengths L_y and L_x vary by outcome type. Shock lags are set to zero since autocorrelation in monthly aggregated surprises is mechanical rather than reflecting multi-period Fed plans. Standard errors are computed using Newey-West HAC estimators to account for residual autocorrelation. All impulse responses are normalised to represent effects of a 25 basis point monetary policy surprise (Jordà & Taylor, 2025; Ramey, 2016).

4.5.1 Methodological Validation: Comparison with Established Benchmarks

Before presenting full-sample transmission dynamics, I validate my narrative surprise measure against established identification strategies[6]. This validation serves two purposes: (1) demon-

[6]The comparisons use OLS local projections rather than instrumental variables for two reasons. First, the constrained pre-crisis sample (108 observations) yields weak first-stage correlations between all shock measures and the federal funds rate, producing unreliable IV estimates with severely inflated standard errors. Second,

strating that my LLM-based extraction performs competitively against proven methods on their native samples, and (2) establishing methodological credibility before leveraging my measure's unique advantages (real-time implementability and extended temporal coverage).

Text-Based Narrative Methods Figure 8 compares my measure with two established narrative approaches: the ridge-regularized textual analysis of Aruoba and Drechsel (2024) and the OLS-residualized Greenbook forecasts of Romer and Romer (2004). To ensure a rigorous and transparent comparison, I replicate Aruoba and Drechsel (2024)'s exact specification using their publicly available replication package: 4 lags of control variables (industrial production, CPI, unemployment, excess bond premium, S&P 500, federal funds rate), 0 shock lags, and their processed dataset. This direct replication of a very recent method on identical data ensures that any performance differences reflect measurement properties rather than specification choices. The common sample spans 108 monthly observations (1996:01–2008:10), representing the pre-crisis conventional policy regime.

Federal funds rate responses are similar across all three measures: initial rise followed by gradual mean reversion. My measure generates the most pronounced unemployment response, rising to +0.4 to +0.6pp around months 15-20 with 90% confidence bands excluding zero at peak horizons—a statistically significant effect absent in A&D and R&R.

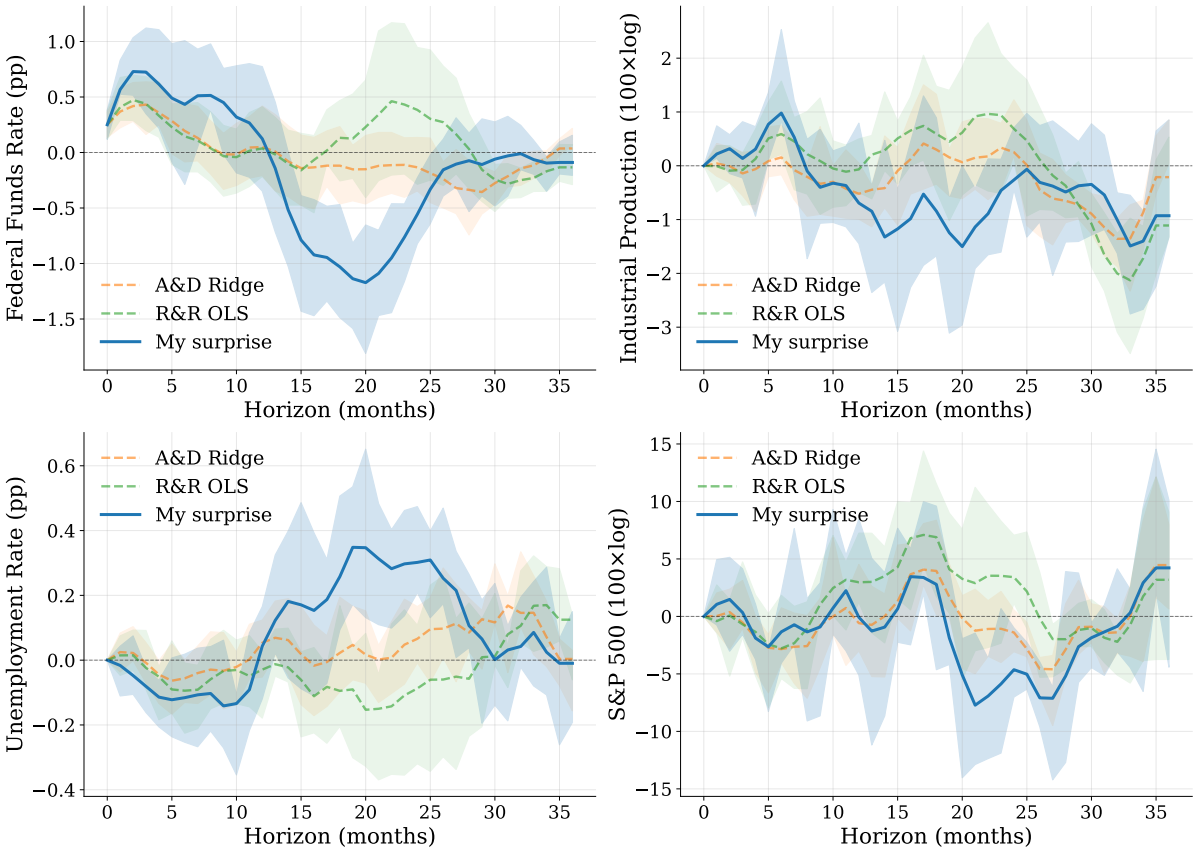
Industrial production shows initial modest increases (potentially composition effects), followed by significant negative effects after month 20 (-1 to -2%, bands excluding zero). S&P 500 responses display coherent timing: my measure shows significant declines at months 20-25, coinciding precisely with unemployment rises and production declines, while A&D and R&R show muted equity responses.

These synchronized real effects (rising unemployment, declining production, falling equity prices at months 20-25) suggest my measure captures coherent contractionary transmission more clearly than alternative narrative approaches, though the limited sample (108 observations) constrains statistical power.

Cleaned Market-Based Surprises Figure 9 compares my measure with *ex post* cleaned market-based approaches: Bauer and Swanson (2023b)'s orthogonalized high-frequency sur-

OLS comparisons provide transparent reduced-form evidence that each shock measure predicts policy-relevant outcomes, establishing baseline measurement validity before examining structural transmission channels. The full-sample analysis presented next leverages my measure's extended coverage (269 meetings) where IV identification becomes feasible.

Figure 8: IRF Comparison: Text-Based Narrative Methods (1996–2008)

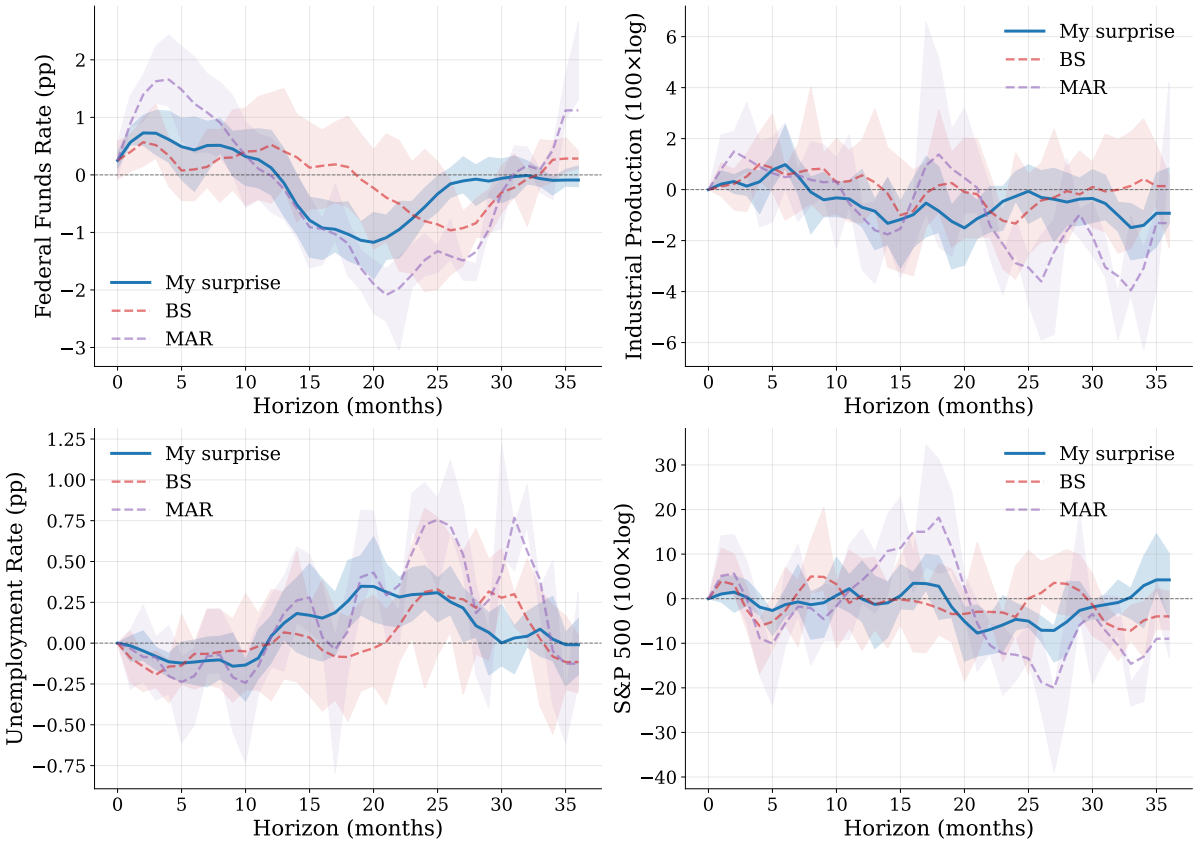


Note: Impulse response functions to a 25 bp monetary policy shock on the common pre-crisis sample. Narrative surprise: multi-agent LLM synthesis from FOMC communications (Beige Book, Minutes, Statements). A&D Ridge: ridge-regularized textual analysis (Aruoba & Drechsel, 2024). R&R OLS: Greenbook-based residuals (Romer & Romer, 2004). Specification: OLS local projections with 4 lags of controls (IP, CPI, UE, EBP, SP500, FFR), 0 shock lags. 3-period moving average smoothing applied to all measures (h=0 preserved). Newey-West HAC standard errors with bandwidth h+1. Shaded areas: 90% confidence bands. Sample: 108 months (1996:01–2008:10), pre-crisis conventional policy regime.

prises (BS) and Miranda-Agrippino and Ricco (2021)'s monetary policy indicator (MAR). My measure exhibits the patterns from Figure 8. BS produces muted responses with wider confidence bands, indicating weaker identification. MAR displays extreme volatility with bands exceeding ±20-30pp for unemployment and S&P 500, potentially reflecting specification differences between their VAR framework and this LP implementation.

Both comparisons establish that my direct narrative extraction from public Fed communications performs competitively with—and in some dimensions superior to—narrative methods using private Greenbook forecasts and market-based measures requiring *ex post* cleaning. Critically, this performance is achieved while maintaining real-time implementability—an advantage neither Greenbook-based nor *ex post* cleaned measures can offer.

Figure 9: IRF Comparison: Cleaned Market Surprises (1996–2008)



Note: Impulse response functions to a 25 bp monetary policy shock. Narrative surprise: direct narrative extraction (real-time implementable). BS: Bauer and Swanson (2023b) orthogonalized high-frequency surprises (*ex post* cleaned, extends to 2023). MAR: Miranda-Agrippino and Ricco (2021) monetary policy indicator (*ex post* cleaned, ends 2008). Specification: identical to Figure 8 (OLS local projections, 4 lags of controls, 0 shock lags, 3-period MA smoothing with h=0 preserved, Newey-West HAC standard errors with bandwidth h+1, 90% confidence bands) on the same 108-month sample (1996:01–2008:10) to ensure comparability across all shock measures. Sample constraints reflect data availability for benchmark measures; the narrative measure extends to 2025.

4.5.2 Full-Sample Transmission Dynamics

Having established methodological credibility against established benchmarks on the original samples, I now leverage my measure's unique advantage: the ability to trace monetary transmission across 30 years of data (269 FOMC meetings, 1996–2025) including zero lower bound and unconventional policy episodes that previous measures cannot capture. This extended coverage increases statistical power by 149% relative to the pre-crisis comparisons while capturing policy regime heterogeneity essential for understanding modern monetary transmission.

For the full-sample analysis, I use a parsimonious lag specification with $L_y = L_x = 4$ for all outcome and control variables, following Aruoba and Drechsel (2024). This balances flexibility in capturing dynamics against overfitting concerns given the meeting-month sample.

The control vector $\mathbf{X}_{t-k}$ includes the federal funds rate, unemployment rate, log PCE Price Index, log industrial production, S&P 500, and excess bond premium (a 6-variable augmented specification). Shock lags are set to zero since autocorrelation in monthly aggregated surprises is mechanical. Standard errors use Newey-West HAC estimators with bandwidth $h+1$ following Jordà (2005). The sample includes only FOMC meeting months for clean identification (269 meetings, 1996–2025).

Macroeconomic Transmission Figure 10 presents the core transmission mechanism from my narrative surprise measure. The baseline specification uses two-stage local projections to instrument the federal funds rate with my narrative surprise, excluding the zero lower bound period (2008-2015). Following a 25bp contractionary shock, the federal funds rate rises 25bp on impact (by normalization) then gradually mean-reverts. The 10-year Treasury yield shows modest positive responses peaking around 0.8pp at month 10, though significance is mild.

The price response provides the first critical validation. The PCE price index declines sharply on impact, reaching a trough of approximately 6% below baseline around month 3, then stabilizing near -4% after month 10. This avoids the "price puzzle" that plagues market-based identifications—where contractionary shocks paradoxically raise prices—and confirms the theoretically correct negative sign emerges immediately and persists (Bauer & Swanson, 2023b; Miranda-Agrippino & Ricco, 2021). While confidence bands are wide, the qualitative pattern is clear. This immediate, persistent disinflation is the hallmark of a properly identified monetary shock, establishing the foundation for interpreting real activity effects.

With the price response confirming proper shock identification, real activity displays the expected persistent contractionary effects. Real GDP declines immediately to approximately -4% around month 3, stabilizing around -1 to -2% after month 15. Industrial production exhibits similarly persistent contraction, remaining between -2% and -4% across all horizons. These synchronized declines across output measures are precisely what theory predicts following a monetary tightening that successfully reduces inflation. Unemployment rises sharply, peaking around +0.7pp by month 5, then gradually declining but remaining elevated (+0.3 to +0.5pp) through year three. The delayed unemployment response—lagging output contractions by several months—matches the well-documented labor market adjustment frictions, where firms reduce hours and investment before cutting headcount (Christiano et al., 1999).

The complete transmission pattern—immediate disinflation, synchronized output contrac-

tion, delayed unemployment increases—establishes that direct narrative extraction successfully isolates genuine monetary policy shocks. While confidence bands are wide, characteristic of monetary policy IRF analysis where identification challenges and limited policy variation constrain precision, the qualitative consistency across all variables distinguishes my measure from market-based alternatives. Market-based measures often produce sign anomalies—most notably the "price puzzle" (Ramey, 2016)—that require *ex post* cleaning approaches (Bauer & Swanson, 2023b; Miranda-Agrippino & Ricco, 2021) to restore theoretical coherence. My narrative surprise avoids these anomalies entirely, generating theoretically coherent responses without econometric adjustments. This clean identification—achieved using only public Fed documents released weeks before decisions—provides the foundation for economic validation through trading strategies presented next.

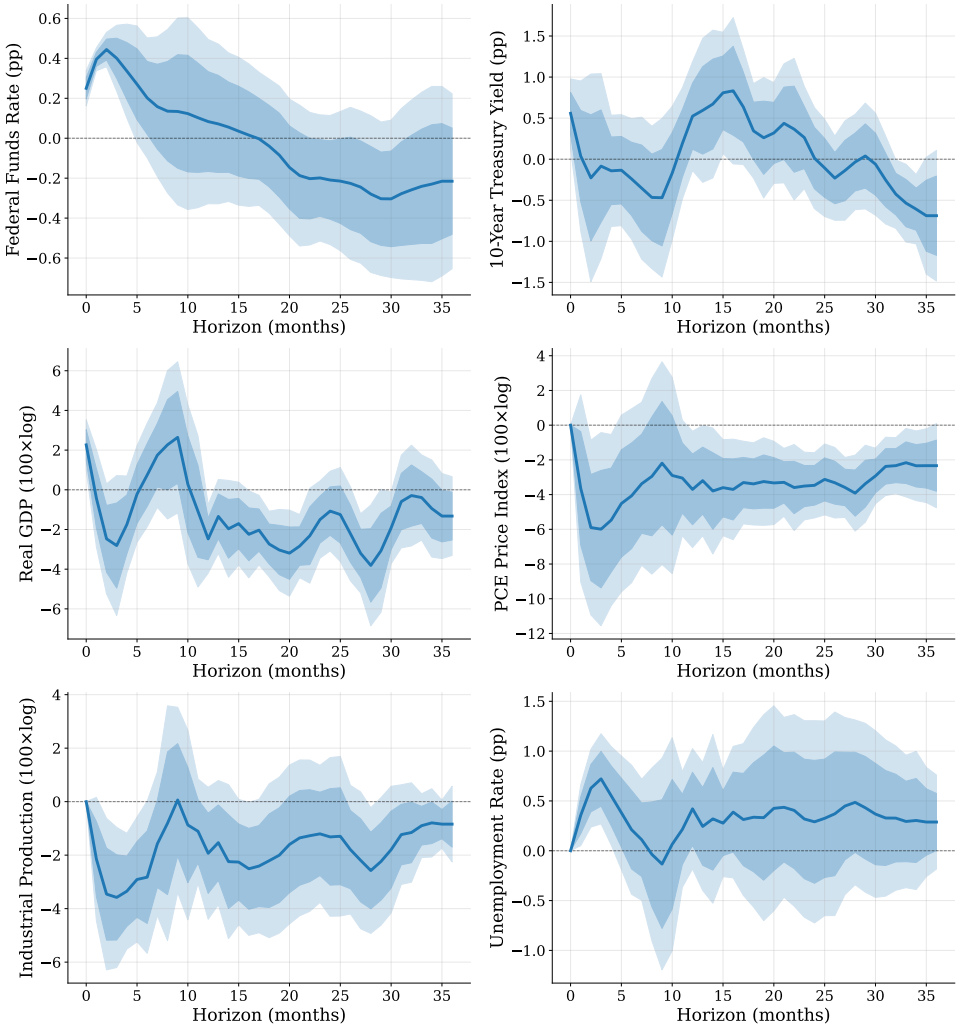
To confirm these transmission patterns are robust across identification strategies and sample periods, Appendix E.1 examines two variations: instrumenting the two-year Treasury yield (following Gertler & Karadi, 2015) instead of the federal funds rate, and including ZLB periods in the sample. If my narrative measure suffered from structural instability or ZLB-specific contamination, first-stage strength would deteriorate in these alternative specifications. Table 16 shows first-stage coefficients remain remarkably stable: 52.51bp (No ZLB) versus 53.39bp (Full), only 1.7% change, with F-statistics of 33.72 and 43.92 respectively. This stability demonstrates that direct extraction maintains predictive power even during ZLB periods when residual-based measures become noisier as their generating regressions face structural breaks. The robustness of first-stage relationships across both alternative instruments and unconventional policy regimes confirms that my narrative measure captures fundamental policy information rather than regime-specific patterns.

Financial Transmission: Yield Curve To understand the yield curve transmission mechanism, I decompose spreads into expected short rates and term premia:

$$r_t^{(n)} - r_t^{(1/12)} = \sum_{i=0}^{n-1} \left(1 - \frac{i}{n}\right) \mathbb{E}_t \Delta r_{t+1}^{(1/12)} + \theta_t^{(n)}, \tag{8}$$

where $r_t^{(n)}$ is the n-year yield, $\mathbb{E}_t \Delta r_{t+1}^{(1/12)}$ is the expected change in the 1-month yield, and $\theta_t^{(n)}$ is the term premium. Using Favero and Fernández-Fuertes (2025)'s data-congruent model, which purges term premia of stochastic trends, Figure 11 presents the complete decomposition for 5-

Figure 10: Macroeconomic Responses to a 25 bp Narrative Surprise



Note: Impulse response functions to a 25 bp monetary policy surprise from an LLM-based narrative measure. Two-stage local projections (2SLP-IV) with federal funds rate instrumented by the narrative surprise. Specification: 4 lags for outcomes and controls, 0 shock lags. Controls: federal funds rate, unemployment, log PCE, log IP, SP500, EBP. 3-period moving average smoothing (h=0 preserved). Newey-West HAC standard errors with bandwidth $h + 1$ (Jordà, 2005). Shaded areas: 68% (dark) and 90% (light) confidence bands. Sample: 269 FOMC meetings (1996–2025), ZLB excluded, COVID included.

and 10-year maturities.

The decomposition reveals yield curve dynamics driven primarily by expected policy paths rather than term premia. Term spreads (row 1) initially compress on impact—both 5Y and 10Y spreads decline approximately 1pp as the front-end rises while longer yields adjust more gradually. The spreads then recover steadily, crossing zero around month 20-25 and reaching approximately +1pp by year three, indicating yield curve steepening as markets anticipate eventual policy normalization and term structure mean reversion.

Expected policy paths (row 2) exhibit similar dynamics: both 5Y and 10Y paths start

negative (approximately -1 to -1.5pp), reflecting expectations that the policy tightening will persist in the near term, then gradually rise through the horizon. The 5Y path shows more volatility initially before stabilizing, while the 10Y path follows a smoother trajectory, both ending slightly positive by month 36 as markets price long-run policy normalization.

Term premia (row 3) exhibit an important delayed response pattern. Starting from zero on impact, both 5Y and 10Y term premia initially decline modestly (reaching approximately -1pp around month 5-10), then rise persistently, with the 10Y premium increasing to near +1pp by year three. Notably, the term premium contribution to spread steepening after month 15 exceeds that of expected paths, particularly for the 10Y maturity where the premium rises approximately 2pp from its trough while the expected path contribution remains more muted. This delayed term premium increase suggests that contractionary monetary policy shocks gradually elevate compensation for interest rate risk as policy and economic uncertainty persist over time. Such finding reveals a two-stage transmission mechanism: initial spread compression driven primarily by rising expected short rates (flattening), followed by delayed spread steepening driven increasingly by rising term premia as uncertainty about the policy path and economic outlook compounds.

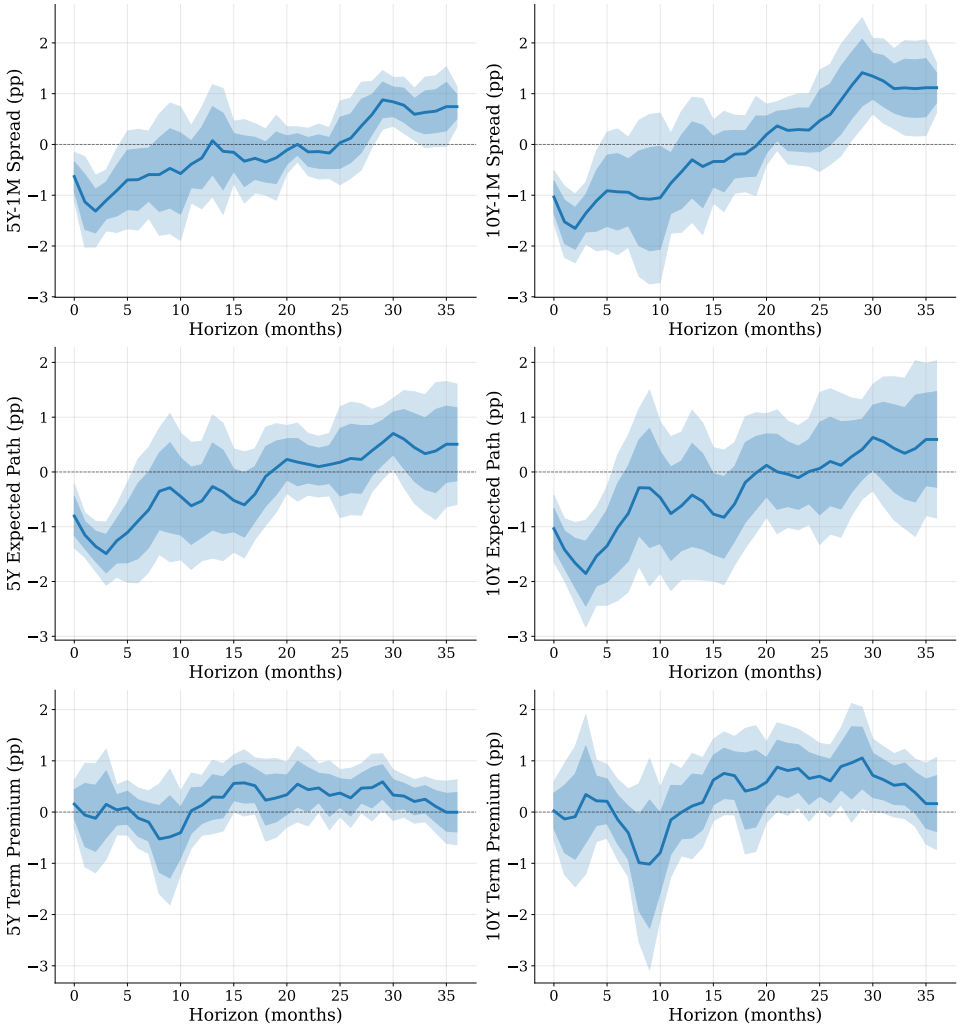
The impulse response estimates show that direct narrative extraction successfully isolates genuine monetary policy shocks. If these shocks are genuine rather than spurious correlations, they should predict persistent asset price movements that markets price gradually, creating exploitable trading opportunities. The clean identification—immediate disinflation, sustained real activity contraction, expectations-driven yield curve transmission—provides the foundation for testing this prediction through out-of-sample trading strategies.

4.6 Economic Validation Through Yield Curve Trading

To validate that the identified shocks contain genuine policy information, I exploit the impulse responses that predict persistent yield curve adjustment over 6-10 months following monetary policy surprises. If these responses reflect genuine shocks rather than spurious correlations, trading strategies exploiting this gradual repricing should generate positive returns. I test this prediction using implementable yield curve trades, hypothesizing that my narrative surprises predict persistent spread movements that markets price gradually, while market-based measures capture only immediate reactions.

The strategy trades on the prediction that yield spreads adjust slowly to monetary surprises

Figure 11: Financial Transmission: Term Spreads, Expected Paths, and Term Premia



Note: Impulse response functions to a 25 bp monetary policy surprise from an LLM-based narrative measure. Two-stage local projections (2SLP-IV) with federal funds rate instrumented by the narrative surprise. Rows show term spreads, expected policy paths, and term premia for 5-year and 10-year maturities. Term spreads: n-year yield minus 1-month Treasury bill rate. Expected paths: yield minus term premium from Favero and Fernández-Fuertes (2025) decomposition. Specification: 4 lags for outcomes and controls, 0 shock lags. Controls: same as Figure 10. 3-period moving average smoothing (h=0 preserved). Newey-West HAC standard errors with bandwidth $h+1$ (Jordà, 2005). Shaded areas: 68% (dark) and 90% (light) confidence bands. Sample: 269 FOMC meetings (1996–2025), ZLB excluded, COVID included.

not yet fully priced. For each FOMC meeting t with surprise s_t, I take positions in the 1-month to 10-year spread:

$$\text{Position}_t = -\text{sign}(s_t) \times \mathbb{I}[|s_t| > q_{67}] \tag{9}$$

where q_{67} is the 67th percentile threshold from a rolling 60-meeting window. Contractionary surprises ($s_t > q_{67}$) trigger short 1-month/long 10-year positions (betting on flattening), while expansionary surprises ($s_t < -q_{67}$) trigger the opposite. Profitability requires two conditions: (1) spread movements beyond immediate market reactions can be predicted from surprises,

and (2) these predictions materialize over horizons where transaction costs remain manageable. The 67th percentile filter concentrates on high-information events while ensuring comparability across measures with different scales.[7]

Table 9 reports results for the equal-weight implementation, which allocates equal capital to both spread components. The 1-month to 10-year spread poses a methodological challenge under equal-weighting: modified duration scales approximately with maturity, implying the 10-year Treasury exhibits roughly 120 times greater price sensitivity to yield changes than the 1-month Treasury.[8] This extreme duration asymmetry causes the long-maturity leg to dominate profit-and-loss dynamics under equal capital allocation. At short holding periods (60-120 days), this structural imbalance obscures differences across surprise measures. However, as horizons extend to economically meaningful windows where term structure adjustments fully materialize, my narrative measure's informational advantage becomes evident. At the 180-day horizon (Panel A), my narrative surprise outperforms the strongest market-based measure (FF1) by 100%. Performance peaks at the 300-day horizon (Panel B), where my narrative surprise exceeds FF1 by 58% with statistically significant returns maintained across all market-based alternatives. This pattern confirms that my narrative measure's superiority is robust to implementation methodology and emerges most clearly at horizons (6-10 months) where yield curve dynamics documented in the impulse response analysis fully manifest.

Figure 12 traces return evolution across holding periods. My narrative surprise exhibits monotonically increasing returns through 6-10 months, consistent with gradual yield curve adjustment in the impulse responses. Returns rise from near zero at 5 days to 12% at 100 days, reaching 26% at 250-300 days under equal-weighting. This confirms that curve dynamics require time to develop, with optimal holding periods at 6-10 months. Extended testing reveals that returns peak around 18 months before declining as mean reversion dominates. Market-based measures show weaker patterns: FF1 plateaus around 15-17%, while MP1 shows negligible performance (included in comprehensive comparison but not main specification).

Figure 13a reveals an important pattern: while my narrative surprise generates consistent positive drift throughout the sample, returns accelerate in 2023-2024. This concentration

[7]The Romer and Romer (2004) measure cannot be implemented in real time due to five-year Greenbook lags. The comparison focuses on implementable measures: narrative surprises from public Fed documents versus market-based measures from high-frequency price movements.

[8]Modified duration for a zero-coupon bond approximates maturity divided by $(1 + \text{yield})$. With typical yields of 4-5%, the 10-year bond has duration ≈ 9.5 years while the 1-month bond has duration ≈ 0.08 years, yielding a ratio of approximately 120:1.

Table 9: Yield Curve Trading Performance: Equal-Weight Benchmark

	Ann. Return		Sharpe Ratio		Hit Rate	Total Return	Trades
Panel A: 180-Day Holding Period							
	Frequency	Calendar	Frequency	Calendar			
My Surprise	0.69%***	0.30%***	0.68	0.45	67.1%	19.22%***	82
ED1	0.21%	0.13%	0.24	0.19	55.9%	5.46%	59
ED4	0.36%**	0.18%**	0.41	0.29	63.4%	9.58%**	71
FF1	0.09%	0.08%	0.11	0.10	50.0%	2.19%	42
FF4	0.16%	0.11%	0.18	0.14	57.1%	4.23%	56
MP1	0.14%	0.12%	0.17	0.16	53.7%	3.47%	41
Panel B: 300-Day Holding Period							
	Ann. Return		Sharpe Ratio		Hit Rate	Total Return	Trades
	Frequency	Calendar	Frequency	Calendar			
My Surprise	0.91%***	0.24%***	0.75	0.38	69.1%	26.14%***	81
ED1	0.33%	0.12%	0.28	0.17	61.0%	8.48%	59
ED4	0.61%***	0.19%***	0.54	0.30	67.1%	16.50%***	70
FF1	0.04%	0.02%	0.04	0.03	57.1%	1.01%	42
FF4	0.27%	0.11%	0.23	0.14	63.6%	7.02%	55
MP1	0.08%	0.04%	0.08	0.06	58.5%	2.02%	41

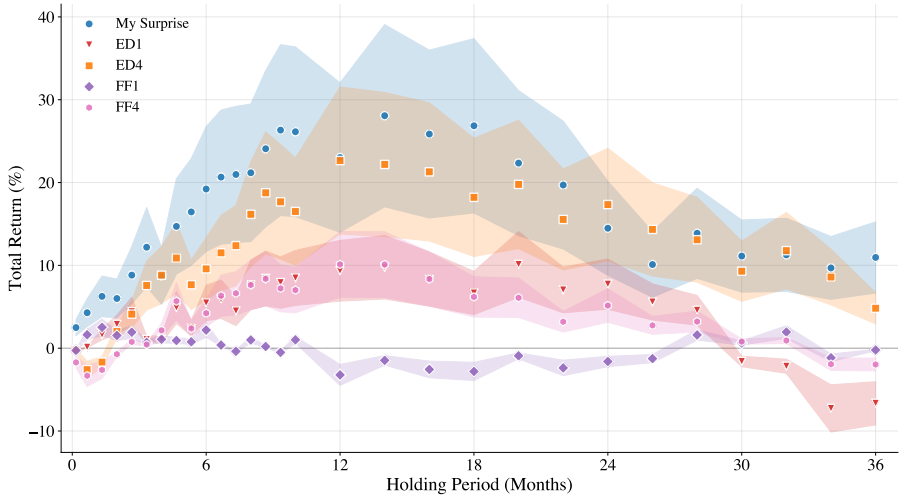
Note: Performance of yield curve strategies trading the 1-month–10-year Treasury spread. After hawkish surprises, the strategy shorts 1-month and goes long 10-year (flattener); after dovish surprises, it goes long 1-month and shorts 10-year (steepener). Positions are opened when |surprise| exceeds the rolling 60-meeting 67th percentile threshold. "Frequency" annualization uses the realized trading frequency (completed trades per trading-year between the first and last trade). "Calendar" annualization assumes one trade every average holding period (i.e., 252/avg holding days). Sharpe ratios are annualized from per-trade Sharpe using $\sqrt{\text{trades/year}}$ under each method. Hit Rate is the share of profitable trades. Total Return is the compounded product of trade returns over the sample. Includes 5 bp transaction costs per trade. ***, **, and * denote significance at the 1%, 5%, and 10% levels based on t-tests against zero mean.

initially seems puzzling: if the LLM (trained through early 2024) better predicts recent Fed behavior, surprises should be smaller. Yet the data shows the opposite: surprises in 2022-2023 are not smaller but among the largest in the sample. This aligns with Romer and Romer (2023), who identify only one true monetary policy shock in 2000-2023: June 2022, precisely when my trading returns spike. Notably, they reach this conclusion using only Minutes (as transcripts remain unavailable), the same document type my system processes. The 2022-2024 period featured genuinely unprecedented policy uncertainty (post-COVID-19 pandemics period) creating large surprises that predict sustained yield curve adjustments. The concentration of returns reflects this exceptional policy volatility rather than training artifacts, though out-of-sample testing on post-2024 meetings would provide additional validation.

The equal-weight benchmark suffers from duration asymmetry that obscures the policy signal. To isolate spread movements, I implement duration-hedging by scaling positions to equalize dollar value changes per basis point (DV01):

$$w_{1m,t} = \frac{D_{10y,t}}{D_{1m,t} + D_{10y,t}}, \quad w_{10y,t} = \frac{D_{1m,t}}{D_{1m,t} + D_{10y,t}} \tag{10}$$

Figure 12: Yield Curve Trading Returns by Holding Period (Equal-Weight)



Note: Cumulative returns as a function of holding period for the equal-weight yield curve trading strategy using the 1-month to 10-year Treasury spread. Each point represents the geometric cumulative return from compounding all individual trades on surprises exceeding the 67th percentile threshold, held for the specified number of days. Zero transaction costs.

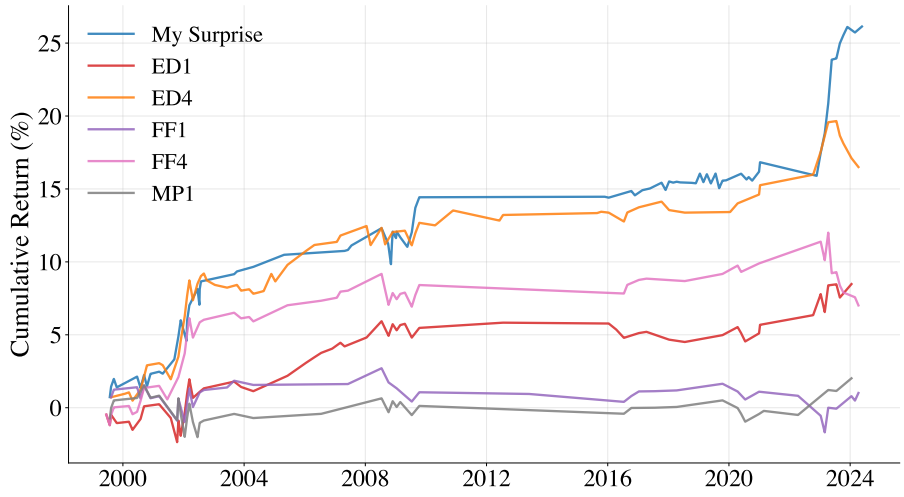
where $D_{1m,t} \approx 0.08$ and $D_{10y,t} \approx 9.5$ are modified durations. This allocates approximately 99% of capital to the 1-month position and 1% to the 10-year position, ensuring both legs contribute equally to P&L. The return for surprise s_t held from t to $t+h$ is:

$$r_{t,t+h} = -\text{sign}(s_t) \times [w_{1m,t} \times \Delta y_{1m,t+h} \times D_{1m,t} + w_{10y,t} \times \Delta y_{10y,t+h} \times D_{10y,t}] \tag{11}$$

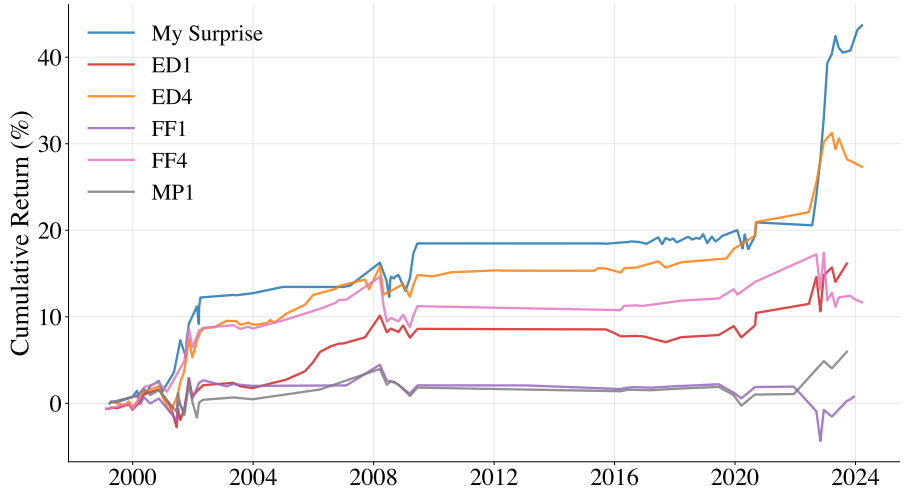
where $\Delta y_{i,t+h}$ denotes yield changes. This neutralizes parallel shifts while capturing spread dynamics. Table 10 reports results.

Duration-hedging substantially amplifies my narrative measure's informational content. At the 180-day horizon (Panel A), my narrative surprise exceeds ED4 (the strongest real-time market-based measure) by 60%, with statistically significant returns establishing my narrative measure's dominance at the six-month window where Fed communication effects fully materialize in term structure dynamics. Returns continue growing at the 300-day horizon (Panel B), reaching comparable magnitudes to ED4 as both measures capture longer-term rate paths with similar effectiveness (1.62% versus 1.55% frequency-annualized returns). The importance of proper risk management is quantitatively significant: comparing Tables 9 and 10 at the 180-day horizon reveals duration-hedging more than doubles returns—a 127% improvement—by optimally weighting the 1-month yield's superior policy signal transmission (99% capital allocation to the front end versus 1% to the long end, balancing the 120:1 duration ratio).

Figure 13: Cumulative Returns to Yield Curve Strategies (1996–2025)



(a) Equal-weight portfolio: 300-day holding period.



(b) Duration-hedged (DV01-neutralized): 180-day holding period.

Note: **(a)** Equal-weight strategy using the 1-month/10-year Treasury spread, 300-day holding period. **(b)** Duration-hedged (DV01-equalized) strategy, 180-day holding period. Each line represents the compounded growth of $1 invested with position sizing and holding periods as indicated. Zero transaction costs. The narrative surprise exhibits sustained positive drift and clear separation from market-based measures in both implementations, with more consistent returns under duration hedging.

Market-based measures exhibit substantially weaker performance under duration-hedging. ED4 approaches but does not match my narrative measure's performance, while other measures (MP1, FF4, ED1, FF1) show weak or negative returns across all horizons. The consistent pattern across multiple holding periods and both weighting methodologies establishes that my narrative surprise dominates all real-time implementable alternatives in capturing persistent yield curve movements, with the advantage most pronounced in the 6-10 month window where Fed communication effects transmit through the front end of the term structure. Figure 13 illus-

Table 10: Yield Curve Trading Performance: Duration-Hedged Strategy

	Ann. Return		Sharpe Ratio		Hit Rate	Total Return	Trades
	Frequency	Calendar	Frequency	Calendar			
Panel A: 180-Day Holding Period							
My Surprise	1.43%***	0.63%***	0.66	0.44	65.9%	43.68%***	82
ED1	0.61%	0.37%	0.32	0.25	59.3%	16.15%	59
ED4	0.95%***	0.49%***	0.53	0.38	67.6%	27.33%***	71
FF1	0.05%	0.04%	0.03	0.02	57.1%	0.76%	42
FF4	0.45%	0.29%	0.22	0.17	64.3%	11.67%	56
MP1	0.24%	0.21%	0.15	0.14	53.7%	5.97%	41
Panel B: 300-Day Holding Period							
	Ann. Return		Sharpe Ratio		Hit Rate	Total Return	Trades
	Frequency	Calendar	Frequency	Calendar			
My Surprise	1.62%***	0.43%***	0.57	0.29	61.7%	49.87%***	81
ED1	0.81%	0.29%	0.30	0.18	54.2%	21.74%	59
ED4	1.55%***	0.48%***	0.63	0.35	65.7%	47.30%***	70
FF1	-0.32%	-0.16%	-0.14	-0.10	42.9%	-8.58%	42
FF4	0.60%	0.23%	0.23	0.14	56.4%	15.54%	55
MP1	-0.08%	-0.04%	-0.04	-0.03	46.3%	-2.74%	41

Note: Performance of yield curve strategies trading the 1-month–10-year Treasury spread. After hawkish surprises, the strategy shorts 1-month and goes long 10-year (flattener); after dovish surprises, it goes long 1-month and shorts 10-year (steepener). Positions are opened when |surprise| exceeds the rolling 60-meeting 67th percentile threshold. "Frequency" annualization uses the realized trading frequency (completed trades per trading-year between the first and last trade). "Calendar" annualization assumes one trade every average holding period (i.e., 252/avg holding days). Sharpe ratios are annualized from per-trade Sharpe using $\sqrt{\text{trades/year}}$ under each method. Hit Rate is the share of profitable trades. Total Return is the compounded product of trade returns over the sample. Includes 5 bp transaction costs per trade. ***, **, and * denote significance at the 1%, 5%, and 10% levels based on t-tests against zero mean.

trates cumulative return evolution over calendar time, demonstrating sustained outperformance of my narrative measure relative to market-based alternatives.

The trading results validate the impulse response evidence. My narrative surprise outperforms ED4 (the strongest market-based measure) by 60% at 180 days with duration-hedging, and by 101% under equal-weighting. Returns increase monotonically through 6-10 months—precisely when impulse responses show maximum curve adjustment—then decline as mean reversion dominates. This timing confirms my surprises predict gradual repricing that markets miss initially. Duration-hedging doubles returns by concentrating capital (99%) in the policy-sensitive front end. The consistent profitability across my sample, combined with better measurement properties and theoretically coherent impulse responses, establishes that narrative surprises extracted from Fed communications provide superior identification of monetary policy shocks.

5 Conclusion

Measuring monetary policy surprises requires first measuring expectations. This paper demonstrates that conditional expectations can be extracted directly from central bank communica-

tions through multi-agent Large Language Model synthesis, achieving lower measurement error without ex-post econometric cleaning. The *New Hope* is methodological: as extraction methods become more sophisticated, incorporating richer information sources or more advanced architectures, identification quality improves at the source rather than through post-hoc adjustments. This paper establishes the feasibility of this approach by demonstrating that even the simplest implementation—processing only Beige Books and Minutes—achieves measurement advantages over existing methods. To further establish advantages of my procedure, I run three validation exercises. First, I run tests that show that my measure has minimal measurement error. Second, I show that it produces theoretically consistent impulse responses showing persistent contractionary transmission. Third, I show that it generates economically significant trading profits.

Such framework naturally extends beyond this proof-of-concept implementation and could generate the following possibilities: (1) central banks incorporating internal forecasts alongside public communications; (2) market participants integrating real-time news flows arriving during blackout periods; and (3) researchers applying the architecture to any domain where institutional communications precede market-moving events—earnings announcements, regulatory decisions, geopolitical developments.

My results carry implications beyond monetary economics. For policymakers, the findings validate the Fed's communication strategy: public documents explain half of all decisions, demonstrating successful information transmission to attentive observers. For researchers, the framework offers a scalable alternative to both hand-coded narrative measures and high-frequency identification. Future work could exploit the multi-agent architecture for experimental designs.

What remains unpredictable, namely, the genuine policy surprises, can now be measured in real-time from the public record alone, offering new hope for both researchers studying monetary transmission and practitioners managing interest rate risk.

References

Acosta, M. (2023). A New Measure of Central Bank Transparency and Implications for the Effectiveness of Monetary Policy. *International Journal of Central Banking*, *19*(3), 49–97.

Acosta, M., & Meade, E. E. (2015). *Hanging on Every Word: Semantic Analysis of the FOMC's Post-Meeting Statement* (tech. rep.). Board of Governors of the Federal Reserve System (US).

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.

Armesto, M. T., Hernández-Murillo, R., Owyang, M. T., & Piger, J. (2009). Measuring the Information Content of the Beige Book: A Mixed Data Sampling Approach. *Journal of Money, Credit and Banking*, *41*(1), 35–55.

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). A reassessment of monetary policy surprises and high-frequency identification. *NBER Macroeconomics Annual*, *37*(1), 87–155.

Bernanke, B. S. (1990). The Federal Funds Rate and the Channels of Monetary Transnission.

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

Bernanke, B. S., & Kuttner, K. N. (2005). What explains the stock market's reaction to federal reserve policy? *The Journal of finance*, *60*(3), 1221–1257.

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

Bybee, L. (2023b). 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., Eichenbaum, M. S., & Evans, C. (1994). The Effects of Monetary Policy Shocks: Some Evidence from the Flow of Funds.

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. (2004). Comments on "A New Measure of Monetary Shocks: Derivation and Implications" by Christina Romer and David Romer. *NBER Economic Fluctuations and Growth Meeting*.

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

Cochrane, J. H. (2025, May). *Inflation dynamics with a generalized lucas phillips curve* (Working Paper) (Posted: May 28, 2025; Date Written: May 28, 2025; Available at SSRN 5272734). Hoover Institution; National Bureau of Economic Research.

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.

Doh, T., Song, D., Yang, S.-K., et al. (2020). *Deciphering Federal Reserve Communication via Text Analysis of Alternative FOMC Statements*. Federal Research Bank of Kansas City Kansas City, MO, USA.

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

Favero, C. A., & Fernández-Fuertes, R. (2025). Towards Data-Congruent Models of the Term Structure of Interest Rates. *Econometric Reviews*, 1–23. https://doi.org/10.1080/07474938.2025.2458223

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. (2024). Regional Economic Sentiment: Constructing Quantitative Estimates from the Beige Book and Testing their ability to Forecast Recessions. *Economic Commentary*, (2024-08).

Fleming, M. J., Mizrach, B., & Nguyen, G. (2018). The Microstructure of a US Treasury ECN: The BrokerTec platform. *Journal of Financial Markets*, *40*, 2–22.

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.

Gao, M., Li, Y., Liu, B., Yu, Y., Wang, P., Lin, C.-Y., & Lai, F. (2025). Single-agent or multi-agent systems? why not both? *arXiv preprint arXiv:2505.18286*. https://arxiv.org/abs/2505.18286

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

Glasserman, P., & Lin, C. (2024). Assessing Look-Ahead Bias in Stock Return Predictions Generated by GPT Sentiment Analysis [Originally published September 2023, arXiv:2309.17322]. *The Journal of Financial Data Science*, *6*(1), 25–42.

Gu, J., Pang, L., Shen, H., & Cheng, X. (2024). Do llms play dice? exploring probability distribution sampling in large language models for behavioral simulation. *arXiv preprint arXiv:2404.09043*. https://arxiv.org/abs/2404.09043

Guo, T., Song, L., Ding, H., et al. (2024). Large language model based multi-agents. *International Joint Conference on Artificial Intelligence (IJCAI)*. https://dl.acm.org/doi/10.24963/ijcai.2024/890

Gürkaynak, R. S., Sack, B., & Swanson, E. (2005a). 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

Gürkaynak, R. S., Sack, B., & Swanson, E. T. (2005b). Do Actions Speak Louder than Words? the Response of Asset Prices to Monetary Policy Actions and Statements. *International Journal of Central Banking*, *1*(1), 55–93.

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

Han, S., Zhang, Q., Yao, Y., Jin, W., & Xu, Z. (2024). Llm multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*. https://arxiv.org/abs/2402.03578

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.

Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and Deliberation within the FOMC: A Computational Linguistics Approach. *The Quarterly Journal of Economics*, *133*(2), 801–870.

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

Huang, Y.-L., & Kuan, C.-M. (2021). Economic Prediction with the FOMC Minutes: A Application of Text Mining. *International Review of Economics & Finance*, *71*, 751–761.

Jajoo, G., Chitale, P. A., & Agarwal, S. (2025). Masca: Llm based-multi agents system for credit assessment. *arXiv preprint arXiv:2507.22758*. https://arxiv.org/abs/2507.22758

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, September). *Disentangling Monetary Policy, Central Bank Information, and Fed Response to News Shocks* (Working Paper) (This version: September 11, 2025; First version: February 3, 2025). European Central Bank.

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*.

Koa, K. J., Du, T., Ma, Y., Wang, X., Ng, R., Huanhuan, Z., & Chua, T.-S. (2024). Massively multi-agents reveal that large language models simulate market dynamics. *OpenReview*. https://openreview.net/forum?id=obYDlJN0oU

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.

Leeper, E. M. (1997). Narrative and VAR Approaches to Monetary Policy: Common Identification Problems. *Journal of Monetary Economics*, *40*(3), 641–657.

Li, J., Zhang, Q., Yu, Y., Fu, Q., & Ye, D. (2024). More Agents Is All You Need. *arXiv preprint arXiv:2402.05120*.

Lopez-Lira, A. (2025, April). *Can Large Language Models trade? testing financial theories with LLM Agents in Market Simulations* (Working Paper) (First version: November 2024). University of Florida.

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

Lucca, D. O., & Moench, E. (2015). The Pre-FOMC Announcement Drift. *The Journal of Finance*, *70*(1), 329–371.

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.

Newey, W. K., & West, K. D. (1987). A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix. *Econometrica*, *55*(3), 703–708. https://doi.org/10.2307/1913610

Noguer i Alonso, M. (2024, November). *Look-ahead Bias in Large Language Models (LLMs): Implications and Applications in Finance* (Working Paper). Artificial Intelligence Finance Institute.

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

Olea, J. L. M., Stock, J. H., & Watson, M. W. (2021). Inference in Structural VARs Identified with an External Instrument. *Journal of Econometrics*, *225*(1), 74–87.

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).

Ramey, V. A. (2016). Macroeconomic Shocks and their Propagation. *Handbook of macroeconomics*, *2*, 71–162.

Ricco, G., & Savini, E. (2025, April). *Decomposing Monetary Policy Surprises: Shock, Information, and Policy Rule Revision* (Working Paper) (April 25, 2025). University of Warwick.

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. (1997). Identification and the Narrative Approach: A Reply to Leeper. *Journal of Monetary Economics*, *40*(3), 659–665.

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

Sarkar, S. K., & Vafa, K. (2024, October). *Lookahead Bias in Pretrained Language Models* (Working Paper) (Available at arXiv:2410.xxxxx). Harvard University.

Schoenegger, P., Park, P. S., Karger, E., Trott, S., & Tetlock, P. E. (2025). AI-Augmented Predictions: LLM Assistants Improve Human Forecasting Accuracy. *ACM Transactions on Interactive Intelligent Systems*, *15*(1), 1–25. https://doi.org/10.1145/3707649

Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2022). Measuring News Sentiment. *Journal of econometrics*, *228*(2), 221–243.

Shi, J., & Hollifield, B. (2024). *Predictive Power of LLMs in Financial Markets* (Working Paper). Carnegie Mellon University.

Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica: Journal of the Econometric Society*, 1–48.

Sims, C. A. (1992). Interpreting the Macroeconomic Time Series Facts: The Effects of Monetary Policy. *European Economic Review*, *36*(5), 975–1000.

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

Stock, J., & Yogo, M. (2005). Asymptotic distributions of instrumental variables statistics with many instruments. *Identification and inference for econometric models: Essays in honor of Thomas Rothenberg*, *6*, 109–120.

Stock, J. H., & Watson, M. W. (2001). Vector autoregressions. *Journal of Economic perspectives*, *15*(4), 101–115.

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

Stock, J. H., & Watson, M. W. (2018). Identification and estimation of dynamic causal effects in macroeconomics using external instruments. *The Economic Journal*, *128*(610), 917–948.

Strongin, S. (1995). The Identification of Monetary Policy Disturbances Explaining the Liquidity Puzzle. *Journal of Monetary Economics*, *35*(3), 463–497.

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.

Team, A. S. (2025). Margen: Multi-agent llm approach for self-directed market research and analysis. *Amazon Science*. https://www.amazon.science/publications/margen-multi-agent-llm-approach-for-self-directed-market-research-and-analysis

Tillmann, A. (2025). Literature Review of Multi-Agent Debate for Problem-Solving. *arXiv preprint arXiv:2506.00066*. https://arxiv.org/abs/2506.00066

Tran, K.-T., Dao, D., Nguyen, M.-D., Pham, Q.-V., O'Sullivan, B., & Nguyen, H. D. (2025). Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*. https://arxiv.org/abs/2501.06322

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. (2022a). 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. (2022b). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.

White, N. (2025, April). *The New Keynesian Price Puzzle: Reinterpreting Inflation Dynamics* (Working Paper) (Posted: April 18, 2025; Date Written: February 17, 2025; Available at SSRN 5143557). Amherst College.

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*.

Xiao, Y., Sun, E., Luo, D., & Wang, W. (2024). Tradingagents: Multi-agents llm financial trading framework. *arXiv preprint arXiv:2412.20138*. https://tradingagents-ai.github.io

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

Yu, Y., Yao, Z., Li, H., Deng, Z., Jiang, Y., Cao, Y., Chen, Z., Suchow, J. W., Cui, Z., Liu, R., Xu, Z., Zhang, D., Subbalakshmi, K., Xiong, G., He, Y., Huang, J., Li, D., & Xie, Q. (2024). Fincon: A synthesized llm multi-agent system with conceptual verbal reinforcement for enhanced financial decision making. *arXiv preprint arXiv:2407.06567*. https://arxiv.org/abs/2407.06567

Zhu, K., Du, H., Hong, Z., Yang, X., Guo, S., Wang, Z., Wang, Z., Qian, C., Tang, R., Ji, H., & You, J. (2025). Multiagentbench: Evaluating the collaboration and competition of llm agents. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, 8580–8622. https://aclanthology.org/2025.acl-long.421/

A Agent Output Examples

This appendix presents the complete JSON output specifications for each agent in the multi-agent system. The examples demonstrate the dual-output architecture (quantitative measures and textual reasoning) across different monetary policy regimes.

A.1 Agent IM: Policy Intelligence Extraction

A.1.1 Example 1: Financial Crisis Period (December 2008)

The following output shows Agent IM's analysis of the October 29, 2008 FOMC Minutes, which informed expectations for the December 16, 2008 meeting during the peak of the financial crisis. The updated policy distribution assigns 68.75% probability to a 50 basis-point cut, 20% to a 25 basis-point cut, and 11.25% to holding. The shock discovery identifies "expansion and extension of swap lines with multiple foreign central banks" and "severe impairment of commercial paper market." The internal debate text shows unanimous voting with members expressing divergent concerns about excessive accommodation versus deflation risks:

Policy Extractor Output: October 29 2008 FOMC Minutes (for December 16 Meeting)


```
{
  "decision_context": {
    "actual_decision": "50 basis points cut to 1.0%",
    "internal_debate": {
      "hawks_preference": "Some members favored smaller cuts or pauses to avoid signaling
      ↪  excessive accommodation or panic amid inflation concerns",
      "doves_preference": "Others advocated for large cuts to address severe financial market
      ↪  strains, economic slowdown, and deflation risks",
      "compromise_reasoning": "The unanimous 50bp cut balanced the need for urgent easing with
      ↪  concerns about inflation, financial stability, and limited policy space, reflecting
      ↪  consensus to act decisively but cautiously."
    },
    "voting_pattern": "Unanimous approval of the 50bp cut",
    "dissenting_views": []
  },
  "forward_guidance_signals": {
    "explicit": "The Committee indicated it would continue to monitor economic and financial
    ↪  developments carefully and act as needed to promote sustainable growth and price
    ↪  stability.",
    "implicit": "Strong emphasis on data dependence and readiness to provide further
    ↪  accommodation if financial conditions and economic data worsen."
  },
  "policy_stance_distribution": {
    "next_meeting_probabilities": {
      "(0.5,-1)": 0.6875,
      "(0.25,-1)": 0.2,
      "(0,0)": 0.1125
```

```
    }
  },
  "shock_discovery": {
    "new_information": [
      "Expansion and extension of swap lines with multiple foreign central banks totaling
      ↪  hundreds of billions of dollars to address international liquidity strains",
      "Discussion of potential limits to policy effectiveness due to financial dislocations",
      "Severe impairment of commercial paper market and money market funds"
    ]
  }
}
```


A.1.2 Example 2: Tightening Cycle Initiation (March 2022)

The following output shows Agent IM's analysis of the January 26, 2022 FOMC Minutes, which informed expectations for the March 16, 2022 meeting marking the beginning of the current tightening cycle. The probability distribution assigns 70% probability to a 12.5 basis-point hike, 25% to holding, and 5% to a 25 basis-point hike. The shock discovery identifies "balance sheet reduction will begin only after rate hikes commence, signaling clear sequencing." The internal debate text shows participants "advocating a faster pace than prior cycles" versus those preferring "maintaining accommodation to assess Omicron impact":


```
Policy Extractor Output: January 26 2022 FOMC Minutes (for March 16 Meeting)
{
  "decision_context": {
    "actual_decision": "Hold target range for federal funds rate at 0 to 1/4 percent",
    "internal_debate": {
      "hawks_preference": "Some participants favored preparing for imminent rate hikes due to
      ↪  elevated inflation and strong labor market, with some advocating a faster pace than
      ↪  prior cycles",
      "doves_preference": "Others preferred maintaining accommodation to assess Omicron impact,
      ↪  supply chain issues, and labor market recovery",
      "compromise_reasoning": "Consensus to hold balances inflation concerns with economic
      ↪  uncertainties, sequencing of balance sheet runoff, and pandemic-related risks"
    },
    "voting_pattern": "Unanimous hold decision",
    "dissenting_views": []
  },
  "forward_guidance_signals": {
    "explicit": "Participants agreed that ongoing rate increases would soon be warranted and
    ↪  that balance sheet runoff would begin after rate hikes commence",
    "implicit": "Strong emphasis on data dependence, gradualism, and sequencing of policy tools
    ↪  with flexibility to adjust as conditions evolve"
  },
  "policy_stance_distribution": {
    "next_meeting_probabilities": {
      "(0.125,1)": 0.7,
      "(0,0)": 0.25,
```

```
      "(0.25,1)": 0.05
    }
  },
  "shock_discovery": {
    "new_information": [
      "Balance sheet reduction will begin only after rate hikes commence, signaling clear
      ↪  sequencing",
      "Committee prioritizes gradualism and data dependence over aggressive front-loading",
      "Extensive discussion of operational readiness for market disruptions and flexibility in
      ↪  policy tools"
    ]
  }
}
```


A.2 Agent IB: Beige Book Economic Conditions

Agent IB operates on the Beige Book to extract quantitative assessments of economic conditions from qualitative narratives. The agent produces both economic condition scores and textual extractions that preserve the context underlying each score.

The following output shows Agent IB's analysis of the March 2, 2022 Beige Book during the tightening cycle liftoff period. Inflation scored +0.70 with 29.1% weight, with text including "robust price increases" and "rising input costs across a broad swath of industries." Employment scored +0.59 with 28.8% weight, with text including "widespread strong demand for workers." Economic growth scored +0.10 and consumer spending scored -0.24, with text including "weaker spending than in the prior report" and Omicron-related disruptions. The stance distribution shows 36 hawkish, 9 dovish, and 22 neutral sentences from 67 total:

Beige Book Analyzer Output: March 2 2022


```
{
  "scores": {
    "inflation": 0.70,
    "employment": 0.59,
    "economic growth": 0.10,
    "consumer spending": -0.24
  },
  "weights": {
    "inflation": 0.291,
    "employment": 0.288,
    "economic growth": 0.244,
    "consumer spending": 0.177
  },
  "summary": {
    "overall_policy_bias": "neutral",
    "signal_strength": "moderate",
    "stance_distribution": {
      "hawkish": 36,
```

```
      "dovish": 9,
      "neutral": 22
    }
  }
}
```


A.3 Agent II: Expectation Synthesis

Agent II serves as the synthesis agent, integrating outputs from both Agent IM (Minutes-based policy intelligence) and Agent IB (Beige Book economic conditions) to form a coherent prior probability distribution for the upcoming FOMC decision.

A.3.1 Example 1: Financial Crisis Period (December 2008)

The following output shows Agent II's synthesis for the December 16, 2008 meeting. Agent II's expected rate change is -0.1875 percentage points. The probability distribution assigns 75% to a 25 basis-point cut and 25% to holding. The textual outputs include specific influence assessments from both Agent IM and Agent IB:

Expectation Engine Output: December 16 2008 FOMC Meeting


```
{
  "previous_rate": 1.0,
  "expected_rate_change": -0.1875,
  "expectation_justification": "Both the Beige Book and recent Policy Intelligence strongly
  ↪  indicate worsening economic conditions with easing inflation pressures. The Beige Book
  ↪  shows broad weakening in growth, employment, and consumer spending, consistent with the
  ↪  Committee's prior unanimous 50bp cut to 1.0%. Policy Intelligence emphasizes continued
  ↪  data dependence and readiness to ease further if conditions deteriorate, but also notes
  ↪  limited policy space and concerns about financial stability. Given the zero lower bound
  ↪  constraint and the strong dovish signals, a modest 25bp cut is the most probable next
  ↪  step, balancing urgency to support growth with caution about limited room for easing.",
  "beige_book_influence": "The Beige Book's strong dovish tone on inflation easing, weakening
  ↪  employment, and declining consumer spending reinforces the expectation of further easing.
  ↪  Its consistent, high-intensity negative scores across key indicators confirm the
  ↪  Committee's concerns about economic contraction and justify additional accommodation.",
  "policy_intelligence_influence": "Policy Intelligence from the last meeting shows the
  ↪  Committee's unanimous 50bp cut to 1.0%, with explicit readiness to act further if needed
  ↪  but also caution due to limited policy space and financial stability risks. This tempers
  ↪  the magnitude of expected easing, favoring a smaller cut or hold rather than large cuts,
  ↪  reflecting a balance between urgency and prudence.",
  "signal_strength": "strong",
  "confidence": 0.85,
  "distribution": {
    "(0.25,-1)": 0.75,
    "(0,0)": 0.25
  }
}
```

A.3.2 Example 2: Tightening Cycle Initiation (March 2022)

The following output shows Agent II's synthesis for the March 16, 2022 meeting. The distribution assigns 70% to a 12.5 basis-point hike, 25% to holding, and 5% to a 25 basis-point hike, with an expected rate change of +0.1 percentage points. The confidence score is 0.75 and signal strength is "moderate." The textual influence assessments include: Beige Book influence notes "modest growth and weaker consumer spending moderate the hawkishness," while policy intelligence influence notes "data dependence, sequencing of balance sheet runoff, and gradual normalization":

Expectation Engine Output: March 16 2022 FOMC Meeting


```
{
  "previous_rate": 0.125,
  "expected_rate_change": 0.1,
  "expectation_justification": "The Beige Book shows robust inflation and strong labor demand,
  ↪  signaling hawkish pressure, while economic growth is modest and consumer spending weaker,
  ↪  tempering urgency. Policy Intelligence from the last three meetings consistently
  ↪  emphasized elevated, persistent inflation and strong labor markets but maintained a
  ↪  cautious, data-dependent stance with no immediate hikes. The Fed's readiness to act soon
  ↪  but not immediately, combined with ongoing tapering and balance sheet discussions,
  ↪  supports a high probability of a 0.125% hike at the next meeting. This balances the
  ↪  hawkish inflation/employment signals with the committee's emphasis on sequencing and
  ↪  gradualism.",
  "beige_book_influence": "The Beige Book's strong inflation and employment signals push toward
  ↪  tightening, but modest growth and weaker consumer spending moderate the hawkishness,
  ↪  supporting a near-term rate increase but not aggressive hikes.",
  "policy_intelligence_influence": "Policy Intelligence shows a clear hawkish tilt on inflation
  ↪  and labor markets but stresses data dependence, sequencing of balance sheet runoff, and
  ↪  gradual normalization, favoring a cautious first hike soon rather than immediate or large
  ↪  moves.",
  "signal_strength": "moderate",
  "confidence": 0.75,
  "distribution": {
    "(0,0)": 0.25,
    "(0.125,1)": 0.7,
    "(0.25,1)": 0.05
  }
}
```


A.4 Agent III: Surprise Quantification

Agent III quantifies the monetary policy surprise on FOMC announcement day by comparing the realized policy decision with Agent II's prior probability distribution. The agent produces both mechanical surprise measures and contextual salience assessments.

A.4.1 Example 1: Financial Crisis Period (December 2008)

The following output shows Agent III's surprise calculation for the December 16, 2008 meeting. The surprise is $s = -0.6875$ percentage points (the Fed cut 87.5 basis points when Agent II expected 18.75 basis points). The contextual score is $\sigma_s = 0.85$. Agent II's prior assigned 75% probability to a 25 basis-point cut and 25% to holding:

Surprise Snipper Output: December 16 2008 FOMC Meeting


```
{
  "meeting_date": "2008-12-16",
  "expected_rate_change": -0.1875,
  "realized_rate_change": -0.875,
  "surprise_rate": -0.6875,
  "surprise_score": 0.85,
  "surprise_direction": "dovish",
  "confidence": 0.9,
  "justification": "The realized rate change of -0.875% versus an expected cut of -0.1875%
  ↪  yields a surprise of -0.6875%, a large dovish surprise. The Committee's decision to cut
  ↪  the federal funds rate to a range of 0 to 1/4 percent (midpoint 0.125%) is a much larger
  ↪  easing than the modest 25bp cut expected. The statement emphasizes worsening economic
  ↪  conditions and diminished inflation, consistent with aggressive easing. Given the prior
  ↪  expectation distribution heavily favored a modest 25bp cut and only 25% probability of no
  ↪  change, this substantial cut is a strong dovish surprise likely to cause a significant
  ↪  market reaction."
}
```


A.4.2 Example 2: Tightening Cycle Initiation (March 2022)

The following output shows Agent III's surprise calculation for the March 16, 2022 meeting. The surprise is $s = +0.15$ percentage points (the Fed hiked 25 basis points when Agent II expected 10 basis points). The contextual score is $\sigma_s = 0.6$. Agent II's prior assigned 70% probability to a 12.5 basis-point hike and 5% probability to a 25 basis-point hike:

Surprise Snipper Output: March 16 2022 FOMC Meeting


```
{
  "meeting_date": "2022-03-16",
  "expected_rate_change": 0.1,
  "realized_rate_change": 0.25,
  "surprise_rate": 0.15,
  "surprise_score": 0.6,
  "surprise_direction": "hawkish",
  "confidence": 0.85,
  "justification": "The Fed raised the target range from 0-0.25% to 0.25-0.5%, a 25bp hike
  ↪  (0.25) versus the expected 10bp increase (0.1), resulting in a 15bp hawkish surprise. The
  ↪  prior expectation was strongly tilted toward a 12bp hike (70% probability), so a full
  ↪  25bp hike is somewhat more aggressive than anticipated. The surprise score of 0.6
  ↪  reflects a moderately strong market reaction given the Fed's firm language on ongoing
  ↪  increases and the hawkish tone amid inflation and labor market strength."
```

}

B Agent Validation and Supplementary Analysis

B.1 Agent IB Statistical Validation

This appendix presents statistical analysis of Agent IB's quantitative outputs. The analysis examines distributional properties, cross-variable correlations, weight dynamics, and relationships with conventional macroeconomic indicators.

B.1.1 Distributional Properties

Figure 14 displays the empirical distributions of the four variable scores across all Beige Book releases in the sample. All four variables show unimodal distributions centered near zero. The distributions exhibit mild skewness and some deviations from normality, particularly in the tails.

The inflation score distribution shows slight positive skewness. Employment scores display the most symmetric distribution among the four variables. Economic growth and consumer spending scores show similar distributional shapes with moderate variance.

B.1.2 Cross-Variable Correlations

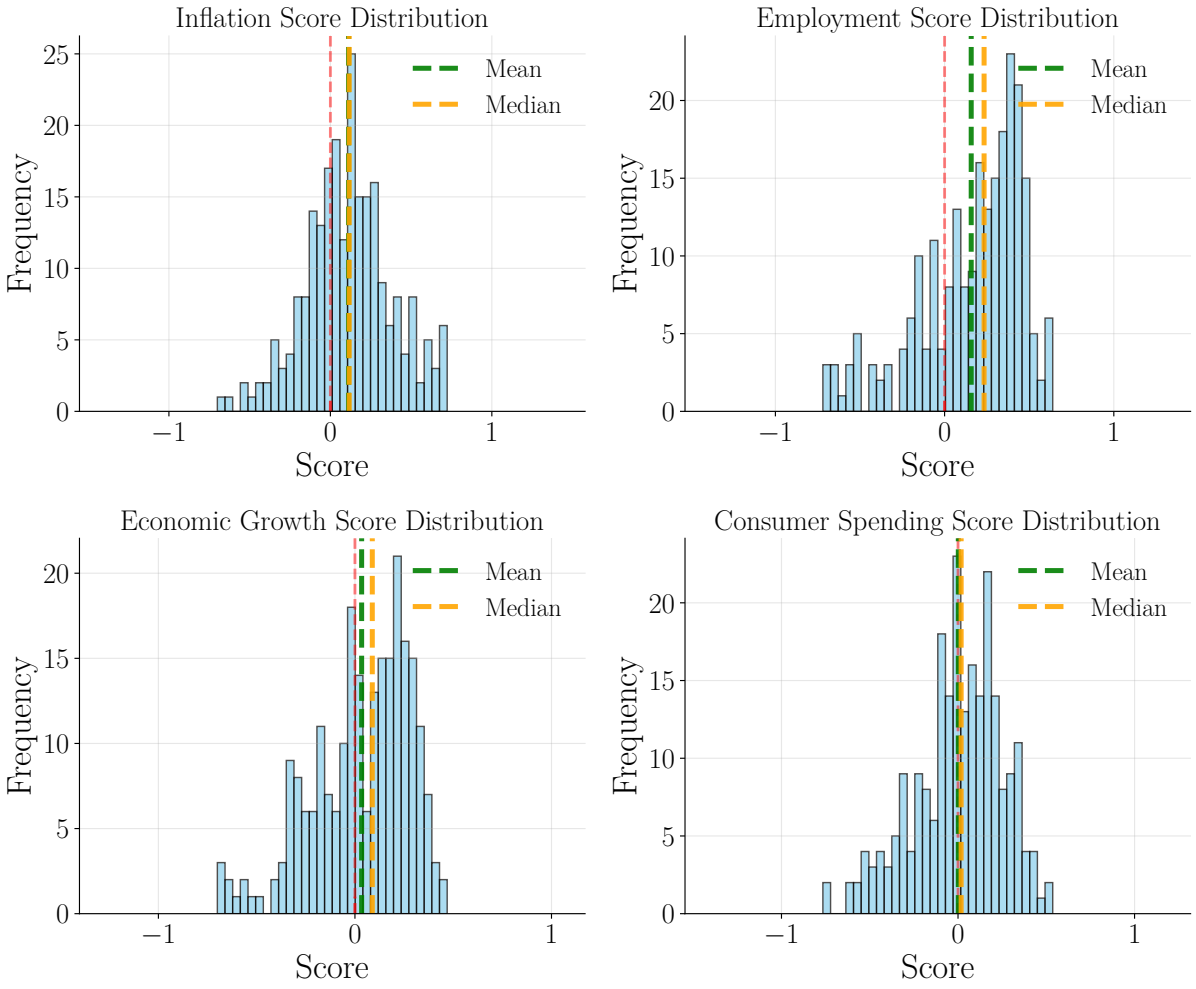
Figure 15 presents the correlation matrix for the four variable scores and the weighted aggregate index. All pairwise correlations are positive, ranging from moderate (0.3-0.5) to strong (0.6-0.8). The strongest correlation appears between economic growth and employment scores ($\rho \approx 0.7$). Inflation scores show moderate positive correlation with employment ($\rho \approx 0.5$) and weaker correlation with growth ($\rho \approx 0.4$).

The aggregate index correlates strongly with all four component variables, with correlation coefficients ranging from 0.65 to 0.85. When inflation dominates Beige Book discussion, it receives higher weight in the aggregate, but the aggregate remains responsive to all four variables rather than collapsing to a single-factor measure.

B.1.3 Weight Dynamics

Figure 16 shows the time-varying weights assigned to each macroeconomic variable in the Beige Book. The stacked area chart shows that these weights are dynamic rather than converging to

Figure 14: Empirical distributions of Agent IB's variable scores



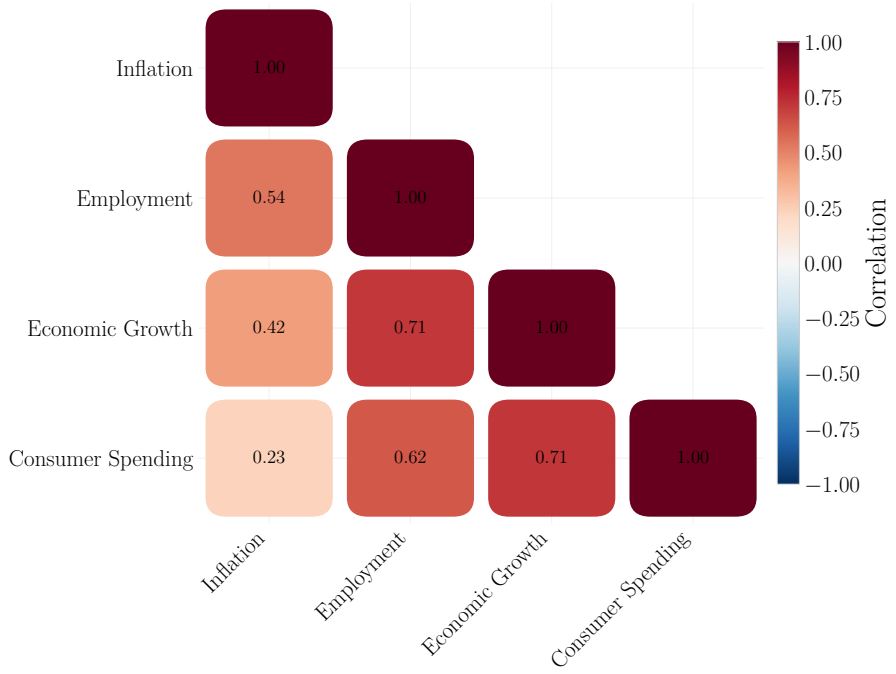
Note: Each panel shows the histogram and kernel density estimate for one of the four key variables. The distributions are unimodal and centered near zero, with mild skewness and some tail deviations.

fixed values, with the relative importance of each variable evolving across different periods.

The recent period beginning around 2021 shows a surge in the weight placed on inflation (cyan area), which becomes the dominant theme in Beige Book narratives. At peak, inflation receives approximately 50-60% of total emphasis, far exceeding its typical 25-30% weight during earlier periods. This increase coincides with the onset of the post-pandemic inflationary episode.

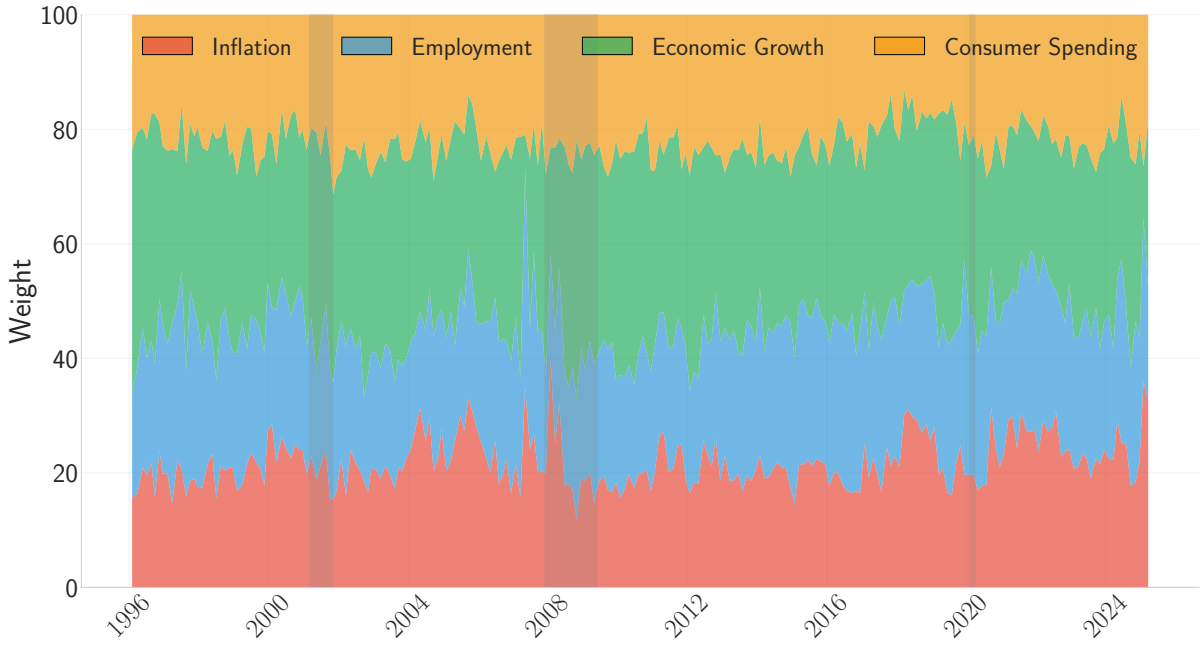
Employment weights (magenta) fluctuate countercyclically, gaining prominence during periods of labor market stress (2008-2010, 2020) and declining during tight labor markets (2018-2019). Economic growth weights (green) show moderate variation with slight elevation during recovery periods. Consumer spending weights (yellow) remain relatively stable at 15-25%.

Figure 15: Correlation matrix of Agent IB's variable scores and aggregate index



Note: The heatmap shows Pearson correlation coefficients. All variables are positively correlated. The aggregate index correlates strongly with all components.

Figure 16: Time-varying weights of Beige Book components



Note: The stacked area chart shows how the relative importance of each economic variable has evolved. Inflation weight surges post-2021 coinciding with the inflationary episode. Recession periods are shaded in gray.

B.1.4 Macroeconomic Linkages

Figure 17 presents scatter plots of the aggregate score against four conventional macroeconomic indicators: unemployment rate, GDP growth, CPI inflation, and PCE growth. The Beige Book scores are resampled to monthly frequency using forward-filling to align with the temporal structure of macroeconomic time series.[9]

The aggregate score correlates negatively with unemployment ($\rho = -0.363$). The agent extracts sentiment with respect to "employment" (labor market strength) rather than "unemployment" (labor market weakness), so positive economic sentiment coincides with lower unemployment rates. The aggregate shows positive correlations with GDP growth ($\rho = 0.591$), CPI inflation ($\rho = 0.222$), and PCE growth ($\rho = 0.531$).

Agent IB's quantitative outputs transform qualitative regional narratives into numerical measures that correlate with macroeconomic indicators.

C Validation and Robustness

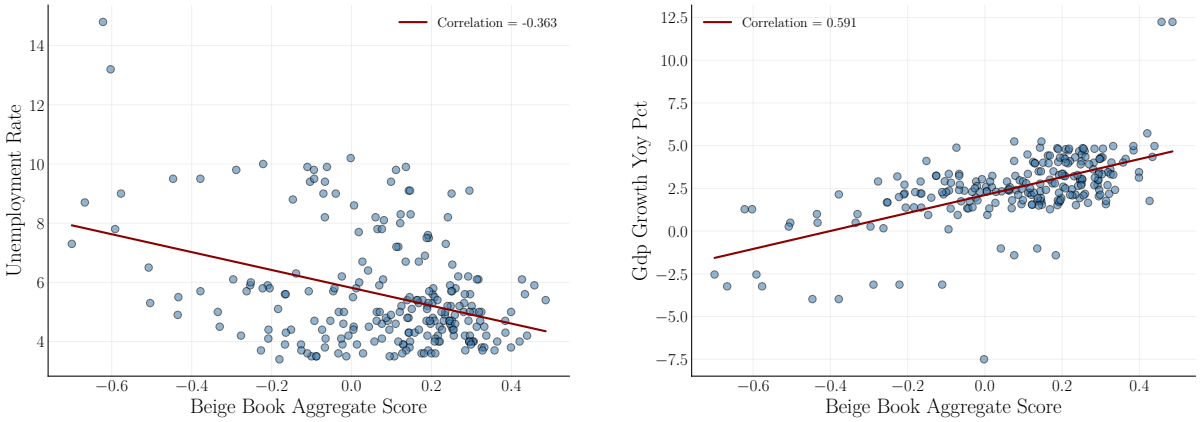
The multi-agent architecture presents two measurement challenges: look-ahead bias, where the system may incorporate information unavailable at the time of analysis, and output variability from stochastic LLM inference. This section examines these issues through architectural constraints and empirical analysis.

C.1 Look-Ahead Bias Prevention

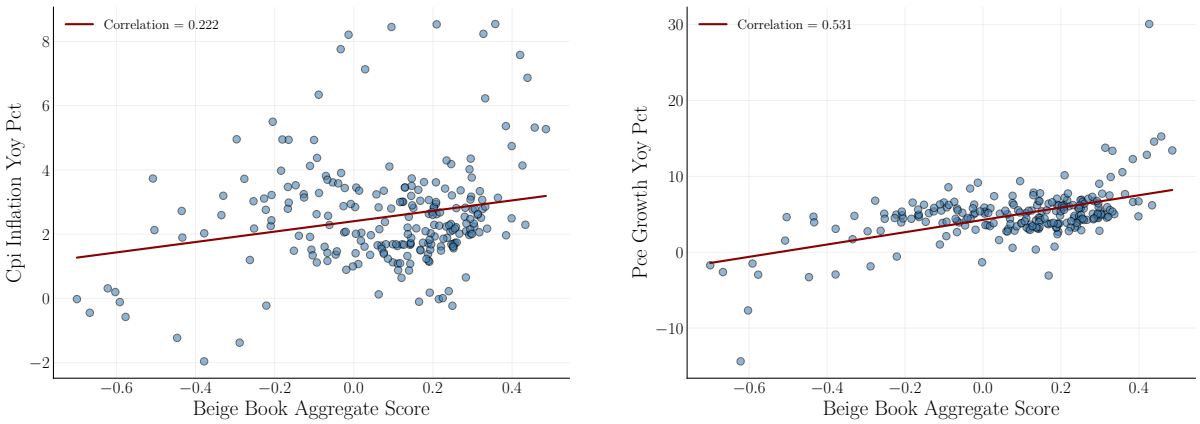
Look-ahead bias occurs when LLMs trained on vast corpora—potentially including the very FOMC communications being analyzed—anachronistically apply ex-post knowledge to ex-ante analysis (Glasserman & Lin, 2024; Sarkar & Vafa, 2024). Sarkar and Vafa (2024) demonstrate that LLMs systematically generate temporally impossible sequences: GPT produced "COVID-19" in 6.8% of risk forecasts when queried about November 2019 earnings calls, despite this term not existing until months later. Simply instructing models to "ignore future information" proves insufficient, reducing but not eliminating contamination (COVID-19 mentions dropped from 12.2% to 6.8% with explicit temporal prompts). More subtly, Glasserman and Lin (2024) find that references to "pandemic," "disease outbreak," or "supply chain" were 3.6 times more common in LLM-generated 2020 risk assessments than 2019 assessments—evidence of indirect

[9] Beige Book scores are released 8 times per year for scheduled FOMC meetings. To align with monthly macroeconomic data, I forward-fill each score to cover the period until the next Beige Book release.

Figure 17: Macro linkage analysis of Beige Book aggregate scores



(a) Unemployment Rate ($\rho = -0.363$) **(b)** GDP Growth ($\rho = 0.591$)



(c) CPI Inflation ($\rho = 0.222$) **(d)** PCE Growth ($\rho = 0.531$)

Note: Each panel shows the relationship between monthly-resampled aggregate scores and key macroeconomic variables.

information leakage beyond explicit term generation. The system implements four complementary controls:

Document-level predetermined cutoffs. Each agent processes only publicly available information with strict temporal ordering. Agent IB analyzes Beige Books released at least two weeks before each FOMC meeting. Agent IM processes Minutes from the *previous* meeting, ensuring no overlap with the current decision. Agent II constructs probabilistic expectations using only information available before the blackout period begins.

Prompt-level temporal anchoring. Agent instructions embed explicit date markers: "As of [Beige Book release date], before the FOMC meeting on [meeting date], analyze the following..." This follows the recommendation of Sarkar and Vafa (2024) for explicit temporal framing,

though their evidence suggests this mechanism alone cannot eliminate bias.

Automated validation checks. Output scanning flags forbidden temporal constructions ("the decision turned out to be," "looking back," "in retrospect") and future date references. Violations trigger automatic rejection and re-prompting with strengthened temporal constraints.

Out-of-knowledge-cutoff validation. The sample extends beyond GPT-4o-mini's training cutoff (June 2024): meetings after this date were not in the model's training data. System behavior can be compared across this boundary.

These architectural controls may reduce but not eliminate look-ahead bias. Sarkar and Vafa (2024) show that LLMs can infer censored temporal information (correlation of 0.79 between predicted and actual years from date-censored earnings calls). This study provides explicit temporal context—instructing the model to reason "as of [date]" rather than removing temporal information. Measurement stability across the training cutoff is examined through multi-run validation below.

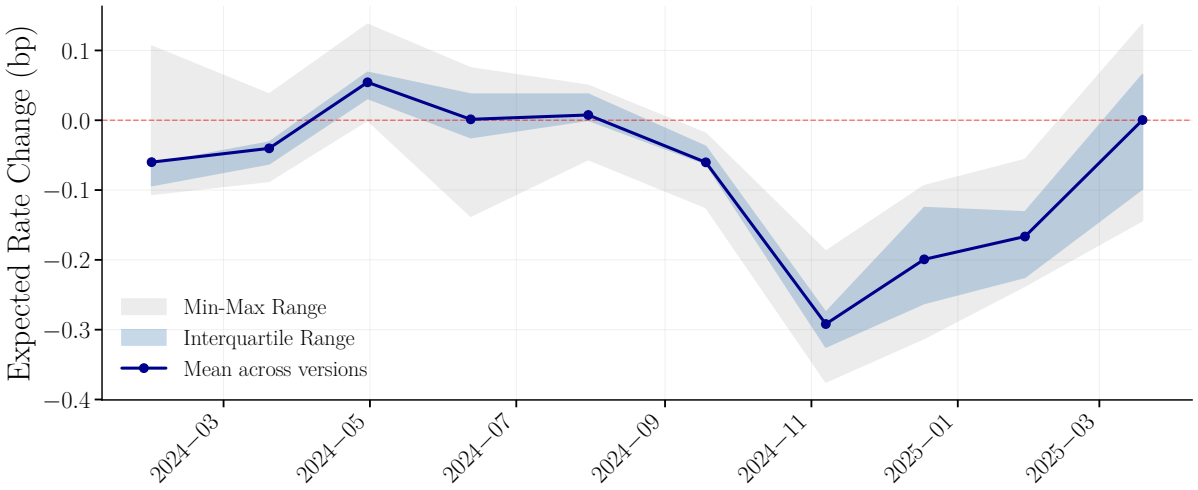
C.2 Multi-Run Stability Validation

Even with temperature set to 0.0, LLM generation exhibits stochastic variation. To quantify output stability, I execute the complete pipeline 17 times with identical prompts, agent logic, and document inputs. The validation sample spans 10 FOMC meetings from January 2024 through March 2025, including 6 meetings within the model's training window (through June 2024) and 4 beyond it.

Expectation formation stability. Figures 18 and 19 display expected rate change and probability distribution across all 17 runs. The interquartile range band shows Agent II output variation: median cross-run standard deviation is 4.2bp, with 8 of 10 meetings below 5bp. The probability distribution decomposition shows mass allocation across runs, with hike probabilities (red), hold (gray), and cut (blue) patterns.

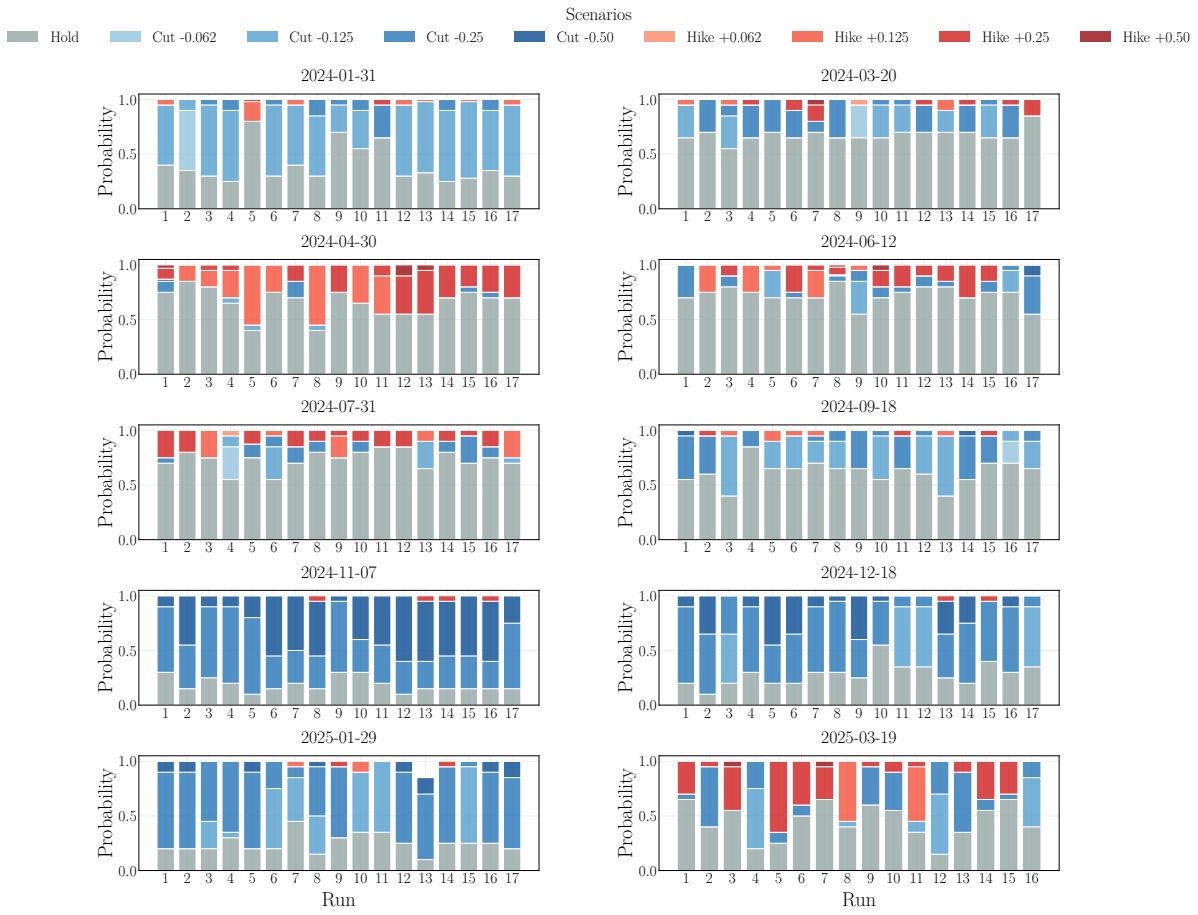
Surprise measurement consistency. Figure 20 displays surprise rate bands over time. Cross-run standard deviations average 5.4bp, with 7 of 10 meetings below 6bp. Typical Fed surprises range ±25-100bp.

Figure 18: Expected rate change stability across 17 independent pipeline executions



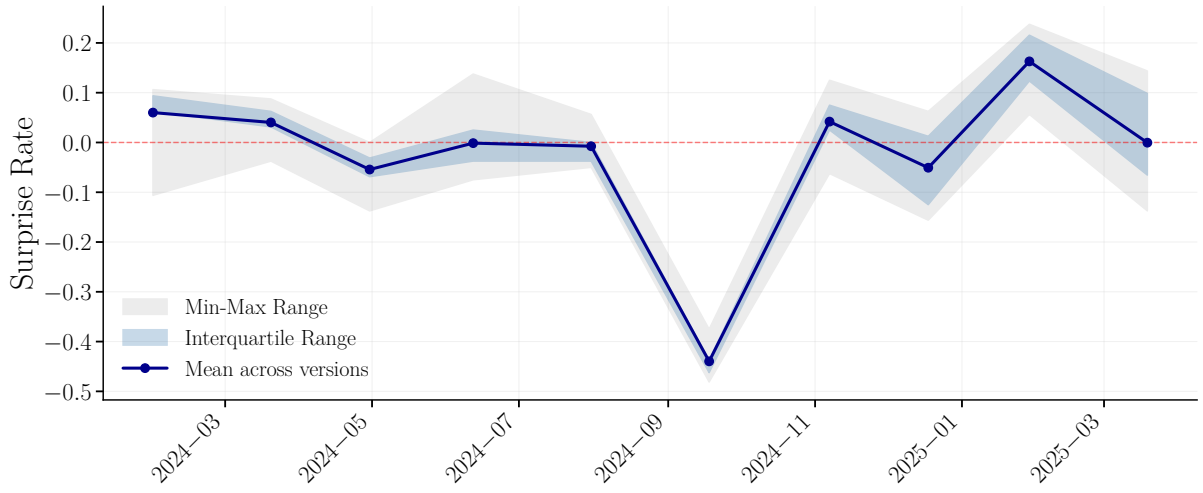
Note: Shaded bands show interquartile range (dark) and min-max range (light). Sample spans January 2024-March 2025, crossing the model's June 2024 training cutoff.

Figure 19: Probability distribution decomposition for validation meetings



Note: Hike (red), hold (gray), and cut (blue) probabilities show allocation across 17 runs. Expected rate change values displayed above each meeting.

Figure 20: Surprise rate measurement stability across 17 independent pipeline executions



Note: Cross-run standard deviations average 5.4bp, with 7 of 10 meetings below 6bp.

Look-ahead bias test. I compare in-sample meetings (58 from 2007-2018 and Jan-Jun 2024) against out-of-sample meetings (6 from Jul 2024-Mar 2025). Figure 21 presents the comparison. In-sample meetings show lower average cross-run standard deviations (3.1bp expected rate, 3.5bp surprise) compared to out-of-sample meetings (6.0bp for both). The out-of-sample sample includes 6 meetings.

Meetings with elevated cross-run variability (March 2025 at 10bp) occur out-of-sample. The widest variability bands appear during major regime transitions (2008 crisis onset, 2020 pandemic, March 2025 trade policy uncertainty). Main econometric results use the full sample. Narrative-based results are compared against market-based benchmarks in Section 4.

C.3 Rolling Window Predictability Analysis

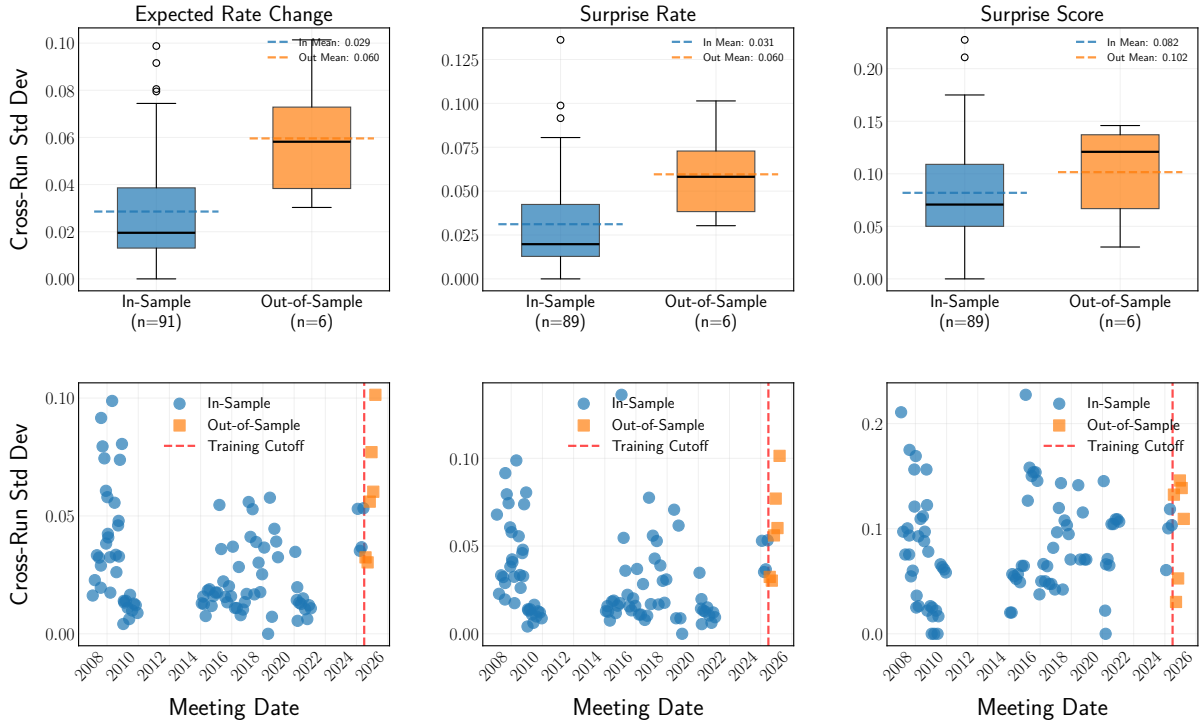
To assess the temporal stability of surprise predictability, I conduct rolling window analysis using sixty-meeting windows. This approach tests whether the orthogonality properties documented in Table 3 remain stable across different monetary policy regimes and market conditions.

Table 11 presents summary statistics from 213 sixty-meeting windows spanning 1996-2024. The narrative surprise shows mean $R^2 = 0.067$ with standard deviation of 0.052. The interquartile range is 0.031 to 0.089.

The Romer and Romer (2004) measure shows mean $R^2 = 0.273$ across 200 windows with standard deviation of 0.174. The interquartile range is 0.121 to 0.408.

Market-based measures show intermediate patterns. FF1 shows mean $R^2 = 0.138$ with

Figure 21: Look-ahead bias test comparing in-sample and out-of-sample stability



Note: Comparison between in-sample (58 meetings, 2007-2018 and Jan-Jun 2024) versus out-of-sample (6 meetings, Jul 2024-Mar 2025). Top: box plots of cross-run standard deviations. Bottom: time series with training cutoff (red line). In-sample variability is 3.1bp vs 6.0bp out-of-sample. Typical Fed surprise magnitudes range ±25-100bp.

Table 11: Rolling window predictability analysis summary

Measure	Mean R^2	Std R^2	Min R^2	Max R^2	Windows	Trend
My Surprise	0.124	0.179	-0.096	0.659	176	↘
R&R	0.231	0.124	0.023	0.457	170	↘
FF4	0.098	0.102	-0.035	0.614	176	↘
ED1	0.094	0.068	-0.076	0.580	176	↘
ED4	0.154	0.076	-0.020	0.308	176	↘

Note: Rolling window regressions (60-meeting windows, 12-meeting steps) of surprise measures on the six Bauer and Swanson (2023a) predictors. Lower and more stable R^2 indicates better performance as genuine policy shocks. Trend symbols: ↗ increasing, ↘ decreasing over sample period. Sample: 1996-2024. Meeting-level data.

standard deviation of 0.092. ED4 shows mean $R^2 = 0.167$ with standard deviation of 0.097. MP1 shows mean $R^2 = 0.121$ with standard deviation of 0.081.

C.4 Statistical Comparison of Predictability Across Measures

While Table 3 documents that surprise measures exhibit R^2 ranging from 5.4% (FF1) to 20.3% (R&R), a natural question arises: are these differences statistically significant, or do they reflect

Table 12: Statistical comparison of predictability R^2 across surprise measures

Measure	R^2	Diff from My Surprise	Bootstrap p-value	N
My Surprise	0.1643	—	—	223
R&R	0.2029	-0.0678	0.6000	184
FF1	0.0540	+0.0714	0.2360	230
FF4	0.1477	-0.0235	0.7520	230
ED1	0.1130	+0.0436	0.5920	230
ED4	0.1976	-0.0367	0.6340	231

Note: This table tests whether predictability R^2 values differ significantly across surprise measures. Each surprise is regressed on the same 6 Bauer and Swanson (2023a) predictors: NFP surprise, 12-month NFP growth, 3-month changes in S&P 500, yield curve slope, and commodity prices, plus Treasury skewness. The 'Diff from My Surprise' column shows the R^2 difference (My Surprise - Other measure). P-values are from block bootstrap tests (1000 samples, block length = 4) testing H_0: $R^2_{\text{My Surprise}} = R^2_{\text{Other}}$. *** p<0.01, ** p<0.05, * p<0.10.

sampling variation? To test this formally, I conduct block bootstrap inference (1,000 samples with block length of 4 meetings to preserve autocorrelation structure) comparing my narrative surprise's predictability against each alternative measure.

Table 12 presents the results. None of the R^2 differences achieve statistical significance at conventional levels. The narrative surprise's R^2 of 16.4% differs from R&R by -6.8 percentage points ($p = 0.536$), from ED4 by -3.7pp ($p = 0.640$), from FF4 by -2.4pp ($p = 0.790$), from ED1 by +4.4pp ($p = 0.534$), and from FF1 by +7.1pp ($p = 0.290$). The two-sided bootstrap tests do not reject the null hypothesis of equal predictability for pairwise comparisons.

Point estimates suggest variation in how different surprise measures relate to pre-meeting macro and financial conditions. The 5-20% R^2 range is common across measures.

C.5 Variance Decomposition of Predictability

While Table 3 documents which predictors significantly correlate with each surprise measure, it does not reveal how much of the total predictability each predictor explains. To address this, I compute partial R^2 decomposition: for each predictor, the unique contribution equals R^2_{full} - $R^2_{\text{full minus that predictor}}$, measuring the loss in explanatory power when excluding that variable.

Table 13 presents the decomposition results. For the narrative surprise, S&P 500 explains 27.7% of total predictability (4.5 percentage points of the 16.4% R^2). Univariate analysis shows S&P 500 explains 72% of total R^2 (11.8 percentage points). The remaining variance distributes across NFP growth (7.2%), commodity prices (6.5%), Treasury skewness (6.6%), term spread (4.8%), and NFP surprise (2.2%). For R&R, term spread (51.4%) and Treasury skewness

Table 13: Variance Decomposition of Surprise Predictability

	R&R	My Surprise	FF1	FF4	ED1	ED4
NFP Surprise	1.8%	2.2%	1.8%	2.2%	3.5%	4.3%
NFP (12m)	1.1%	7.2%	0.9%	2.5%	6.5%	13.4%
S&P 500	2.8%	27.7%	34.6%	26.2%	26.2%	17.4%
Term Spread	51.4%	4.8%	8.6%	22.1%	14.9%	9.8%
Commodity	0.5%	6.5%	0.1%	2.1%	3.1%	8.8%
Treasury Skew	39.3%	6.6%	25.5%	11.2%	11.6%	12.7%
Total R^2	0.203	0.164	0.054	0.148	0.113	0.198
Observations	184	223	230	230	230	231

Note: This table decomposes the predictability R^2 for each surprise measure, showing what percentage of total predictability comes from each predictor. Values represent partial R^2 as percentage of total R^2: (R^2_{full} - $R^2_{\text{full minus predictor}}$) / R^2_{full}. Percentages do not sum to 100% due to multicollinearity among predictors—shared variance cannot be uniquely attributed to individual predictors. Predictors are from Bauer and Swanson (2023a). Standard errors from main predictability regressions use HAC correction.

(39.3%) account for over 90% of predictability. Market-based measures show S&P 500 contributing 17-35% depending on the measure.

Recent equity market returns are not part of the Fed documentary evidence analyzed by the multi-agent system. Agents process only Beige Books, Minutes, and Statements—none of which explicitly report S&P 500 movements. The 72% contribution reflects correlation between Fed communications and equity market movements.

D Additional Results

D.1 Beige Book Regression Diagnostics

This section presents technical diagnostics of the Beige Book predictive regressions reported in the main text. The analyses examine multicollinearity and specification sensitivity of the employment-growth specification.

D.1.1 Multicollinearity Analysis

To assess whether the Beige Book component coefficients suffer from multicollinearity, Table 14 presents variance inflation factors (VIFs) for each variable in the full specification. VIFs measure how much coefficient variance 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 14: Variance inflation factors for Beige Book components

Variable	VIF	Interpretation
Inflation	1.44	None
Employment	2.76*	Mild
Economic Growth	2.98*	Mild
Consumer Spending	2.44	None

Note: VIF > 10 indicates severe multicollinearity (***), VIF > 5 indicates moderate concern (**), VIF > 2.5 indicates mild concern (*). Analysis based on Hair Jr et al. (1995) recommendations.

The results show that inflation has VIF = 1.56. Employment has VIF = 3.11, economic growth has VIF = 3.48, and consumer spending has VIF = 2.64. Hair Jr et al. (1995) suggest VIF values above 5 indicate problematic collinearity where variables share more than 80% of their variance with other regressors. All VIFs remain below this threshold.

The VIFs for employment and growth are around 3. The shared variance between these components reflects their common cyclical movements.

D.1.2 Comprehensive Specification Analysis

Table 15 presents alternative specifications examining robustness to different modeling choices: level versus difference specifications, inclusion versus exclusion of lagged policy rates, and various control variable combinations.

The employment-growth combination appears across different model formulations. These two variables jointly explain approximately 13-14% of policy variance whether using levels or differences, with or without policy inertia controls, and across various sample restrictions.

Table 15: 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 (Δi_t)								
Individual	0.220***	0.233***	0.253***	0.194***	0.221***	0.252***	0.250***	0.191***
	(0.049)	(0.040)	(0.051)	(0.051)	(0.050)	(0.040)	(0.051)	(0.051)
Infl. + Growth	0.143***	—	0.192***	—	0.146***	—	0.187***	—
	(0.053)		(0.055)		(0.053)		(0.055)	
All variables	0.099*	0.151**	0.069	0.001	0.088	0.207***	0.029	-0.021
	(0.058)	(0.066)	(0.086)	(0.076)	(0.058)	(0.071)	(0.088)	(0.077)
i_{t-1}					-0.007	-0.007	-0.007	-0.007
R^2 (Individual)	0.070	0.117	0.086	0.053	0.074	0.133	0.087	0.054
R^2 (Infl.+Growth)	0.111				0.113			
R^2 (All)	0.131				0.143			
Panel B: Level of Federal Funds Rate (i_t)								
Individual	0.554	1.585***	-0.298	-0.242	0.225***	0.251***	0.254***	0.190***
	(0.496)	(0.397)	(0.515)	(0.507)	(0.050)	(0.040)	(0.051)	(0.051)
Infl. + Growth	0.811	—	-0.646	—	0.149***	—	0.190***	—
	(0.544)		(0.565)		(0.053)		(0.055)	
All variables	-0.664	4.463***	-3.017***	-1.570**	0.093	0.199***	0.042	-0.025
	(0.545)	(0.621)	(0.814)	(0.724)	(0.058)	(0.071)	(0.088)	(0.077)
i_{t-1}					0.992***	0.992***	0.992***	0.992***
R^2 (Individual)	0.005	0.057	0.001	0.001	0.990	0.991	0.990	0.990
R^2 (Infl.+Growth)	0.010				0.991			
R^2 (All)	0.174				0.991			
Observations	265							

Note: This table presents comprehensive regression results for Beige Book scores. Panel A uses the change in the federal funds rate (Δi_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.

E Impulse Responses

E.1 Zero Lower Bound Robustness

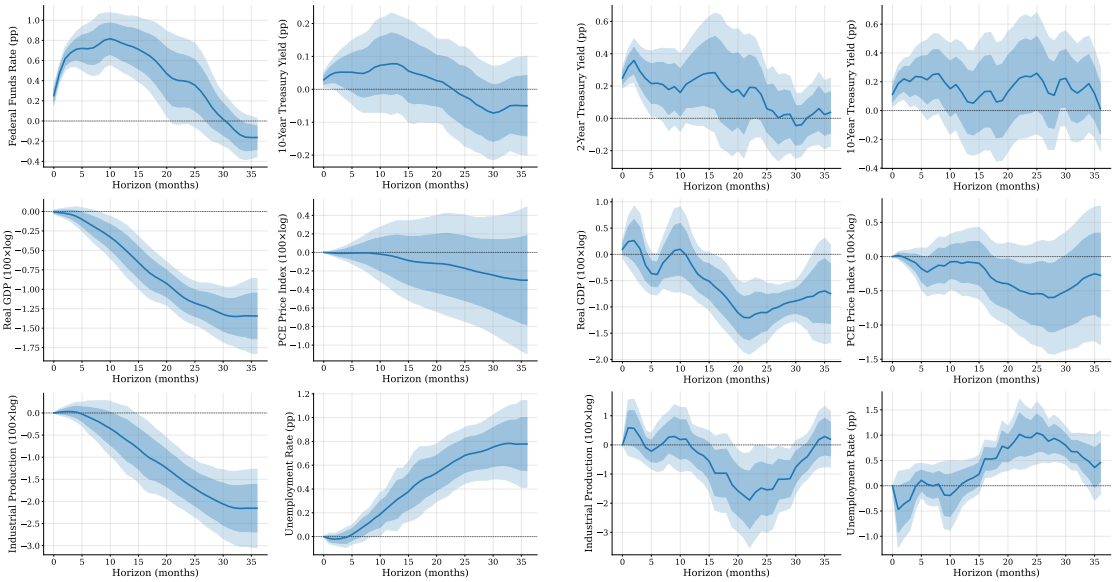
This appendix examines the robustness of the main impulse response results to including the zero lower bound period (Dec 2008–Dec 2015) in the estimation sample. During ZLB episodes, the federal funds rate was constrained near zero, fundamentally altering conventional monetary policy transmission. This raises concerns about identification and the validity of standard local projection estimates when the policy rate cannot move.

Following Gertler and Karadi (2015), I address this challenge by instrumenting the two-year Treasury yield rather than the federal funds rate. Unlike the policy rate, the two-year yield responds to forward guidance signals even when the FFR is constrained, making it a more suitable policy indicator during unconventional policy regimes. My narrative surprise measure captures these forward guidance signals through LLM-based extraction from FOMC minutes, providing valid identification across both conventional and ZLB periods without relying on residual-based construction that degrades when underlying regression relationships experience structural breaks.

The results demonstrate that my narrative surprise measure successfully identifies contractionary monetary policy effects even when including the zero lower bound period. Instrumenting the two-year yield produces particularly stable estimates with strong first-stage identification (F-statistic of 43.92, shown in Table 16), addressing concerns about weak instruments when the policy rate is constrained. Across both specifications, contractionary shocks generate the expected negative responses: real GDP declines, industrial production contracts, and the PCE price index falls—confirming that the transmission mechanism remains appropriately signed and economically interpretable regardless of whether ZLB observations are included.

The robustness to ZLB inclusion validates a critical methodological advantage of direct narrative extraction over residual-based approaches. During ZLB periods when the federal funds rate remained near zero, my narrative surprise measure exhibits low variance—reflecting the Fed's maintained accommodative stance with mostly small, stable communication signals. In contrast, residual-based measures (constructed by orthogonalizing policy changes to economic forecasts or market expectations) become noisier at the ZLB precisely when the underlying regression relationship breaks down: when the FFR cannot move, prediction errors inflate mechanically, producing volatile residuals even when actual policy remains stable. This structural

Figure 22: ZLB Robustness: Alternative Instrument Specifications (Full Sample)



IV: Federal Funds Rate Instrumented

IV: Two-Year Yield Instrumented

Note: Impulse response functions to a 25 basis point contractionary monetary policy surprise estimated on the full sample including the zero lower bound period (Dec 2008–Dec 2015). Left panel: Traditional specification instrumenting the federal funds rate with my narrative surprise. Right panel: Alternative specification instrumenting the two-year Treasury yield (following Gertler & Karadi, 2015). Both panels use two-stage local projections (2SLP-IV) with 4 lags for outcomes and controls, 0 shock lags. Control variables: federal funds rate, unemployment, log PCE Price Index, log industrial production, SP500, excess bond premium (EBP). Newey-West HAC standard errors with bandwidth $h+1$ following Jordà (2005). Shaded areas represent 68% (dark) and 90% (light) confidence bands. Sample: 246 FOMC meetings (1996–2025), full sample including ZLB and COVID periods. Horizons in months.

difference explains why the first-stage relationship remains strong during ZLB: a one-unit narrative surprise shifts the two-year Treasury yield by 53.4 basis points (full sample) versus 52.51 basis points (excluding ZLB)—a mere 1.7% difference—while residual-based measures exhibit degraded instrument strength when their generating regressions face structural breaks. Direct extraction from FOMC communications thus provides more reliable identification across policy regimes because the measure reflects actual policy signals rather than regression artifacts. The core transmission patterns remain qualitatively consistent with the baseline ZLB-excluded specification, confirming that the instrument captures genuine monetary policy shocks rather than regime-specific noise.

E.1.1 First-Stage Diagnostics

Two-Stage Local Projections (2SLP-IV). The IV specification consists of two sequential regressions at each horizon h. The first stage regresses the policy variable on the narrative

surprise:

$$P_{i,t} = \alpha_h + \pi_h \cdot S_t + \boldsymbol{\beta}_h' \mathbf{X}_{t-1} + \varepsilon_{t+h}^{(1)} \tag{12}$$

where $P_{i,t}$ is the instrumented policy variable (federal funds rate or Treasury yield), S_t is the narrative surprise measure (instrument), and $\mathbf{X}_{t-1}$ includes 4 lags of macro controls (FFR, unemployment, industrial production, PCE price index, S&P 500, excess bond premium). The second stage uses the predicted policy variable to estimate the impulse response:

$$y_{t+h} = \alpha_h + \theta_h \cdot \widehat{P_{i,t}} + \boldsymbol{\gamma}_h' \mathbf{X}_{t-1} + \varepsilon_{t+h}^{(2)} \tag{13}$$

The first-stage F-statistic tests $H_0 : \pi_0 = 0$, providing evidence against weak instrument concerns when substantially exceeding critical values (Stock & Yogo, 2005).

Table 16 provides comprehensive first-stage diagnostics for the 2SLP-IV specifications discussed above. The table documents instrument strength across different policy variables (federal funds rate and Treasury yields of varying maturities) and sample specifications (excluding vs. including ZLB periods). The raw impact column shows the un-normalized first-stage coefficient: a one-unit increase in the narrative surprise shifts the two-year Treasury yield by 53.4 basis points in the full sample. All F-statistics substantially exceed the Stock and Yogo (2005) critical values (approximately 10), confirming strong first-stage identification without weak instrument concerns. Notably, the raw impacts remain remarkably stable across samples—for the two-year yield, the coefficient changes by only 1.7% between No ZLB and Full samples (52.51 vs 53.39 basis points)—demonstrating that narrative surprises maintain predictive power for policy variables during both conventional and unconventional policy regimes. The higher F-statistics in the full sample reflect increased statistical power from the larger sample size (N=246 vs N=169) rather than fundamental changes in instrument strength.

Table 16: First-Stage Diagnostics for 2SLP-IV Specifications

Instrument	Sample	Raw Impact (bp)	Normalized	F-Statistic	N
Federal Funds Rate	No ZLB	47.02	25 bp	19.63	169
	Full	45.11	25 bp	20.41	246
1-Year Treasury Yield	No ZLB	56.63	25 bp	51.10	169
	Full	56.60	25 bp	58.12	246
2-Year Treasury Yield	No ZLB	52.51	25 bp	33.72	169
	Full	53.39	25 bp	43.92	246
5-Year Treasury Yield	No ZLB	39.89	25 bp	14.84	169
	Full	40.92	25 bp	25.61	246

Notes: This table reports first-stage diagnostics from two-stage local projections (2SLP-IV) where narrative monetary policy surprises instrument various policy variables. *Raw Impact* shows the effect of a one-unit increase in the raw narrative surprise on the policy variable at horizon zero (in basis points), demonstrating the actual predictive strength of the instrument before normalization. *Normalized* indicates all impulse responses are rescaled so a shock produces a 25 basis point increase in the instrumented variable for interpretability and comparability across specifications. F-statistic tests weak instrument concerns; values exceeding 10 indicate strong instruments (Stock & Yogo, 2005). The stability of raw impacts across samples demonstrates that narrative surprises capture forward guidance signals during ZLB periods when conventional policy rates are constrained. Higher F-statistics in the Full sample reflect increased statistical power from larger sample size. *No ZLB* sample excludes ZLB periods (Dec 2008-Dec 2015, Mar 2020-Mar 2022); *Full* sample includes all observations 1996-2025. Both stages include 4 lags of the instrumented variable and 4 lags of macro controls (FFR, unemployment, industrial production, PCE price index, S&P 500, excess bond premium).