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: May 13, 2026

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 analysing Statements, Minutes, Beige Books, and press conferences released before each FOMC meeting, the system elicits conditional expectations that yield less noisy surprises than market-based measures. These surprises produce theoretically consistent impulse responses, with contractionary shocks generating persistent disinflation, and carry directional information about the policy path that high-frequency announcement surprises miss.

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, Martin Fankhauser, Nicola Gennaioli, Nicolás Guíñez, Alejandra Inzunza, Mohammad R. Jahan-Parvar, Tommaso Monacelli, David Murakami, Fernando Pérez-Cruz, Angelo Ranaldo, Damiano Sandri, Kevin Schneider, Ivan Shchapov, Jakob Ahm Sørensen, Isabella M. Wolfskeil for their insightful comments and suggestions. Special thanks go to participants of the following conferences and workshops: ASSA/AFA Annual Meeting, AI and Society Conference (Bocconi University), Wolfe Research European Quantitative and Macro Investment Conference, Conference on Artificial Intelligence in the Macroeconomy (University at Albany/UCSB-LAEF), Workshop in Empirical Macroeconomics (University of Innsbruck), 1st Lausanne PhD Macroeconomics Conference (HEC Lausanne), and PhD Alumni Conference (Bocconi University); and to seminar participants at the Bank of England, Bank of Spain, Bank of Italy, CUNEF, Bilkent University, HEC Montréal, and the Bank for International Settlements. 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: ruben-fernandezfuertes.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 shock 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 Statements, press conferences, Minutes, and Beige Books 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. Across 272 FOMC meetings from 1996 to 2026, this nonparametric, nonlinear extraction of conditional expectations from narrative text yields less noisy surprises that explain 46.5% of policy-rate variation, roughly three times the 14.5–16.9% explained by standard market-based measures. The residual variation captures the part of policy decisions that was not predictable from the Fed's public communications. By measuring surprises against the information set frozen weeks before the meeting, I isolate what was genuinely unexpected from the Fed's public messaging, while market measures incorporate all information up to announcement moments and may include flows that contaminate identification.

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 pre-meeting 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's blackout period, my surprises are predetermined relative to announcement-day information flows, mitigating the contamination issues that affect high-frequency measures. An optional news stage can incorporate inter-meeting financial-press coverage during the pre-FOMC blackout, further reducing the predictability of the surprise from standard pre-meeting controls.

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 public pre-meeting documents, 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.

Three sets of findings characterize the resulting surprise: its measurement properties, its macroeconomic transmission, and the directional content it carries beyond standard high-frequency measures. First, the surprises exhibit minimal measurement error: realised rate changes move nearly one-for-one with the model-implied surprise, consistent with approximately Bayesian updating and small contamination leakage. The remaining gaps from the Fed's full information set are quantifiable rather than hidden: internal Greenbook forecasts add explanatory power beyond my measure, quantifying a private-information wedge between public documents and the Fed's internal staff forecasts, while the news stage absorbs the public information that arrives during the blackout window. The measure also carries genuine path information: it predicts subsequent rate changes but not subsequent surprises, a pattern consistent with forward-guidance content rather than stale expectations.

Second, impulse-response analysis (Ramey, 2016) reveals transmission patterns that are the-

oretically coherent across macroeconomic and financial variables. Following a contractionary surprise, price levels decline immediately and persistently, avoiding the "price puzzle" that often plagues market-based identifications, real GDP and industrial production show sustained contractionary effects, and unemployment rises with the expected delay. The consistent contractionary signs across all variables distinguish my measure from market-based alternatives that often produce sign anomalies requiring *ex post* instrumental-variable corrections. The yield curve decomposes into two transmission channels, an initial spread compression driven by rising expected short rates, followed by a delayed steepening as the policy cycle is digested by the expected path, revealing that the surprise affects both expectations and risk compensation.

Third, the residual carries directional information about the policy path that the four standard announcement-window surprises (FF1, FF4, ED1, ED4) do not linearly span. A span test rejects 81.5% of the LLM-surprise variance as orthogonal to those four factors, and a yield-curve local projection localises that orthogonal content to the front of the curve, where forward guidance is priced. An asset-pricing test built from the local-projection sign change (a 1m/2y equal-notional flattener signed by the surprise) delivers per-trade yield-spread returns that survive a battery of placebo, threshold, weighting, maturity-pair, and cross-LLM robustness checks, and an orthogonal-residual variant is statistically indistinguishable from the raw signal, so the payoff is not a re-projection of the linear high-frequency span. The exercise is measurement validation, not tradable alpha: the result is concentrated in active rate-cycle episodes and on the dovish side, and the paired LLM–ED4 horizon comparison is statistically weak on the small paired sample — the short-horizon gap does not reach significance even before any multiple-testing correction.

The multi-agent LLM framework processes the Federal Reserve's public communication timeline to form probabilistic expectations before each FOMC decision. Before each meeting, four documents become publicly available in sequence: the previous meeting's FOMC Statement and Chair's press conference, the Minutes from that meeting (released several weeks later), and the current Beige Book (released roughly two weeks before the upcoming meeting). This structured communication system motivates a multi-agent pipeline: dedicated decoders quantify the policy-relevant content of each document type, a forecaster synthesises these inputs through sequential probability updates into a distribution over potential Fed actions, and a surprise extractor compares the final pre-meeting prior with the realised decision. An optional news-stage agent fills the pre-FOMC blackout window using inter-meeting financial-press coverage. This architecture

mirrors the Fed's actual communication timeline rather than treating documents in isolation.

Following the sampling-and-voting evidence in Li et al. (2024), and drawing on the broader multi-agent LLM literature surveyed in 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. Section 3 formalises the resulting decomposition: the measured surprise equals a policy innovation relative to the Fed's internal information, plus a private-information wedge, plus a public non-document wedge, plus extraction error, so the residual gaps are explicit rather than hidden. 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 retained predictability 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, 2021) or misspecified reaction functions (Bauer & Swanson, 2023a, 2023b) (henceforth, B&S), with recent evidence from Ricco and Savini (2025) favouring 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 identification. 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 (De Fiore et al., 2024; Gambacorta et al., 2024; Hansen & Kazinnik, 2023) focuses on characterising tone rather than constructing counterfactual expectations for surprise measurement. A separate strand maps central-bank text to forecast revisions via supervised machine-learning models trained on Greenbook texts (Ahrens & McMahon, 2021; Ahrens et al., 2024); these methods extract quantitative signals from speeches but do not construct meeting-level priors against which to measure surprise. 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 analyse 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 (Statements, press conferences, Minutes, and Beige Books released before blackout periods) to address 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 the sampling-and-voting evidence in Li et al. (2024), alongside the broader multi-agent literature surveyed in 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 272 FOMC meetings' worth of Statements, press

conferences, Minutes, and Beige Books from 1996 to 2026, 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 and the optional news stage. Section 3 develops the signal-extraction framework that formalises the wedge decomposition. Section 4 presents the data sources. Section 5 validates measurement properties of the narrative surprise. Section 6 extends to macroeconomic and financial impulse responses and a yield-curve asset-pricing diagnostic of the surprise's residual content. Section 7 discusses implications for monetary economics research and the application of Large Language Models to systematic document processing.

2 Methodology

2.1 General Framework

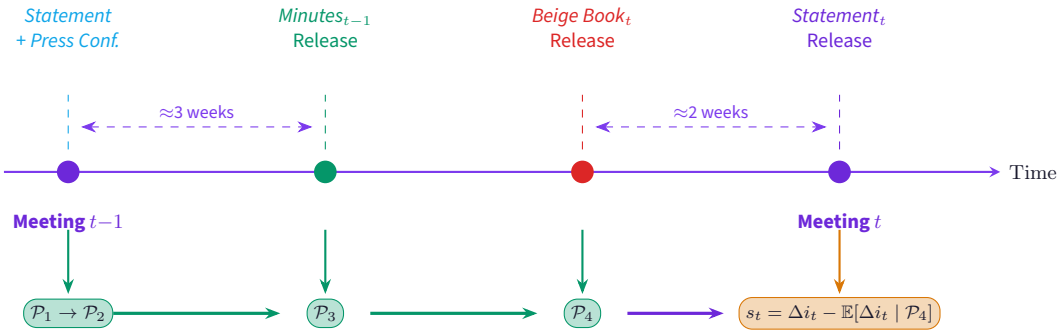
The Federal Reserve's staggered document release schedule provides a natural sequence for updating expectations over the next FOMC rate decision as each official communication becomes public. For each of the eight *scheduled* FOMC meetings per year, up to four key documents become publicly available in sequence:[1] the *FOMC Statement* from meeting $t-1$, announcing the policy decision and forward guidance language; the *Press Conference* from the same day, consisting of the Chair's Q&A with journalists;[2] the *Minutes* released approximately three weeks after meeting $t-1$, documenting deliberations and the distribution of views; and the *Beige Book* released approximately two weeks before meeting t, containing qualitative economic assessments from the 12 Districts. I report a schematic of the FOMC's communication releases in Figure 1.

This temporal structure motivates a pipeline that mirrors the sequential release of information rather than processing all documents simultaneously. Dedicated *Decoder* agents extract structured signals from each document as it becomes publicly available, and these signals are fed into a *Forecaster* that maintains a probability distribution over the upcoming rate decision. Each document expands the information set, constituting a filtration: at stage k, the Forecaster

[1] The sample also includes unscheduled emergency decisions (e.g., January 2008, March 2020). For these meetings the standard document sequence is incomplete: no Beige Book precedes the decision and the Minutes from the prior meeting may not yet have been released. The pipeline handles missing stages by passing the prior through unchanged, so $\mathcal{P}_k = \mathcal{P}_{k-1}$ for any absent document.

[2] Press conferences began in April 2011 and were held quarterly until January 2019, when they became available after every meeting. The pipeline skips this stage when no transcript exists, passing $\mathcal{P}_1$ through as $\mathcal{P}_2$.

Figure 1: FOMC Communication Timeline and Filtration Structure



Note: Temporal sequence of Federal Reserve communications and the corresponding filtration stages for meeting t. On meeting day $t-1$, the Statement and Press Conference produce priors $\mathcal{P}_1$ and $\mathcal{P}_2$. The Minutes$_{t-1}$ (released ≈3 weeks later) update to $\mathcal{P}_3$. The Beige Book$_t$ (released ≈2 weeks before meeting t) produces the final prior $\mathcal{P}_4$. At meeting t, the surprise is computed as $s_t = \Delta i_t - \mathbb{E}[\Delta i_t \mid \mathcal{P}_4]$.

produces an updated posterior $\mathcal{P}_k$ with expected rate change $\mathbb{E}[\Delta i_t \mid \mathcal{P}_k]$. The *Surprise Extractor* then computes the policy innovation $\hat{s}_t = \Delta i_t - \mathbb{E}[\Delta i_t \mid \mathcal{P}_4]$ using the final posterior.[3]

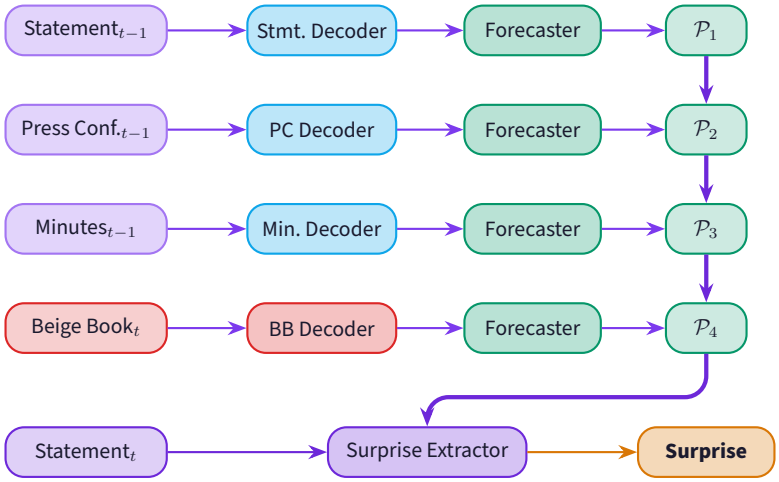
The rationale for this decomposition is both economic and computational. Each document type exhibits distinct linguistic properties requiring specialized analytical focus: Beige Book narratives demand sentiment extraction from qualitative regional reports, Minutes require parsing of internal Committee dynamics and forward guidance signals, and the Forecaster must synthesize these inputs into a probability distribution over the next rate decision. A single prompt processing all documents simultaneously must juggle multiple conflicting objectives while maintaining temporal consistency and minimizing look-ahead bias. The pipeline comprises six LLM-facing agents organized in three functional roles: four *Decoders* that extract structured information from each document type, a *Forecaster* that performs sequential probability updating, and a *Surprise Extractor* that computes the monetary policy surprise. Figure 2 illustrates the architecture.

For each meeting t, the four Decoders process the documents in the chronological order shown in Figure 2, the Forecaster updates $\mathcal{P}_1 \to \mathcal{P}_4$ at each step, and the Surprise Extractor compares $\mathbb{E}[\Delta i_t \mid \mathcal{P}_4]$ with the realized FOMC decision. Each component is detailed in turn below.

Existing textual analysis methods face a tension between scale and context when processing this documentary architecture. Supervised machine-learning approaches that map central-bank text to forecasts (Ahrens & McMahon, 2021; Ahrens et al., 2024; Aruoba & Drechsel, 2024),

[3]The hat on $\hat{s}_t$ reflects that the surprise is computed from the LLM's approximate posterior $\mathcal{P}_4$ rather than from the true conditional expectation $\mathbb{E}[\Delta i_t \mid \mathcal{B}_t]$. When the context is unambiguous, I occasionally write s_t for brevity.

Figure 2: Sequential Filtration Architecture



Note: Sequential filtration pipeline for meeting t. Four Decoder agents process documents in chronological order of public availability: Statement$_{t-1}$ and Press Conference$_{t-1}$ (released on meeting day $t-1$), Minutes$_{t-1}$ (released $\approx$3 weeks later), and Beige Book$_t$ (released $\approx$2 weeks before meeting t). Each Decoder feeds the Forecaster, which updates the prior distribution: $\mathcal{P}_1 \to \mathcal{P}_2 \to \mathcal{P}_3 \to \mathcal{P}_4$. The final prior $\mathcal{P}_4$ incorporates all pre-meeting public information. The Surprise Extractor compares $\mathbb{E}[\Delta i_t \mid \mathcal{P}_4]$ with the realized decision from Statement$_t$ to compute the monetary policy surprise.

alongside single-model LLM analyses (De Fiore et al., 2024; Gambacorta et al., 2024; Hansen & Kazinnik, 2023) improve semantic comprehension over keyword-frequency methods but typically reduce a document to a single quantitative summary, struggling to maintain coherent reasoning across the full documentary timeline while respecting the distinct analytical requirements each document type demands. The sequential architecture addresses this by assigning each document to a specialized decoder, while the Forecaster handles the synthesis task of updating beliefs as new information arrives.

Applying an LLM to historical documents raises a specific concern: extraction may reflect what the model absorbed during training rather than what the document says, in which case pre-cutoff results need not generalise. The pipeline uses DeepSeek-v3.1 (671B), whose July 2024 knowledge cutoff partitions the sample into in-training and out-of-training meetings and supplies a built-in falsification test, since a memorisation-driven pipeline would deteriorate on post-cutoff meetings. Architectural safeguards complement that comparison: strict document-level temporal cutoffs, prompt-level time anchors, sequential release ordering, and automated future-reference checks. The pipeline has been re-run on five other LLM families with distinct training cutoffs, and the results hold (Appendices B.1.1, D.2, and D.2). Section 3 formalises the leakage parameter ℓ_j and Section 2.9 reports the diagnostics.

2.2 Statement Decoder

The Statement Decoder processes the FOMC Statement released on each meeting day. The Statement is the Fed's most concise and carefully worded communication: typically 300–600 words that announce the rate decision, characterize economic conditions, and signal the likely future policy path. Because the Statement is crafted by consensus, every word change between consecutive meetings carries informational content.

The decoder extracts three categories of information. First, the *rate decision*: the target federal funds rate or range, and the change from the previous meeting. Second, *language changes*: a systematic comparison of the current Statement against the previous one, identifying additions, deletions, and modifications in key phrases, particularly those related to the economic outlook, risk assessment, and forward guidance. Third, *commitment signals*: the strength and type of forward guidance embedded in the Statement, classified along the Campbell et al. (2012) Delphic–Odyssean spectrum through linguistic proxies (outlook-based language such as "expects" and "likely" versus commitment-based language such as "until" and "at least").

These extractions feed directly into the Forecaster's first update stage, forming the initial prior $\mathcal{P}_1$ for the next meeting's decision.

2.3 Press Conference Decoder

The Press Conference Decoder analyzes the Chair's Q&A session held shortly after the Statement release. While the Statement represents the Committee's consensus language, the press conference reveals the Chair's individual interpretation, emphasis, and willingness to go beyond the prepared text. Journalists' questions often probe ambiguities in the Statement, and the Chair's responses can either reinforce or subtly diverge from the written message.

The decoder focuses on three dimensions. First, *divergence from Statement*: instances where the Chair's language is more hawkish, more dovish, or more uncertain than the Statement text, signaling the Chair's private assessment. Second, *commitment reinforcement*: whether the Chair strengthens or hedges the Statement's forward guidance when pressed by journalists. Third, *notable responses*: answers that reveal information about the Committee's reaction function, risk assessment, or internal disagreements not captured in the consensus Statement.

The decoder is calibrated to distinguish genuine signals from institutional boilerplate: during a tightening cycle, reaffirming the Committee's commitment to price stability is expected language that carries no incremental information, while explicit signals of *additional* tightening

beyond the current stance represent meaningful hawkish shifts. These signals update the prior from $\mathcal{P}_1$ to $\mathcal{P}_2$, capturing information that would be missed by analyzing the Statement alone.

2.4 Minutes Decoder

The Minutes Decoder operates on the Minutes released three weeks after meeting $t-1$, the most detailed publicly available window into Committee deliberations before meeting t and a deliberate communication tool that shapes market expectations between meetings.

The decoder extracts *policy intelligence* (internal debates, forward guidance, and risk assessments) absent from the same-day Statement. The core outputs include: (i) the distribution of hawkish and dovish views within the Committee, including the intensity of disagreement and compromise reasoning; (ii) forward guidance signals classified along the Campbell et al. (2012) taxonomy as Delphic (outlook-based, e.g. "expects", "likely") or Odyssean (commitment-based, e.g. "until", "at least"), each reported as a continuous intensity score on $[0,1]$;[4] (iii) new policy-relevant information surfaced in Minutes but not in the same-day Statement; and (iv) the Committee's own assessment of risks and uncertainties facing the economy.

FOMC Minutes follow a consistent deliberative structure that the decoder exploits (Figure 3). A segmenter groups the document into three strategic units, *Staff* (economic and financial reviews plus outlook), *Committee* (participants' discussion of conditions and outlook), and *Action* (policy decision and dissents), separating the informational backdrop from internal Committee dynamics and the decision rationale so each can be extracted with targeted prompts. Each unit is processed independently and the results are merged; when the full document fits within the model's context window, a single-pass extraction is used instead.[5]

These extractions update the prior from $\mathcal{P}_2$ to $\mathcal{P}_3$. Table 1 presents two validation tests. The internal check (column 1) regresses the magnitude of the $\mathcal{P}_1 \to \mathcal{P}_3$ revision on the prior Minutes' guidance type:[6]

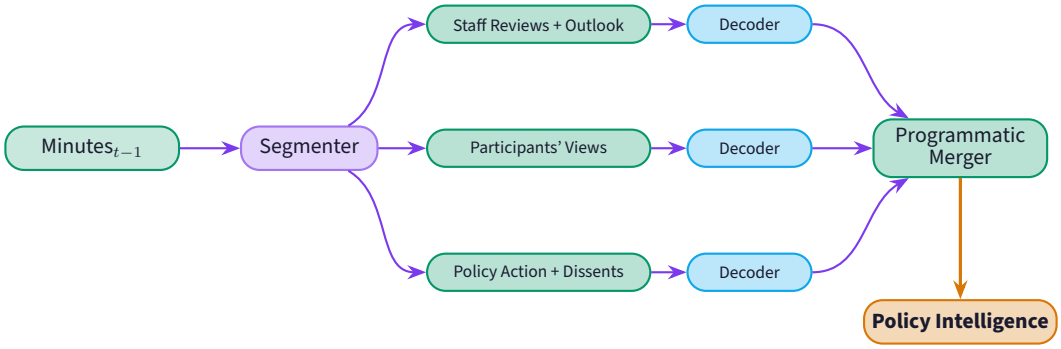
$$|\mathrm{E}[\Delta i_t \mid \mathcal{P}_3] - \mathrm{E}[\Delta i_t \mid \mathcal{P}_1]| = \alpha + \beta\, odyssean_{t-1} + \gamma\, delphic_{t-1} + \delta\, dissents_{t-1} + \phi\, \mathrm{E}[\Delta i_t \mid \mathcal{P}_1] + \varepsilon_t. \tag{1}$$

[4] Scores are independent intensities, not a composition; in the v30.5/DeepSeek-v3.1 sample they range over $[0, 0.7]$ and $[0, 0.9]$ respectively. The extraction prompt and JSON schema are included in the replication package.

[5] The segmenter handles both the modern format (post-2009, explicit section headers) and the legacy format (pre-2009, identified through narrative transition patterns).

[6] The dependent variable uses the cumulative $\mathcal{P}_1 \to \mathcal{P}_3$ rather than $\mathcal{P}_2 \to \mathcal{P}_3$ because press conferences (the $\mathcal{P}_1 \to \mathcal{P}_2$ stage) became routine only in 2011 and were quarterly until 2019, so $\mathcal{P}_2$ is undefined for roughly half the sample; restricting to meetings with a press conference gives the same sign pattern on n=81.

Figure 3: Minutes Section-Based Processing Architecture



Note: When the Minutes exceed the model's context window, the decoder exploits the document's deliberative structure via section-based processing. A segmenter splits the Minutes into five substantive sections, grouped into three strategic units: Staff Reviews (economic and financial reviews plus outlook), Participants' Views (committee discussion), and Policy Action (decision rationale and dissents). Each group is processed independently, and partial results are merged programmatically. When the full document fits within the context window, a single-pass extraction is used instead.

A negative β on odyssean guidance would indicate that stronger commitment-based language in the prior Minutes had already anchored markets, leaving less work for the Minutes-stage update; a positive γ on delphic guidance would indicate that outlook-based language, by contrast, moves expectations because it conveys data-dependent information that markets had not already priced. Estimation confirms the first reading: odyssean guidance enters at -0.014, significant at the 1% level, while delphic guidance and the dissent count are not statistically significant.

The external checks (columns 2–3) regress the realized FOMC decision Δi_t, which is not a pipeline output, on the hawkishness signal:

$$\Delta i_t = \alpha + \beta\, \textit{hawk-dove}_{t-1} + \gamma\, \textit{debate}_{t-1} + \delta\, \textit{dissents}_{t-1} + \phi\, \mathrm{E}[\Delta i_t \mid \mathcal{P}_1] + \varepsilon_t, \tag{2}$$

with column (2) estimating a parsimonious version that drops $\textit{debate}_{t-1}$ and $\mathrm{E}[\Delta i_t \mid \mathcal{P}_1]$ and column (3) estimating the full specification. The parsimonious specification (column 2: hawk-dove and dissents) explains 36% of the variation in realized decisions, with hawk-dove balance loading at $+0.293$, significant at the 1% level. Conditioning on $\mathrm{E}[\Delta i_t \mid \mathcal{P}_1]$ in column (3) shows that hawk-dove balance still carries incremental predictive content beyond what the prior Statement already signalled (coefficient $+0.074$, significant at the 1% level), and the prior loads at $+0.814$ at the 1% level. Debate intensity enters negatively at the 10% level (-0.199) in this specification; the dissent count remains insignificant. Together, the two specifications give a coherent picture: the *type* of guidance predicts how much the prior moves (column 1), and the *direction* of hawkishness predicts the realised decision both unconditionally and incrementally

Table 1: Minutes Decoder Validation: Internal Consistency and Realized Policy Decision

		External Validation	
	$\lvert\Delta\mathcal{P}_1 \to \mathcal{P}_3\rvert$	Δi_t	Δi_t
	(1)	(2)	(3)
Debate intensity$_{t-1}$			-0.199^{*}
			(0.108)
Hawk-dove balance$_{t-1}$		0.293^{***}	0.074^{***}
		(0.058)	(0.027)
Odyssean guidance$_{t-1}$	-0.014^{***}		
	(0.005)		
Delphic guidance$_{t-1}$	-0.006		
	(0.009)		
Num. dissents$_{t-1}$	0.000	-0.010	0.025
	(0.001)	(0.013)	(0.017)
$\mathrm{E}[\Delta i_t \mid \mathcal{P}_1]$	0.008		0.814^{***}
	(0.010)		(0.113)
Constant	0.010^{***}	0.040^{**}	0.041^{*}
	(0.004)	(0.018)	(0.021)
N	206	230	208
R^2	0.084	0.361	0.611
Adj. R^2	0.065	0.356	0.604

Note: (1) DV: $\lvert\mathrm{E}[\Delta i_t \mid \mathcal{P}_3] - \mathrm{E}[\Delta i_t \mid \mathcal{P}_1]\rvert$ (bp). Newey-West HAC (4 lags).
(2) DV: realized rate change Δi_t (bp). Newey-West HAC (4 lags).
(3) Adds $\mathrm{E}[\Delta i_t \mid \mathcal{P}_1]$ to isolate Minutes content beyond the prior Statement signal. Newey-West HAC (4 lags).

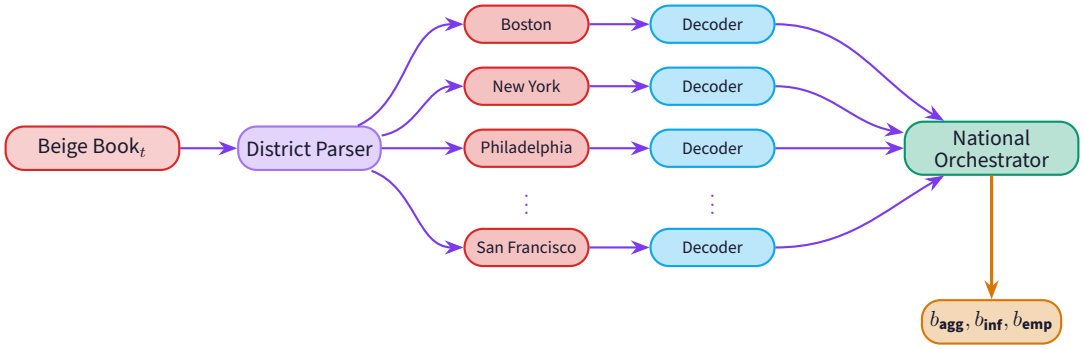
to the prior Statement (columns 2–3).

2.5 Beige Book Decoder

The Beige Book Decoder operates on the Beige Book released approximately two weeks before each FOMC meeting, converting qualitative economic narratives from the twelve Federal Reserve districts into structured, policy-relevant signals that feed the final prior update $\mathcal{P}_3 \to \mathcal{P}_4$.

The decoder exploits the Beige Book's geographic structure (Figure 4). A first-tier parser splits the raw document along its twelve Federal Reserve district sections, and each section is processed independently and in parallel by a district-level decoder that produces a structured intelligence report: dual mandate assessments for inflation and employment, any auxiliary topics the district emphasizes, a communication analysis of tone and framing, and identified causal mechanisms linking local conditions to policy implications. A second-tier national orchestrator then synthesizes the twelve district reports into aggregate national scores, weighting districts by approximate GDP shares and by confidence levels. The orchestrator does not simply average: it identifies geographic patterns, flags bellwether districts (Cleveland and Chicago for manufacturing, New York for financial conditions, Dallas for energy), and notes disagreements across regions. This design exploits the document's inherent regional structure rather than resorting to

Figure 4: Beige Book Regional Processing Architecture



Note: The Beige Book Decoder exploits the document's geographic structure. A parser splits the raw text into twelve Federal Reserve district sections. Each district is processed independently in parallel, producing district-level dual mandate and auxiliary topic scores. The National Orchestrator synthesizes district assessments into aggregate scores (b_{agg}, b_{inf}, b_{emp}) using Equation 3.

arbitrary text chunking, ensuring that each district's narrative is analyzed in full context. Appendix B.3.2 examines whether this geographic decomposition carries independent identification content beyond the national aggregate.

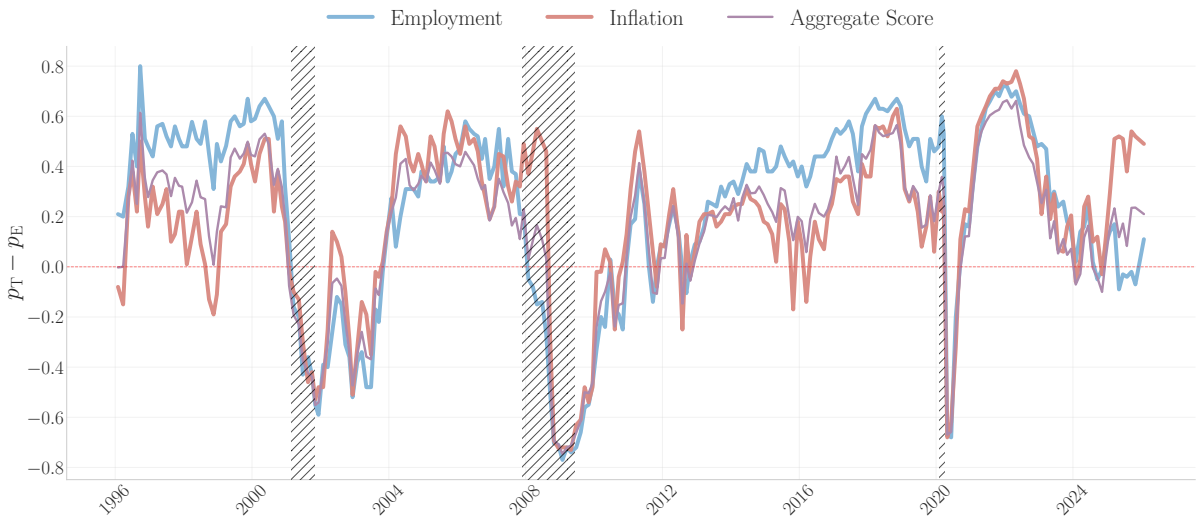
Scoring follows a *narrative-first* protocol. District decoders write a qualitative assessment of each economic signal and then assign a policy probability simplex $\boldsymbol{\pi} = (p_{\text{tighten}}, p_{\text{neutral}}, p_{\text{ease}})$, reflecting the probability that the signal, considered in isolation, would push the FOMC toward tightening, holding, or easing. The net policy bias is then computed deterministically as $b = p_{\text{tighten}} - p_{\text{ease}} \in [-1, 1]$, with the sign indicating direction and the magnitude reflecting strength. This factorization separates qualitative reasoning from quantification, preventing the false precision that arises when LLMs are asked to output continuous scores directly.

The decoder produces two score categories. The *dual mandate* scores target the Fed's statutory objectives: an inflation score $b_{\text{inf}} \in [-1, 1]$ and an employment score $b_{\text{emp}} \in [-1, 1]$, with positive values indicating hawkish-leaning conditions. *Auxiliary topics* discovered from the text (consumer demand, housing, manufacturing, credit conditions, energy, and others) are scored on the same scale; the topic set varies across meetings. The orchestrator also produces a geographic synthesis, an integrated causal narrative, a risk balance, and conditional triggers ("if labour softens materially, bias shifts to hold"), all of which enter the Forecaster's context.

These components combine into an aggregate sentiment score

$$b_{\text{agg}} = w_{\text{inf}} \cdot b_{\text{inf}} + w_{\text{emp}} \cdot b_{\text{emp}} + \sum_{j=1}^{J} w_j \cdot b_j, \tag{3}$$

Figure 5: Beige Book Dual Mandate Scores



Note: Inflation score (red), employment score (blue), and weighted aggregate score (black, computed via Equation 3) extracted from Beige Book text. Scores are computed as $p_{\text{tighten}} - p_{\text{ease}}$ from district-level policy probability simplexes, yielding continuous values in $[-1, 1]$ where positive values indicate hawkish-leaning conditions. The aggregate additionally incorporates auxiliary topics (Figure 27). Recession periods are shaded in gray.

where $\{b_j\}_{j=1}^{J}$ are the auxiliary topic scores and the weights satisfy $w_{\text{inf}} + w_{\text{emp}} + \sum_j w_j = 1$. The weights are not fixed: the decoder assigns them at each meeting based on the relative policy relevance each dimension receives in the Beige Book discussion, with an explicit justification for the allocation. Inflation and employment jointly carry 85–95% of the weight throughout the sample, but the inflation share rises sharply during 2021–2022 and falls during the post-2008 disinflationary period; Appendix B.3.1 reports the full weight dynamics and the auxiliary topic scores.

Figure 5 shows that the dual mandate scores track business cycle dynamics. Employment drops sharply during the 2008–2009 recession and remains depressed through 2010, while inflation stays near zero through the late 1990s and early 2000s before rising strongly during the 2022 inflationary episode. The aggregate score declines during recessions and rises during expansions, with the gap to the individual mandate scores reflecting the auxiliary topics' contribution.

Appendix B.3 examines the implications of the regional architecture. Geographic disagreement across districts, FOMC voting rotation, and heterogeneous district slopes introduce no exploitable bias (Sections B.3.2–B.3.4), validating the GDP-weighted aggregation. The mandate weights themselves, however, are state-dependent: they shift systematically with the macro regime (Section B.3.5).

Table 2: Beige Book Information Content and Policy Momentum

	(1)	(2)	(3)	(4)	(5)
Panel A: Level Inertia					
i_{t-1}	-0.008	-0.019***	-0.020***	-0.015***	-0.018***
	(0.006)	(0.006)	(0.006)	(0.006)	(0.006)
$\text{BB}^{\text{agg.}}$	—	0.275***	—	—	—
		(0.037)			
$\text{BB}^{\text{empl.}}$	—	—	0.241***	—	0.176**
			(0.033)		(0.087)
$\text{BB}^{\text{infl.}}$	—	—	—	0.261***	0.388***
				(0.038)	(0.099)
R^2	0.007	0.175	0.168	0.154	0.180
Adj. R^2	0.003	0.169	0.162	0.147	0.171
Obs.	272	272	272	272	272
Panel B: Change Inertia					
Δi_{t-1}	0.609***	0.545***	0.550***	0.551***	0.541***
	(0.048)	(0.052)	(0.052)	(0.051)	(0.052)
$\text{BB}^{\text{agg.}}$	—	0.102***	—	—	—
		(0.034)			
$\text{BB}^{\text{empl.}}$	—	—	0.086***	—	0.019
			(0.030)		(0.074)
$\text{BB}^{\text{infl.}}$	—	—	—	0.101***	0.205**
				(0.034)	(0.087)
R^2	0.372	0.393	0.391	0.392	0.397
Adj. R^2	0.370	0.388	0.386	0.387	0.390
Obs.	271	271	271	271	271

Note: This table shows the progression from policy inertia to the full dual-mandate specification. Panel A uses the lagged rate level (i_{t-1}); Panel B uses the lagged rate change (Δi_{t-1}). Column (1): inertia only. Column (2): weighted Beige Book aggregate. Columns (3) and (4): individual mandate components. Column (5): both mandate variables. Standard errors in parentheses. ***, **, and * denote significance at 1%, 5%, and 10% levels. Time window: 1996-01 to 2026-01.

I now assess whether the decoder produces economically sensible scores by estimating

$$\Delta i_t = \alpha + \rho\, x_{t-1} + \gamma\, d_{\text{BB},t} + \boldsymbol{\beta}'\mathbf{b}_t + \varepsilon_t \tag{4}$$

where x_{t-1} is either the lagged rate change Δi_{t-1} or the lagged rate level i_{t-1}, $d_{\text{BB},t}$ is an indicator equal to one when a dedicated Beige Book is available for meeting t,[7] and $\mathbf{b}_t$ expands progressively from the weighted aggregate BB_t to the mandate components ($\text{BB}_t^{\text{infl}}, \text{BB}_t^{\text{empl}}$) and their interaction. Beige Book scores are available approximately two weeks before each meeting, so this is a predictive regression.

Panel A conditions on the rate level. Without inertia to absorb persistence, the aggregate

[7]When the FOMC sets a target range, I use the midpoint for calculations. Approximately 30 meetings (inter-meeting actions and emergency decisions) lack a dedicated Beige Book. To avoid conflating "no Beige Book" with "neutral economic conditions," scores are demeaned within the dedicated-BB sample before zero-filling non-BB meetings; $d_{\text{BB},t} = 0$ then captures departure from average BB conditions rather than a neutral economy.

alone explains 19.0% of variance, and a suppressor effect emerges: the rate level is insignificant unconditionally but enters negatively once Beige Book scores are included, because the scores absorb the cyclical component and reveal the underlying tendency for the Fed to ease from elevated levels.

Panel B conditions on the lagged rate change, capturing the empirical persistence of rate decisions (Rudebusch, 2002). Momentum alone explains 37.2% of variance; the Beige Book aggregate remains highly significant after absorbing this control. Entering the two mandate components separately, neither is individually significant because the subindices are highly collinear ($\text{corr}(b_{\text{inf}}, b_{\text{emp}}) = 0.79$, with 88% of meetings on the NE–SW diagonal; see Appendix B.2.4). Column (5) re-adds them jointly: because the terms enter additively, the signal compounds when inflation and employment agree and partially cancels when they conflict, with inflation dominating in Panel B and the two carrying roughly equal weight in Panel A.

2.6 Forecaster

The Forecaster synthesizes Decoder outputs through four sequential probability updates that mirror, but do not literally implement, Bayesian updating (see Section 3 for the as-if framing). At stage $k = 1$, it forms an initial distribution $\mathcal{P}_1$ from the Statement Decoder. At stages $k \in \{2, 3, 4\}$, it takes the previous posterior $\mathcal{P}_{k-1}$, the structured output of the stage-k Decoder, and a *regime narrative* summarizing the recent policy cycle (tightening, easing, or holding), and returns an updated posterior $\mathcal{P}_k$. The previous meeting's realized outcome and a narrative interpretation generated by the Surprise Extractor (Section 2.8) are passed into $\mathcal{P}_1$ as cross-meeting context, giving the pipeline memory across meetings.

Each $\mathcal{P}_k$ is a probability mass function over discretized changes in the federal funds rate target, covering the direction and magnitude of the next decision.[8] Prompting is *narrative-first*: before assigning probabilities, the model must explain how the new document should shift beliefs relative to $\mathcal{P}_{k-1}$ in light of the prevailing regime. This ordering reduces anchoring on the inherited numeric prior and ensures that each update reflects the incremental information in the new document.

Figure 6 illustrates the filtration on two meetings spanning the LLM training horizon. Panel A shows the June 2022 FOMC, in the early stages of the post-pandemic tightening cy-

[8]The scope is restricted to conventional policy: the distribution excludes balance sheet operations, forward guidance about the future path, and unconventional tools.

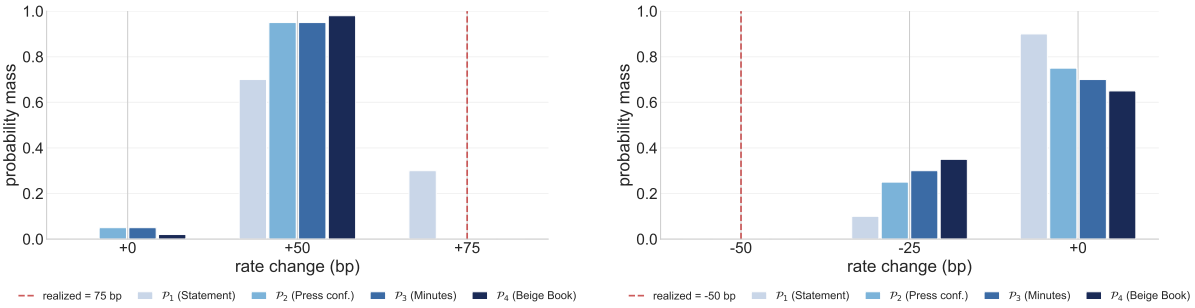
cle and squarely within DeepSeek-V3.1's training window. The Statement Decoder reads the May 2022 communications as committing the Committee to half-point hikes, producing a prior split between +50 and +75 bp ($\mathcal{P}_1$: $E = +57.5$ bp). The Press Conference Decoder then registers Chair Powell's explicit signal that "a 75 basis point increase is not something the Committee is actively considering," and the +75 mass collapses to near zero ($\mathcal{P}_2$: $E = +47.5$ bp). The Minutes confirm a Committee aligned around +50 ($\mathcal{P}_3$: $E = +47.5$ bp); the Beige Book leaves the central tendency essentially unchanged ($\mathcal{P}_4$: $E = +49.0$ bp), against a realized 75 bp hike. The residual surprise is +26 bp, the largest hawkish surprise of the cycle: a documented commitment was overturned in the two days before the meeting following a Wall Street Journal report indicating that policymakers would consider a larger move, an information channel that lies outside the four official documents the pipeline reads.

Panel B shows the September 2024 FOMC, post-cutoff and therefore out of sample for the LLM. The Statement Decoder reads a baseline that still leans toward a hold ($\mathcal{P}_1$: $E = -2.5$ bp); each subsequent document shifts probability mass monotonically toward a 25 bp cut, with the Press Conference, Minutes, and Beige Book each adding roughly ten percentage points to the easing bin ($\mathcal{P}_2$: $E = -6.25$ bp; $\mathcal{P}_3$: $E = -7.5$ bp; $\mathcal{P}_4$: $E = -8.75$ bp), against a realized 50 bp cut. The realization falls outside the prior's support, and the residual surprise is -41.25 bp. The two examples make the same point in two regimes: each document moves the distribution in the direction of the realization, and what remains at $\mathcal{P}_4$ is the part of the decision the documents do not foreshadow. The out-of-sample case is hard to reconcile with a pure-memorization explanation: the LLM cannot have seen the September 2024 outcome at training time, yet the filtration walks toward it stage by stage rather than jumping to it.

Across the full sample, the average marginal update is largest at the Beige Book stage, reflecting fresh real-economy information not present in the policy-focused earlier documents, and smallest at the Press Conference stage; cumulative updates grow with the span, with $\mathcal{P}_1 \to \mathcal{P}_4$ capturing total pre-meeting information content. Appendix C.1 reports the full distribution of update magnitudes across all stage pairs.

Box 1 traces verbatim LLM outputs at the Beige Book stage of the July 2022 FOMC: the Decoder's synthesis enters the Forecaster's context window unedited, alongside $\mathcal{P}_3$ and the regime description.

Figure 6: Sequential Updating Across the LLM Training Horizon



A. In-sample: June 2022 (+75 bp).

B. Out-of-sample: September 2024 (−50 bp).

Note: Each strip shows the probability mass function over rate-change outcomes at one filtration stage, from $\mathcal{P}_1$ (Statement) to $\mathcal{P}_4$ (Beige Book). Colors indicate policy direction (red: cut, grey: hold, blue: hike); the dashed line marks the realized decision. Panel A is in-sample for DeepSeek-V3.1; Panel B is post-cutoff. Residual surprises at $\mathcal{P}_4$ are +26 bp (a WSJ leak two days before the meeting overturned a documented commitment, see text) and −41.25 bp (the realized 50 bp cut lies outside the prior's support).

2.7 When the Fed Doesn't Speak: News in the Pre-FOMC Blackout

Once the Beige Book is released, the Fed enters its pre-FOMC media blackout: no new Statement, press conference, Minutes, or Beige Book reaches the public until the meeting. The window lasts roughly 10–14 days. The baseline pipeline stops at $\mathcal{P}_4$ at this point. An optional fifth stage, $\mathcal{P}_5$, fills the blackout by expanding the information set to include inter-meeting news arriving in the same window, drawing on the financial press to capture what the Fed itself has stopped saying.

News articles are sourced from FactSet StreetAccount (see Section 4) and filtered for macro-relevant content. Five specialized extractors process the articles by domain in parallel: *data releases* (hard economic indicators with actual, consensus, prior, and surprise direction); *Fed communications* (speeches and testimony by FOMC members, with speaker, voting status, key message, and deviation from the most recent Statement); *macroeconomic outlook* (press narratives on growth, inflation, and labor dynamics, tagged by direction and confidence); *financial conditions* (credit spreads, bank stress, energy shocks, and geopolitical disruptions, rated by systemic severity); and *market expectations* (implied probabilities from futures markets and named-source forecasts from dealer banks). Across post-2004 meetings, the pipeline extracts a median of 20 data-release records per window, and 76 meetings include verbatim quotes from FOMC officials.

A synthesizer consolidates the five domain reports into a single coherent assessment, identifying the dominant signal, conflicts across domains, and the net policy implication. The

July 27, 2022 FOMC: Beige Book stage ($\mathcal{P}_3 \to \mathcal{P}_4$; realized +75 bp)

PRIOR ENTERING STAGE ($\mathcal{P}_3$, after Minutes): {+50 bp : 40%, +75 bp : 60%} $E[\Delta i_t \mid \mathcal{P}_3] = +65.0$ bp

Beige Book Decoder : verbatim output:

> `The national economic picture reveals persistent inflationary pressures coexisting with emerging demand destruction in interest-sensitive sectors. [...] Inflation remains broad-based with particular strength in services, energy, and food, driven by ongoing cost pressures and wage growth well above historical averages. [...] The clear transmission of monetary policy is evident through housing market cooling and manufacturing softening, yet these demand-side adjustments have not yet translated into meaningful labor market rebalancing or core services inflation moderation.`

Forecaster ($\mathcal{P}_3 \to \mathcal{P}_4$) : verbatim reasoning:

> `The Beige Book aligns with` $\mathcal{P}_3$`'s hawkish stance: inflation pressure (72% tighten) and labor tightness (68% tighten) remain high, supporting aggressive hikes. However, stronger-than-expected secondary softening in housing (66% ease) and consumer demand (55% ease) slightly reduces the probability of a 75 bp hike in favor of 50 bp, as these effects may temper urgency without undermining the primary mandate focus.`

UPDATED DISTRIBUTION ($\mathcal{P}_4$): {+50 bp : 55%, +75 bp : 45%} $E[\Delta i_t \mid \mathcal{P}_4] = +61.25$ bp
REALIZED: +75 bp (✓ direction correct; $|s_t| = 13.75$ bp genuine surprise)

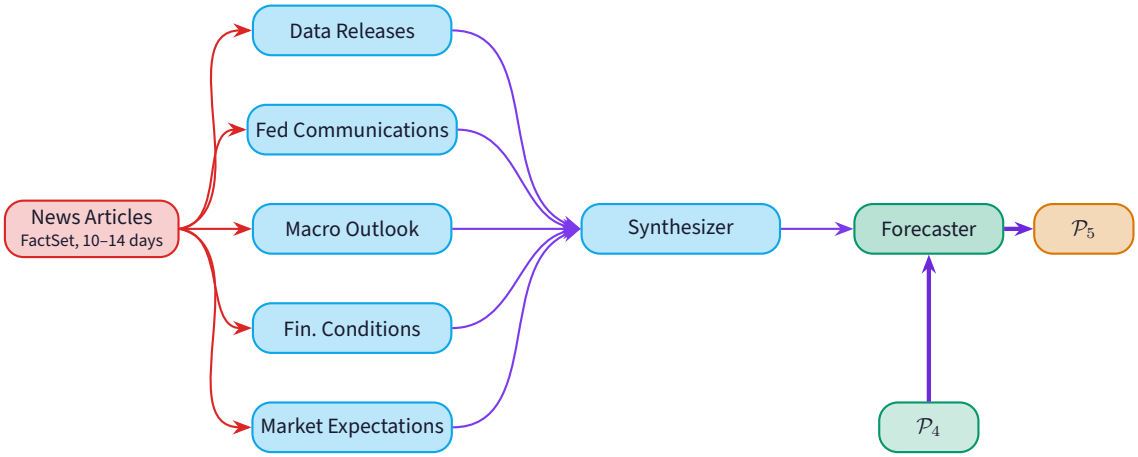
Box 1: Verbatim LLM outputs for the Beige Book stage ($\mathcal{P}_3 \to \mathcal{P}_4$), July 2022 FOMC. The Beige Book Decoder's synthesis (first shaded block) enters the Forecaster's context window verbatim; the Forecaster produces a natural-language justification (second shaded block) before committing to probabilities. Both outputs are reproduced from the database (v30.1, `deepseek-v3.1:671b`); ellipses indicate excerpts across sibling narrative fields, with minor punctuation edits for readability.

synthesizer weights channels rather than averaging them, so a single high-severity signal can override consensus across the other four. The Forecaster then updates $\mathcal{P}_4 \to \mathcal{P}_5$ using the same narrative-first protocol as the earlier stages. When $\mathcal{P}_5$ is enabled, the Surprise Extractor uses it as the reference distribution in place of $\mathcal{P}_4$.

The March 2020 emergency cut illustrates the value of channel weighting. Strong pre-virus data (NFP +225K above consensus, ISM manufacturing in expansion) were classified by the synthesizer as backward-looking and overridden by the financial conditions channel: coronavirus systemic severity, Treasury yields at record lows, oil prices down 16%. The Forecaster delivered $\mathcal{P}_5$ with $\mathbb{E}[\Delta i_t] = -49$ bp against the realized -50 bp emergency cut. A pure averaging of the five channels would have anchored the prior near zero; the synthesizer's editorial layer is what allows $\mathcal{P}_5$ to follow the regime break.

Figure 8 traces the $\mathcal{P}_1 \to \mathcal{P}_5$ evolution for six episodes selected to span the range of news-

Figure 7: Pre-FOMC Blackout: News-Update Architecture



Note: Architecture of the news-stage update. News articles are processed by five domain extractors in parallel (blue nodes). A synthesizer consolidates the five reports into a single coherent assessment. The Forecaster combines this synthesis with the document-only prior $\mathcal{P}_4$ (entering from below) and produces $\mathcal{P}_5$ using the same narrative-first protocol as stages $\mathcal{P}_1$–$\mathcal{P}_4$.

stage behavior: large corrections where $\mathcal{P}_5$ closes most of the gap to a realized policy break, sharpening of an already near-correct $\mathcal{P}_4$ onto a near-certain outcome, partial updates that leave residual uncertainty, and one case (December 2008) in which the realized decision falls outside the prior's support entirely.

Across post-2004 meetings, the news stage cuts mean absolute forecast error roughly in half, with the largest gains concentrated in the heaviest revisions and no movement at all when the synthesizer judges the inter-meeting signal too weak to act on. The news stage therefore concentrates its work where it is needed and stays out of the way otherwise. Appendix C.2 reports the full accuracy-by-update-magnitude breakdown alongside calibration, serial-correlation, and predictability diagnostics that compare $\mathcal{P}_4$- and $\mathcal{P}_5$-based surprises as identified shocks.

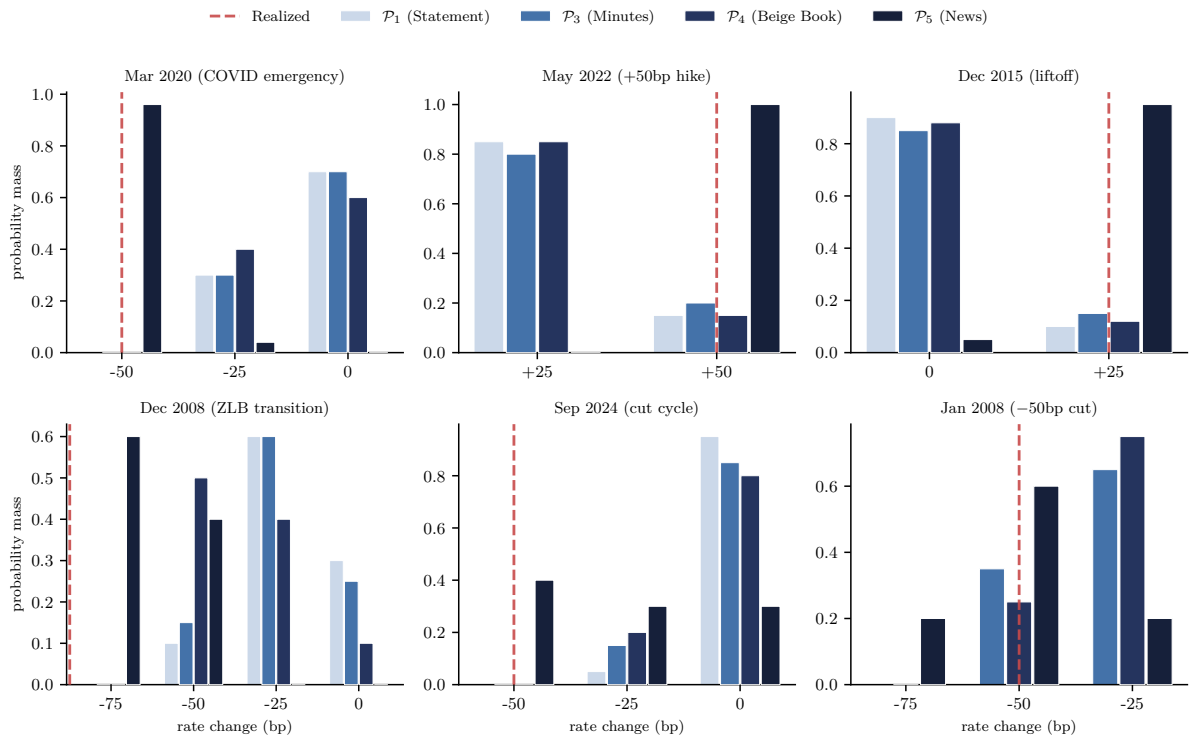
2.8 Surprise Extractor

The Surprise Extractor quantifies the monetary policy surprise on FOMC announcement day. The agent receives the Forecaster's final prior $\mathcal{P}_4$ and expected rate change, extracts the realized policy decision from the FOMC Statement text for meeting t, and computes the deviation:

$$s_t = \Delta i_t^{\text{realized}} - \mathbb{E}[\Delta i_t \mid \mathcal{P}_4]$$

where $\Delta i_t^{\text{realized}}$ is the announced rate change and $\mathbb{E}[\Delta i_t \mid \mathcal{P}_4] = \sum_i d_i \, p_i$ is the Forecaster's probability-weighted expectation, with d_i indexing the support of rate-decision outcomes (e.g.,

Figure 8: Filtration Evolution: Key FOMC Episodes



Note: Probability distributions over the rate decision at each filtration stage for six key FOMC meetings. Overlapping bars show $\mathcal{P}_1$ (statement), $\mathcal{P}_3$ (minutes), $\mathcal{P}_4$ (Beige Book), and $\mathcal{P}_5$ (inter-meeting news), with increasing opacity reflecting later stages. Dashed vertical lines mark the realized decision. Episodes span the range of news-stage contributions: large corrections (Mar 2020, Jan 2008), sharpening of near-certain outcomes (Dec 2015, May 2022), partial updates where residual uncertainty remains (Sep 2024), and a case where the decision lay beyond the prior support entirely (Dec 2008 ZLB transition).

± 25 bp, ± 50 bp, no change) and p_i the corresponding $\mathcal{P}_4$ posterior probability.

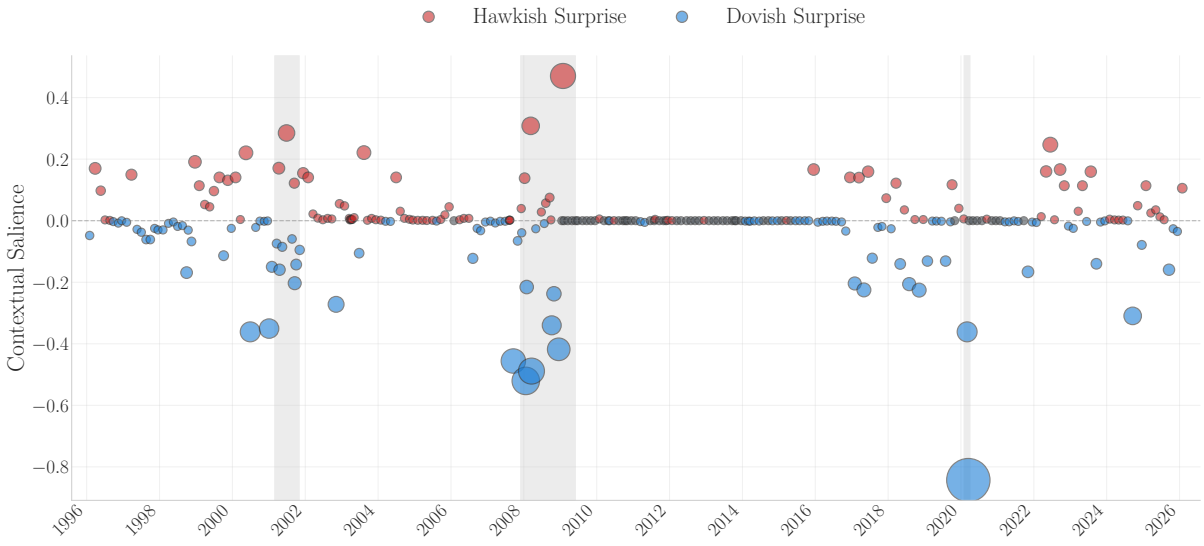
Beyond s_t, the agent produces two ancillary outputs. A narrative summary of the decision is passed forward to the next meeting's Decoders and Forecaster as cross-meeting context, helping the pipeline track evolving policy dynamics. A salience score $\mu_t \in [0, 1]$ rates the announcement's novelty, but is largely mechanical: the prior's probability mass at the realised decision, $p(\text{actual} \mid \mathcal{P}_4)$, alone explains 77% of its variation (Appendix C.1.4). Identification therefore relies on s_t throughout, in keeping with market-based and regression-based alternatives.

2.9 Validation

The pipeline is validated along three pillars: stability under repeated execution, robustness to look-ahead bias, and coherence with external benchmarks. Each is summarised here; full diagnostics are in Appendix B.

Stability. Six independent executions of the pipeline at temperature zero produce essentially

Figure 9: Monetary Policy Surprises



Note: Monetary policy surprises $s_t = \Delta i_t^{\text{realized}} - \mathbb{E}[\Delta i_t \mid \mathcal{P}_4]$ over the sample period. Positive values indicate hawkish surprises (tighter than expected), negative values indicate dovish surprises (looser than expected). The largest surprises cluster during regime transitions (2008, 2015 liftoff, 2020 pandemic, 2022 tightening initiation).

the same series. Median cross-run dispersion is roughly 2 bp for both expected rate changes and surprises, an order of magnitude smaller than the 25 bp typical policy increment. Elevated dispersion is concentrated at genuine regime breaks (the 2001 recession, the 2008–09 ZLB transition with a peak in January 2009, the March 2020 pandemic, and the 2021 inflation-regime pivot), where Fed communication itself was more ambiguous (Appendix B.1.1).

Look-ahead bias. If the model were retrieving memorised outcomes rather than reasoning from documents, meetings beyond its training cutoff would show *higher* cross-run dispersion. Across five model families with cutoffs ranging from July 2024 to January 2025 (DeepSeek-v3.1, Gemma4, Qwen3.6, GPT-4.1-mini, GPT-5-mini), in-sample and out-of-sample distributions overlap and none of the five t-tests is significant. A second check follows from the filtration design: stage-by-stage updates $\mathcal{P}_1 \to \mathcal{P}_4$ are large and document-driven, a behaviour memorisation alone cannot produce, since the realised outcome never enters the filtration prompts and appears only at the Surprise Extractor. Architectural controls (document-level temporal cutoffs, prompt-level date anchors, automated future-reference checks) reinforce the empirical null (Appendix B.1.1).

External benchmark. The framework agrees with professional forecasters on both the central forecast and the level of policy uncertainty. The LLM's rate expectation correlates with the market-implied expectation at 0.78 across the full sample, rising to 0.84 in the post-2019

every-meeting press-conference regime. Cross-forecaster disagreement on the three-month rate, a survey-based uncertainty proxy constructed independently of the LLM, co-moves with the magnitude of the narrative surprise at a correlation of 0.43, significant at the 0.1% level: meetings that forecasters found hard to call also produce larger LLM surprises (Appendix B.2).

The critical question is whether these measures improve upon existing identification strategies. I now examine the framework's empirical performance: first validating individual agent outputs, then assessing the resulting surprise measures through measurement quality diagnostics, impulse response analysis, and out-of-sample trading performance.

3 Signal Extraction Framework

An *as-if* Bayesian signal extraction framework organises the empirical analysis. It defines the object produced by the pipeline, maps each component to a specific empirical test, and characterizes the conditions under which extraction quality degrades. This section presents the key ideas; Appendix E provides the full model and derivations.

3.1 Information Sets and the Measured Surprise

The framework reads the LLM's output through an *as-if* Bayesian lens: I write down what a Bayesian aggregator would produce on the same extracted signals and use it as the interpretive benchmark, and the calibration test in Section 3.2 then asks whether the pipeline's outputs are *observationally equivalent* to that benchmark, a behavioural property of the outputs rather than a claim about the model's internal computation. Let t index FOMC meetings and define three nested information sets just before the policy decision:

$$\mathcal{B}_t \equiv \sigma(\text{public Fed documents processed by the pipeline before } t), \tag{5}$$

$$\mathcal{M}_t \equiv \sigma(\text{broader public information available by } t^-), \tag{6}$$

$$\mathcal{G}_t \equiv \sigma(\text{policymakers' internal assessments and deliberations by } t^-), \tag{7}$$

with $\mathcal{B}_t \subseteq \mathcal{M}_t \subseteq \mathcal{G}_t$. Let $\theta_t \equiv \mathbb{E}[\Delta i_t \mid \mathcal{G}_t]$ denote the expected rate decision conditional on the Fed's internal information. The realized rate change adds an unpredictable innovation,

$$\Delta i_t = \theta_t + u_t, \qquad u_t \perp \mathcal{G}_t. \tag{8}$$

Each document j provides a noisy signal $d_{jt} = \theta_t + \varepsilon_{jt}$ with precision $\tau_j(R_t)$ that increases with the transparency regime R_t. The pipeline does not pass raw documents to the Forecaster. Specialized Decoder agents first extract structured summaries, and the Forecaster updates beliefs from these summaries alone. This two-stage architecture acts as an information bottleneck against contamination: the LLM's training data may encode prior knowledge c_{jt} about meeting t (memorised outcomes, post-hoc narratives) that bypasses the documentary record. The Decoder's structured-output protocol attenuates this channel, but a fraction $\ell_j \in [0, 1]$ may still propagate through:

$$S_{jt} = \theta_t + \varepsilon_{jt} + \ell_j \, c_{jt} + \varepsilon_{jt}^{dec}. \tag{9}$$

The Forecaster adds its own noise ν_{jt}^{for}, giving an effective precision for each document stage:

$$\tilde{\tau}_j = \left(\tau_j^{-1} + (\kappa_j^{dec})^{-1} + (\kappa_j^{for})^{-1} \right)^{-1}. \tag{10}$$

Extraction quality depends on document informativeness (τ_j), Decoder capability (κ_j^{dec}), and Forecaster capability (κ_j^{for}). The leakage term $\ell_j c_{jt}$ is treated as a non-classical channel: it is not a Gaussian noise component and would inflate variance only if c_{jt} varies meaningfully across meetings, but it can introduce mean-bias when c_{jt} correlates with θ_t. After processing up to four documents, the LLM posterior $\mathcal{P}_4$ yields the measured surprise:

$$\hat{s}_t = \Delta i_t - m_{4t}, \tag{11}$$

where m_{4t} is the posterior mean. Adding and subtracting the true conditional means given $\mathcal{M}_t$ and $\mathcal{B}_t$ decomposes the surprise into four terms:

$$\begin{aligned} \hat{s}_t = \underbrace{u_t}_{\text{policy innovation}} &+ \underbrace{(\mathbb{E}[\theta_t \mid \mathcal{G}_t] - \mathbb{E}[\theta_t \mid \mathcal{M}_t])}_{\xi_t^{priv}\text{: private-information wedge}} \\ &+ \underbrace{(\mathbb{E}[\theta_t \mid \mathcal{M}_t] - \mathbb{E}[\theta_t \mid \mathcal{B}_t])}_{\xi_t^{pub}\text{: public non-document wedge}} + \underbrace{(\mathbb{E}[\theta_t \mid \mathcal{B}_t] - m_{4t})}_{\eta_t\text{: extraction error}}. \end{aligned} \tag{12}$$

Each term maps to a specific empirical margin. The conditioning set $\mathcal{B}_t$ is explicit and dated, so the wedges are interpretable rather than hidden inside announcement-window prices or *ex post* cleaning regressions. The extraction-error term η_t is a catch-all for everything the LLM gets wrong about its own conditioning set, including classical Bayesian-update noise and any leakage

bias $\sum_j w_{jt}\ell_j c_{jt}$; Appendix E unpacks η_t into these two components explicitly.

3.2 Testable Implications

The framework generates two implications that the empirical sections test directly, plus a sequential-learning property used to organise the appendix.[9]

The first implication concerns forecast efficiency. If the pipeline aggregates signals as a Bayesian would on the extracted summaries, with negligible leakage ($\ell_j \approx 0$), two complementary regressions of the realised rate change should each yield a unit slope: regressing Δi_t on the prior mean $\mathbb{E}[\Delta i_t \mid \mathcal{P}_4]$ tests whether the prior tracks the truth on average, and regressing Δi_t on the surprise $\hat{s}_t$ tests whether the surprise carries the rate-change information one-for-one. Table 3 reports a slope of 1.000 in the expectation regression and Table 6 reports 1.005 in the surprise regression, both statistically indistinguishable from unity. The estimates are consistent with the joint hypothesis of small leakage and observational equivalence to a Bayesian, though they do not separately identify either condition.

The second implication concerns transparency and degradation. Surprise variance decreases when transparency raises document precision τ_j or when better LLMs raise extraction precision κ_j^{dec}, κ_j^{for}. The measure degrades when documents are unavailable, uninformative, poorly extracted, or when contamination bypasses the bottleneck. The Consensus Economics validation (Appendix B.2.3) confirms the transparency margin: the correlation between the LLM prior and the market-implied expectation rises from $r = 0.38$ in the 2011–2018 partial-transparency regime (mechanically compressed by the zero lower bound) to $r = 0.84$ in the post-2019 every-meeting press-conference regime. The contamination margin is bounded by the look-ahead diagnostics summarised in Section 2.9, which find no evidence of higher cross-run dispersion outside the training window.

The remaining margins of the decomposition map onto specific results developed later: predictability of $\hat{s}_t$ from non-document public variables ($R^2 = 0.166$, falling to 2.9% with the $\mathcal{P}_5$ news extension) bears on the public non-document wedge ξ_t^{pub}; the 11.2-percentage-point Greenbook increment constrains the precision of the Fed's private signal ξ_t^{priv}. By construction $\hat{s}_t$ is a documentary-conditioned monetary policy shock (the unanticipated component of the rate decision given $\mathcal{B}_t$); the LP and IV estimands recover the response to that shock as it is

[9] Under approximately Gaussian-Bayesian updating, positive effective precision at each stage ($\tilde{\tau}_j > 0$) implies monotone entropy decline from $\mathcal{P}_1$ to $\mathcal{P}_4$; the distributional analysis in Appendix C.1.2 confirms this in the data.

delivered in the meeting-day announcement bundle, with the same scope as Gertler and Karadi (2015) and broader than the ex-post-cleaned object targeted by Bauer and Swanson (2023b) and Miranda-Agrippino and Ricco (2021). Detail appears with the impulse responses in Section 6.

4 Data

The primary inputs to the pipeline are FOMC Statements, press conference transcripts, Minutes, and Beige Books, all sourced from the Federal Reserve's public website. The sample spans 272 FOMC meetings from January 1996 through March 2026. Statement and Minutes coverage begins in 2000; earlier meetings (1996–1999) have partial document availability. Press conference transcripts are available from April 2011 and were held quarterly until January 2019, after which every meeting includes a press conference. Beige Books are available for approximately 239 of the 272 meetings; the remaining meetings (inter-meeting actions and emergency decisions) lack a dedicated Beige Book release. The pipeline processes whatever documents are available for each meeting, skipping filtration stages when a document is missing.

The optional $\mathcal{P}_5$ stage (Section 2.7) draws on news text from FactSet StreetAccount, a curated financial-news service whose editorial team distils real-time wire output into concise, market-focused summaries of macroeconomic data releases, Fed communications, financial-conditions news, and dealer commentary. Coverage begins in mid-2003 and reaches sufficient density from 2004 onward, which sets the effective sample for $\mathcal{P}_5$ (178 meetings); each pre-FOMC blackout window contains a median of roughly 12–18 articles per day after macro-topic filtering, of which 3–5 per day are policy-relevant.

For validation and comparison, the analysis incorporates several external data sources. The updated Romer and Romer (2004) series of monetary policy shocks is obtained from Acosta (2023).[10] Market-based surprise measures are obtained from Jarociński and Karadi (2020).[11] Fed Funds futures (FF1–FF4) are sourced from LSEG DataScope Tick History from 1996 onwards; pre-1996 data come from Gürkaynak et al. (2005). 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 over a 30-minute window as

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

[11] 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.

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.

Two ex-post cleaned shock series enter as comparators: Bauer and Swanson (2023a), which residualises market-based surprises on contemporaneous macro releases,[12] and Miranda-Agrippino and Ricco (2021), which residualises on internal Greenbook forecasts.[13] Both are matched to FOMC meeting dates. Remaining macro and financial data come from FRED, the Federal Reserve Board, and standard databases.

5 Results

5.1 Forecaster Synthesis Performance

The Forecaster synthesizes the previous meeting's statement and (from April 2011 onward) the Chair's press conference, the previous meeting's minutes, and the current Beige Book to form complete probability distributions over rate decisions. Table 3 reports a sequence of nested OLS regressions of the realized rate change on progressively richer subsets of the moments of those distributions, building up to the full posterior mean as the sole regressor in column (5).

Adding the Beige Book aggregate to the rate level raises R^2 from 0.007 to 0.203 (columns 1–2), confirming that regional economic narratives carry substantial rate-decision content. Once the Forecaster's first moment enters (column 5), it subsumes the Beige Book aggregate, raises R^2 to 0.542, and yields a slope coefficient of 1.000, statistically indistinguishable from unity (Proposition 1).[14] The standard deviation carries significant standalone predictive power (column 3) but loses all significance once the mean enters (columns 7–8): wider distributions correlate with larger rate moves during easing episodes, but this is a mean-variance correlation that $\mathrm{E}[\Delta i_t \mid \mathcal{B}_t]$ absorbs, not independent information.

Column (6) tests whether the slope coefficient depends on Beige Book availability. The interaction $\mathrm{E}[\Delta i_t \mid \mathcal{B}_t] \times d_{\mathrm{BB}}$ is insignificant in the pooled sample, but this masks a regime-

[12] Monthly series, Feb 1988–Dec 2023 (298 meeting months in sample). Source: https://www.michaeldbauer.com/files/monetary_policy_surprises_data.xlsx.

[13] Monthly series, Jan 1991–Dec 2009 (159 meeting months), bounded by Greenbook confidentiality at the time of their analysis.

[14] Throughout the empirical sections I write $\mathrm{E}[\Delta i_t \mid \mathcal{B}_t]$ for the LLM's reported posterior mean m_{4t}; this is the pipeline's approximation to the true documentary conditional expectation rather than the conditional expectation itself, with the gap absorbed into the extraction-error term η_t of Section 3.

Table 3: Forecaster Statistical Moments and Monetary Policy

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
i_{t-1}	−0.008	−0.018**	−0.005	−0.018***	−0.005	−0.006	−0.003	−0.003
	(0.008)	(0.007)	(0.009)	(0.007)	(0.004)	(0.004)	(0.006)	(0.006)
$\text{BB}^{\text{agg.}}$		0.320***	0.317***	0.324***	0.001	−0.006	0.001	−0.006
		(0.077)	(0.072)	(0.077)	(0.040)	(0.040)	(0.039)	(0.041)
$\sigma[\Delta i_t \mid \mathcal{B}_t]$			−1.101***				−0.323	−0.284
			(0.409)				(0.307)	(0.299)
$\text{Skew}[\Delta i_t \mid \mathcal{B}_t]$				−0.007*				0.003
				(0.004)				(0.003)
$\text{E}[\Delta i_t \mid \mathcal{B}_t]$					1.000***	0.690**	0.590*	0.606*
					(0.096)	(0.295)	(0.344)	(0.342)
d_{BB}						0.040	0.045	0.044
						(0.030)	(0.032)	(0.032)
$\text{E}[\Delta i_t \mid \mathcal{B}_t] \times d_{\text{BB}}$						0.328	0.404	0.406
						(0.338)	(0.373)	(0.373)
R^2	0.007	0.203	0.246	0.210	0.542	0.544	0.548	0.549
Adj. R^2	0.003	0.197	0.238	0.202	0.537	0.536	0.537	0.537
N	272	272	272	272	272	272	272	272

Note: OLS regressions of FOMC rate changes Δi_t on moments of the Forecaster's $\mathcal{P}_4$ posterior. $\text{E}[\Delta i_t \mid \mathcal{B}_t]$ is the posterior mean; $\sigma[\Delta i_t \mid \mathcal{B}_t]$ and $\text{Skew}[\Delta i_t \mid \mathcal{B}_t]$ are the posterior standard deviation and skewness. $\text{BB}^{\text{agg.}}$ is the Beige Book aggregate score; d_{BB} indicates whether a dedicated Beige Book is available for meeting t. Column (5) tests the calibration coefficient $H_0 : \beta = 1$ on the posterior mean. Newey-West HAC standard errors (4 lags) in parentheses. ***, **, *: 1%, 5%, 10%.

dependent contribution. On the common sample of $n = 189$ meetings, the Beige Book raises the probability mass on the realized outcome by +12.1pp for cuts (improving in 68% of cut meetings) and +1.95pp for hikes (improving in 56%), while slightly reducing it for holds (−1.78pp). The cut effect concentrates in the major recessionary easing episodes (2008, 2020), when regional conditions deteriorated faster than top-down Committee communications could acknowledge; this is consistent with the Beige Book's distinctive role as the only *bottom-up, real-economy* input with timing fresh to meeting M, while the statement, press conference, and minutes are *top-down Committee* documents released around meeting $M - 1$. Holds dominate the sample numerically, so the average effect washes out, while the Beige Book's contribution is concentrated in recessionary cuts (with a smaller positive contribution to hikes). Per-stage entropy and KL waterfall decompositions appear in Appendix C.1.2.

Table 4 evaluates the Forecaster's predictions against a horse race of simple benchmarks and an encompassing regression. Column (1) establishes the LLM's standalone performance: $\text{E}[\Delta i_t \mid \mathcal{B}_t]$ explains 53.9% of rate-change variance using only qualitative Fed documents.

As benchmarks, I consider a simple momentum rule (the previous meeting's rate change) and the slope of the yield curve. The momentum rule captures 37.2% of the variance (column 2), while yield-curve slopes are substantially more informative at 64.5% (column 3), consistent with

Table 4: LLM Text Forecast vs Naive Numerical Benchmarks

	(1)	(2)	(3)	(4)	(5)
$\mathrm{E}[\Delta i_t \mid \mathcal{B}_t]$	1.004***				0.381***
	(0.078)				(0.102)
Δi_{t-1}		0.609***		0.218***	0.063
		(0.083)		(0.062)	(0.074)
Slope (1Y − FFR)			0.656***	0.527***	0.419***
			(0.057)	(0.051)	(0.060)
Slope (2Y − FFR)			−0.287***	−0.218***	−0.162***
			(0.038)	(0.034)	(0.037)
R^2	0.539	0.372	0.645	0.674	0.694
Adj. R^2	0.537	0.370	0.642	0.671	0.689
N	272	271	272	271	271

Note: OLS regressions of FOMC rate changes Δi_t on LLM text-based forecast and naive numerical predictors. $\mathrm{E}[\Delta i_t \mid \mathcal{B}_t]$ is the LLM's expected rate change from qualitative Fed documents only (no market data). Slopes are Gurkaynak-Sack-Wright zero-coupon yields minus the federal funds rate, measured on the last trading day before each meeting. Column (5) is the encompassing test: both text and numerical predictors remain significant, confirming partially non-overlapping information content. Newey-West HAC standard errors (4 lags) in parentheses. ***, **, *: 1%, 5%, 10%.

the term structure encoding the market's comprehensive expectations of the future policy path.

Adding the LLM forecast to yield-curve slopes raises R^2 from 67.4% to 69.4%, with the LLM forecast entering at 0.381, significant at the 1% level (column 5). The text-based expectation also absorbs historical momentum: the coefficient on the lagged rate change Δi_{t-1} collapses from 0.218 (significant at the 1% level) in column (4) to a statistically insignificant 0.063, indicating that the LLM subsumes the persistence channel. While financial variables offer superior predictive power, yield curves conflate policy expectations with unobservable risk premia (Adrian et al., 2013; Cochrane, 2011) and non-policy flows. I rely on the LLM forecast because it conditions only on a known sequence of time-stamped public documents; that observable structure is what makes the resulting residual a document-conditioned innovation in the sense of Section 3 (derivations in Appendix E).

5.2 Properties of the Surprise

This subsection asks whether the narrative surprise behaves as a disciplined forecast error relative to its conditioning set. Three properties are tested: predictability from public non-document predictors, forecast efficiency (slope unity on the realized rate change), and incremental information beyond the other available measures.

Table 5: Predictability of FOMC Meeting Surprises from Bauer and Swanson (2023a) Predictors

	Narrative			Market-Based			
Variable	LLM	LLM ($\mathcal{P}_5$)	R&R (2004)	FF1	FF4	ED1	ED4
NFP Surprise	0.004	-0.000	0.065	0.002**	0.004**	0.004***	0.005***
	(0.004)	(0.003)	(0.060)	(0.001)	(0.002)	(0.001)	(0.002)
Nonf. Payrolls (12m)	0.004	-0.005	0.032	0.003**	0.005**	0.006***	0.009***
	(0.011)	(0.008)	(0.027)	(0.002)	(0.002)	(0.002)	(0.003)
S&P 500 (3m)	0.040***	0.014	-0.018	0.004	0.006*	0.007*	0.011***
	(0.013)	(0.012)	(0.015)	(0.003)	(0.004)	(0.004)	(0.004)
Term Spread (3m)	-0.017*	0.000	-0.048***	-0.001	-0.008***	-0.004	-0.006*
	(0.010)	(0.009)	(0.013)	(0.002)	(0.003)	(0.003)	(0.003)
Comm. Index (3m)	-0.003	-0.003	-0.001	0.003	0.008*	0.008**	0.011***
	(0.009)	(0.010)	(0.016)	(0.003)	(0.004)	(0.003)	(0.003)
Treasury Skewness	0.027**	0.010*	0.056***	0.004	0.005	0.005**	0.009***
	(0.011)	(0.006)	(0.012)	(0.003)	(0.003)	(0.003)	(0.003)
R^2	0.166	0.029	0.205	0.039	0.099	0.088	0.142
Observations	223	223	201	298	298	351	349

Note: Excluding ZLB meetings (January 2009 – December 2015), LLM $R^2 = 0.194$, LLM ($\mathcal{P}_5$) $R^2 = 0.040$, R&R $R^2 = 0.345$, confirming that predictability is not mechanically driven by the zero lower bound.

5.2.1 Predictability from Public Predictors

What a predictability test says about the surprise depends on its conditioning set. A surprise conditioned on the full public information set should be unpredictable from any pre-meeting public variable. Because the narrative surprise conditions only on documents, predictability from non-document public predictors should measure the share of public information that lies outside the documentary record.

Table 5 regresses narrative and market-based surprises on the six Bauer and Swanson (2023a) predictors: NFP surprise, 12-month employment growth, 3-month S&P 500 returns, term spread, commodity prices, and Treasury market skewness. The LLM ($\mathcal{P}_4$) column covers 223 FOMC meetings (1996–2024); the LLM+News ($\mathcal{P}_5$) column covers the same 223 meetings, with $\mathcal{P}_5 = \mathcal{P}_4$ for the (relatively few) meetings that lack inter-meeting StreetAccount coverage so the news stage has no content to ingest; the R&R column covers 201, reflecting Greenbook coverage limits.

The LLM surprise is moderately predictable, sitting within the range of market-based measures and slightly below R&R. The R&R comparison is the informative one: R&R conditions on the Greenbook, the Fed staff's private macro forecast, which should already absorb much of the public macro signal in the B&S predictors. That the LLM achieves comparable predictability without conditioning on private Fed forecasts is consistent with the LLM extracting the policy-relevant content of public documents efficiently. The patterns across individual predictors are themselves informative: 3-month S&P 500 returns predict the LLM surprise but not

Table 6: Measurement Validity: Slope of Δi_t on $\hat{s}_t$ ($H_0 : \beta = 1$)

	Narrative	Market-Based			
	LLM	FF1	FF4	ED1	ED4
Coefficient	1.005	2.367***	1.825**	1.682*	1.324
$H_0 : \beta = 1$	(0.091)	(0.484)	(0.358)	(0.360)	(0.308)
R^2	0.465	0.158	0.169	0.163	0.145
Observations	272	218	218	218	219

Note: Each column reports a separate univariate OLS regression of the realised FOMC rate change Δi_t on one surprise measure. Stars test $H_0 : \beta = 1$ (unbiased measurement), not $H_0 : \beta = 0$. Newey–West HAC standard errors (4 lags) in parentheses. The narrative surprise covers the full sample of FOMC meetings since 1996; the market-based measures begin in 1996 (Fed Funds Futures) and 1990 (Eurodollar futures), with N varying across columns due to surprise-series availability. *, **, ***: 10%, 5%, 1% rejection of $\beta = 1$.

R&R, indicating that the LLM does not fully condition on recent equity-market dynamics; NFP surprises and payroll growth are insignificant for both narrative measures and load instead on market-based futures, consistent with NFP releases falling inside the FOMC blackout window.

Within the signal-extraction framework of Section 3, this predictability reflects the public non-document wedge ξ_t^{pub}: the predictors lie in $\mathcal{M}_t \setminus \mathcal{B}_t$. Two diagnostics make extraction error η_t unlikely as the main driver: the slope coefficient on the realised rate change is indistinguishable from unity (Proposition 1), and residualising the surprise on the six B&S predictors leaves the impulse responses essentially unchanged (Appendix D.1.4). Incorporating inter-meeting StreetAccount articles via $\mathcal{P}_5$ closes the documentary window, and the joint significance of the predictors disappears at the 0.1% level; the news stage reverses $\mathcal{P}_4$'s direction in only 4.5% of meetings, so the dominant mode is magnitude correction rather than contradiction (Appendix C.2).

5.2.2 Forecast efficiency

To assess forecast efficiency, I estimate

$$\Delta i_t = \alpha + \beta \hat{s}_t + u_t. \tag{13}$$

As derived in Appendix E, the population slope can be written as

$$\beta = 1 + \frac{\mathrm{Cov}(m_{4t}, \eta_t)}{\mathrm{Var}(\hat{s}_t)}, \tag{14}$$

where m_{4t} is the LLM's posterior mean at stage $\mathcal{P}_4$ (Section 3). Under the signal extraction framework, $\beta = 1$ holds when the LLM aggregates its extracted signals as a Bayesian would and leakage is negligible (Proposition 1). The finding $\beta \approx 1$ is consistent with approximately

Table 7: Incremental Explanatory Power: Surprise Measures and the Realized Rate Change

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
LLM	1.005***	0.655***	0.594***	0.595***	0.614***	0.614***	0.597***
	(0.091)	(0.112)	(0.112)	(0.118)	(0.111)	(0.111)	(0.105)
R&R (2004)		0.545***	0.514***	0.514***	0.492***	0.492***	0.503***
		(0.107)	(0.107)	(0.105)	(0.098)	(0.098)	(0.098)
FF4			0.842*	0.844**	1.949***	1.968***	1.429***
			(0.450)	(0.418)	(0.550)	(0.614)	(0.536)
FF1				-0.007	2.846**	2.850**	2.658**
				(0.527)	(1.327)	(1.294)	(1.320)
MP1					-2.428***	-2.425***	-1.997**
					(0.836)	(0.858)	(0.840)
ED1						-0.022	-0.272
						(0.527)	(0.409)
ED4							0.455***
							(0.172)
R^2	0.465	0.577	0.603	0.603	0.639	0.639	0.647
Observations	272	175	175	175	175	175	175

Bayesian updating and small contamination leakage, though it does not separately identify either channel.[15]

Table 6 reports the result. The narrative surprise yields a coefficient ($\beta = 1.005$, s.e. $= 0.091$) statistically indistinguishable from unity, consistent with approximately Bayesian updating and small leakage (Proposition 1).[16] Market-based measures show $\beta > 1$, with FF4 and FF1 significantly exceeding unity, reflecting a different mechanism: futures-based surprises conflate unanticipated policy actions with information effects and, in the case of longer-horizon instruments, risk premia that push the coefficient above unity even absent measurement noise (Ricco & Savini, 2025).

5.2.3 Incremental Information vs. Other Measures

On the conventional-policy sample, the narrative surprise and R&R are complementary. Table 7 shows that both remain highly significant when combined (column 3), with R&R adding 12.2pp of explanatory power beyond the full LLM sample baseline (column 3 vs column 1). Because R&R coverage ends in 2018 (Greenbook confidentiality lag) while the LLM panel ex-

[15] This test is closely related to the Mincer and Zarnowitz (1969) forecast efficiency test, which regresses Δi_t on m_{4t} and tests $\alpha = 0$, $\beta = 1$. Table 3 reports $\beta = 1.000$ (s.e. $= 0.096$) in that specification, consistent with forecast rationality.

[16] The decomposition $\Delta i_t \equiv m_{4t} + \hat{s}_t$ holds by construction. What is not automatic is orthogonality: as shown in Section 3, $\beta = 1$ requires $\text{Cov}(m_{4t}, \hat{s}_t) = 0$, which follows from the law of iterated expectations if the LLM aggregates as a Bayesian would and leakage is negligible, but need not hold for an arbitrary processor. Once orthogonality is established, $R^2(m_{4t}) + R^2(\hat{s}) = 1$ follows, and the empirical counterparts $0.542 + 0.465 = 1.007$ are close to that benchmark, with the small excess over unity reflecting finite-sample noise rather than violation of the identity.

tends to 2026, column 2 reports the sample-matched LLM-only baseline on the LLM ∩ R&R intersection ($N = 178$). The matched-sample LLM-only $R^2 = 0.402$ is below the full-sample $R^2 = 0.463$ because the post-2018 LLM observations include the 2022–2024 active-cycle meetings on which the LLM extracts strongly. The structural Greenbook private-information wedge — R^2(col 3) − R^2(col 2) = 18.3pp — is therefore larger than the cross-sample 12.2pp comparison suggests, with the difference attributable to sample composition rather than to the underlying $\mathcal{G}_t \setminus \mathcal{M}_t$ gap. The LLM coefficient attenuates from 1.005 (full sample) to 0.935 (matched sample) to 0.645 when R&R enters, which is expected: both measures respond to the same underlying policy signals, so partial overlap is natural rather than a sign of misspecification. The 18.3pp structural increment is consistent with private Greenbook information in $\mathcal{G}_t \setminus \mathcal{M}_t$: R&R identification conditions on the Fed's internal forecasts, which contain policy-relevant content beyond the broader public information set. Adding all four high-frequency market-based measures beyond this narrative-Greenbook baseline raises R^2 by 4.8pp in total (column 7 vs column 3), consistent with most policy information flowing through official channels well before announcement-day pricing, as documented by Lucca and Moench (2015). A natural extension of this framework would apply the same LLM extraction methodology directly to Greenbook documents, constructing expectations from $\mathcal{G}_t$ rather than $\mathcal{B}_t$ and thereby measuring the private-information wedge narratively rather than as a residual econometric increment.

5.3 What Kind of Shock Is It?

Two further questions determine what the narrative surprise captures: does it survive at the zero lower bound when the policy rate cannot move, and does it load on the pure monetary policy or the central bank information component of FOMC announcements?

5.3.1 Identification Under Policy Rate Constraints

The Romer and Romer (2004) approach constructs a surprise as the residual from regressing the actual federal funds rate change on Greenbook forecasts: $\hat{s}_t^{R\&R} = \Delta i_t - \hat{\mathbb{E}}^{GB}[\Delta i_t]$. R&R is the natural narrative benchmark; Miranda-Agrippino and Ricco (2021) ends in 2012 and Bauer and Swanson (2023a) ends in 2019, so both are excluded from the ZLB analysis but enter the pre-crisis comparison (Appendix D.1.5). When the policy rate moves freely, the R&R residual identifies an unexpected deviation in the observed instrument. When it is constrained at the zero lower bound, $\Delta i_t = 0$ for almost every meeting by construction: the left-hand side loses its

Table 8: Statistical Moments Comparison During ZLB Period

Measure	N	Mean	Std Dev	Skew.	Kurt.	Min	Median	Max	Zero%	Pos%
LLM	51	1.2	7.4	6.04	38.65	-3.8	0.0	49.5	56.9	17.6
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. Skew. and Kurt. are sample skewness and kurtosis. Zero% = percentage of zero observations, Pos% = percentage of positive observations.

variation and the residual reduces mechanically to $-\hat{\mathbb{E}}^{GB}[\Delta i_t]$, approximating a counterfactual desired-rate gap under a reaction function estimated in a different regime.

The LLM approach behaves differently. The Forecaster reads what the Fed actually communicated, including forward guidance, balance-sheet signals, and commitment language, and forms a distribution over rate decisions directly from the text. It does not require the policy rate to move to register a non-trivial surprise: persistent accommodation communicated through language generates systematically negative surprises when the Fed signals more dovishness than a mechanical hold would imply. The ZLB is thus the regime where the two measures' properties diverge most sharply.

Table 8 confirms the divergence. R&R produces a near-symmetric distribution centred on negative-Greenbook-expectation territory, with a majority of ZLB meetings classified as positive surprises by construction: the sign tracks whether the Fed was internally forecasting cuts that did not materialise, not any actual policy deviation. The LLM measure clusters at zero, reflecting the Forecaster correctly anticipating the persistent ZLB hold from forward-guidance and balance-sheet language; the residual mass is asymmetric, with more dovish than hawkish surprises (the communicated stance more accommodative than the prior on rate decisions implied), and the right tail picks up rare regime-defining events such as the December 2015 liftoff. Sign here is the rate-decision forecast error, $s_t = \Delta i_t - \mathbb{E}[\Delta i_t \mid \mathcal{P}_4]$: at the ZLB $\Delta i_t = 0$ for almost every meeting, so a negative s_t records a meeting where the documents implied a positive expected change that did not materialise (and conversely for positive s_t), with forward-guidance language doing the work of moving $\mathbb{E}[\Delta i_t \mid \mathcal{P}_4]$. This asymmetry is by design: a measure that reads forward guidance directly should register zero whenever the language merely confirms continued accommodation. The relevant orthogonality condition is that surprises are unpredictable from prior information, not that they have zero mean in every regime subsample.

Table 9 shows the overlap directly. Across the full ZLB sample the two measures correlate at 0.460, significant at the 1% level, which might suggest they are picking up much of the same

Table 9: Correlation Robustness Analysis During ZLB Period

Sample	N	Correlation	Description
Full Sample	51	0.460** (0.001)	All FOMC meetings during the 2009–2015 zero-lower-bound regime.
Trimmed (1%-99%)	48	0.398** (0.005)	Drops observations outside the 1st and 99th percentiles of either measure to test outlier sensitivity.
No Key Episodes	49	0.060 (0.683)	Drops the 2009-01 QE1 announcement and the 2015-12 liftoff to remove the two large regime-transition meetings.

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

policy variation. The story changes once the two regime-defining meetings are removed, the January 2009 QE1 announcement and the December 2015 liftoff: the correlation collapses to 0.060, statistically indistinguishable from zero. The series therefore agree only at the meetings where the funds-rate target itself moves enough for R&R to register a surprise in its own right; elsewhere, when the rate is pinned and information flows through forward guidance and balance-sheet language, they share essentially no common variation.

Bügel et al. (2026) address the ZLB problem within R&R by substituting the Wu and Xia (2016) shadow rate for the conventional funds rate, and their diagnostics confirm the resulting shocks pass the serial-correlation and predictability tests that the conventional R&R series fails at the ZLB. LLM extraction provides an alternative that does not require a shadow-rate model: forward guidance, balance-sheet operations, and commitment language are processed directly from the text rather than approximated through a scalar rate substitute.

5.3.2 Channel Attribution: MP versus CBI Components

Jarociński and Karadi (2020) identify a pure monetary policy shock (MP: yields rise, stocks fall) and a central bank information shock (CBI: yields and stocks rise together) within a Bayesian VAR with sign restrictions. The decomposition is structural and external to any of the announcement-based measures considered here, which makes it a natural benchmark for consistency: a candidate MP shock should load one-for-one on the J&K MP component and have zero loading on the CBI component. Table 10 regresses three surprise measures on these two components without a constant.

The two market-based measures fail both legs of the test. M-A&R and B&S each reject $\beta_{\text{MP}} = 1$ at the 1% level and reject $\beta_{\text{CBI}} = 0$ at the 1% level. Their MP coefficients are attenuated and their CBI coefficients are economically large, so the cleaning step has not delivered a series that lines up with the structural MP shock. This is unsurprising given construction: M-A&R

Table 10: J&K Decomposition: Testing for CB Information Contamination

	M-A&R	B&S	LLM
β_{MP}	0.499***	0.740***	0.841**
	(0.064)	(0.088)	(0.360)
β_{CBI}	0.595***	0.584***	0.904
	(0.181)	(0.177)	(0.654)
R^2	0.508	0.662	0.110
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. LLM refers to the narrative surprise extracted 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.

and B&S are derived from the same announcement-window price changes that the J&K sign restrictions discipline, so any residual information-effect content in those prices passes through.

The LLM surprise is the only column where the joint null is not rejected, with standard errors three to four times wider than the announcement-window measures'. The non-rejection is therefore partly a low-power result rather than sharp identification of a pure MP shock, and the relative advantage strengthens in the post-2008 sample where forward guidance, balance-sheet language, and commitment text become the binding policy channels that announcement-window prices cannot mechanically span (Appendix D.1.7 reports the sub-sample table). Calibration against realized rate moves (Table 6) and the IRF diagnostics in Section 6.1 carry the absolute quality argument; this table is a relative consistency check whose informativeness depends on the policy regime.

The Bauer-Swanson robustness result in Appendix D.1.4 addresses a logically separate concern, predictability of the surprise from pre-meeting public observables, and shows that residualizing the surprise on the six B&S predictors leaves the impulse responses unchanged.

Together the two diagnostics support a partial-MP-identification reading. The narrative surprise registers monetary policy information through what the Fed communicates rather than through what the rate does, which is why it survives the ZLB where the rate-based R&R residual loses its left-hand-side variation. On the J&K decomposition, it is the only series that does not reject the structural pure-MP loading; the cleaned market-based alternatives reject both legs.

We stop short of claiming sharp identification of a pure MP shock: the J&K standard errors are three to four times wider on the LLM than on the announcement-window measures, so the non-rejection is partly a low-power result and a meaningful CBI admixture cannot be ruled out from this test alone. The honest scope claim is that the LLM behaves more like a structural MP shock than the cleaned market-based alternatives do, with the relative advantage strongest in the post-2008 regime where forward guidance, balance-sheet language, and commitment text are the binding policy channels. The next subsection sharpens this characterisation along the predictive and decomposition dimensions, where the surprise emerges as a *target-anchored communication shock*: linearly aligned with current-meeting target news, predictive of the future policy path, and orthogonal to the linear announcement-window basis.

5.4 What Does the LLM Surprise Capture?

I test two properties: the surprise's predictive power for future rate changes, and its lack of positive serial correlation. A third test, the decomposition on the Gürkaynak et al. (2005) target and path factors, appears in Appendix D.1.6. These tests suggest the LLM measure functions as a *target-anchored communication surprise*: it loads on the current rate decision, conveys persistent policy-stance information from pre-announcement Fed documents, and is not a noisy proxy for announcement-window derivatives.

I test whether the narrative surprise predicts future policy actions by estimating

$$\Delta i_{t+k} = \alpha + \beta\, \hat{s}_t + \varepsilon_{t+k}, \qquad k = 1, \ldots, 4 \tag{15}$$

where Δi_{t+k} is the rate change at the k-th subsequent meeting and $\hat{s}_t$ is the surprise at filtration stage $j \in \{\mathcal{P}_1, \mathcal{P}_3, \mathcal{P}_4\}$; $\mathcal{P}_2$ is excluded because press conferences began only in April 2011, which would roughly halve the available sample.[17]

All three filtration stages predict the next meeting's rate change and the cumulative six-meeting path (Tables 11–12), with $\mathcal{P}_1$ delivering the strongest single-meeting loading and a one-percentage-point cumulative six-meeting effect above unity; coefficients increase monotonically with horizon, consistent with a persistent policy-stance signal rather than a single meeting's news. The $\mathcal{P}_1 > \mathcal{P}_3 > \mathcal{P}_4$ ordering at every horizon aligns with FOMC statements as the primary

[17] Results are qualitatively similar when $\mathcal{P}_2$ is included on its post-2011 subsample. All three stages are estimated on the common subsample of $N = 186$ meetings where $\mathcal{P}_1$, $\mathcal{P}_3$, and $\mathcal{P}_4$ are jointly available, so coefficients are directly comparable.

Table 11: Forward Prediction: Future Rate Changes

	$\mathcal{P}_1$		$\mathcal{P}_3$		$\mathcal{P}_4$	
k	$\hat{\beta}$	R^2	$\hat{\beta}$	R^2	$\hat{\beta}$	R^2
1	0.379**	0.054	0.348**	0.049	0.323**	0.042
	(0.158)		(0.150)		(0.152)	
2	0.290**	0.031	0.262*	0.028	0.248*	0.025
	(0.133)		(0.140)		(0.139)	
3	0.220*	0.018	0.199	0.016	0.116	0.005
	(0.129)		(0.131)		(0.126)	
4	0.364**	0.050	0.316**	0.041	0.202*	0.016
	(0.147)		(0.146)		(0.119)	
N	186		186		186	

Note: Bivariate OLS regressions: $\Delta r_{t+k} = \alpha + \beta\, \hat{s}_t^{P_j} + \varepsilon_{t+k}$, where k is the number of meetings ahead. Newey-West HAC standard errors in parentheses (bandwidth k). ***, **, and * denote significance at 1%, 5%, and 10% levels.

source of path surprises (Gürkaynak et al., 2005): subsequent stages absorb path information into the prior, leaving $\mathcal{P}_4$'s residual smaller on the trajectory but still positive. This predictive content for the policy path coexists with the surprise's near-zero loading on the linear GSS path factor (Appendix D.1.6): the path-relevant content the documents carry need not be in the linear span of announcement-window derivatives, and the forward-prediction regressions show that this content is informative about future policy even when the announcement-window basis does not price it.

Forward prediction of rate changes is necessary but not sufficient for a forward guidance interpretation: a stale expectation would also predict future rate changes if the Fed's policy path is persistent (Rudebusch, 2002). To rule this out, I test whether surprises predict their own future values:

$$\hat{s}_{t+k} = \alpha + \beta\, \hat{s}_t + \varepsilon_{t+k}, \qquad k = 1, \ldots, 3 \tag{16}$$

Table 13 reveals no significant serial correlation at any horizon or filtration stage: the surprise predicts where rates *go* but not where future *surprises* go, consistent with forward-guidance content rather than mechanical staleness.

Table 12: Forward Prediction: Cumulative Rate Path

	$\mathcal{P}_1$		$\mathcal{P}_3$		$\mathcal{P}_4$	
k	$\hat{\beta}$	R^2	$\hat{\beta}$	R^2	$\hat{\beta}$	R^2
1	0.379**	0.054	0.348**	0.049	0.323**	0.042
	(0.158)		(0.150)		(0.152)	
2	0.668***	0.052	0.610**	0.048	0.571**	0.041
	(0.253)		(0.256)		(0.254)	
3	0.877**	0.045	0.798**	0.040	0.675*	0.028
	(0.374)		(0.384)		(0.376)	
4	1.248**	0.056	1.122**	0.049	0.883*	0.030
	(0.505)		(0.516)		(0.488)	
5	1.538**	0.058	1.385**	0.052	1.055*	0.029
	(0.598)		(0.602)		(0.554)	
6	1.725***	0.055	1.553**	0.048	1.149*	0.026
	(0.666)		(0.665)		(0.605)	
N	186		186		186	

Note: Bivariate OLS regressions: $\sum_{j=1}^{k} \Delta r_{t+j} = \alpha + \beta\, \hat{s}_t^{P_j} + \varepsilon_{t+k}$, where k is the cumulative horizon in meetings. Newey-West HAC standard errors in parentheses (bandwidth k). ***, **, and * denote significance at 1%, 5%, and 10% levels.

6 Macroeconomic and Financial Transmission

6.1 Impulse Responses to Monetary Policy Surprises

The narrative surprise produces a coherent macroeconomic and financial transmission pattern. Contractionary surprises generate persistent disinflation without the price puzzle, sustained output and industrial-production contraction, and a delayed rise in unemployment, with signs consistent across all variables and no *ex post* cleaning required. The yield curve responds in two phases: an initial compression driven by rising expected short rates, then recovery and steepening past zero as the cycle is digested.

I examine the dynamic effects of narrative surprises on macroeconomic and financial variables using two-stage local projections with instrumental variables (2SLP-IV; Jordà, 2005). The first stage instruments the federal funds rate with the narrative surprise:

$$\text{FFR}_t = \alpha^{(1)} + \pi \cdot \hat{s}_t + \sum_{j=1}^{L} \gamma_j^{(1)} \text{FFR}_{t-j} + \sum_{k=1}^{L} \boldsymbol{\delta}_k^{(1)\top} \mathbf{X}_{t-k} + u_t. \tag{17}$$

The second stage regresses each outcome on the fitted policy rate at horizon h:

$$y_{t+h} = \alpha_h^{(2)} + \beta_h \cdot \widehat{\text{FFR}}_t + \sum_{j=1}^{L} \gamma_{h,j}^{(2)} y_{t-j} + \sum_{k=1}^{L} \boldsymbol{\delta}_{h,k}^{(2)\top} \mathbf{X}_{t-k} + \varepsilon_{t+h}, \tag{18}$$

where y_{t+h} is the outcome at horizon h, $\hat{s}_t$ is the narrative surprise (instrument), and $\mathbf{X}_{t-k}$ in-

Table 13: Serial Correlation Test: Future Own Surprises

	$\mathcal{P}_1$		$\mathcal{P}_3$		$\mathcal{P}_4$	
k	$\hat{\beta}$	R^2	$\hat{\beta}$	R^2	$\hat{\beta}$	R^2
1	−0.024	0.001	0.015	0.000	−0.026	0.001
	(0.120)		(0.128)		(0.128)	
2	0.090	0.008	0.073	0.005	0.096	0.009
	(0.066)		(0.065)		(0.072)	
3	−0.064	0.004	−0.059	0.003	−0.072	0.005
	(0.067)		(0.057)		(0.070)	
N	186		186		186	

Note: Bivariate OLS regressions: $\hat{s}^{P_j}_{t+k} = \alpha + \beta\, \hat{s}^{P_j}_{t} + \varepsilon_{t+k}$, where k is the number of meetings ahead. Newey-West HAC standard errors in parentheses (bandwidth k). ***, **, and * denote significance at 1%, 5%, and 10% levels. Under the null that the surprise is a genuine innovation, $\hat{\beta} = 0$ at all horizons.

cludes macro controls. Identification requires $\mathbb{E}[\varepsilon_{t+h} \mid \hat{s}_t, \mathbf{X}_{t-k}, \mathrm{FFR}_{t-j}] = 0$: the narrative surprise is exogenous to future macroeconomic innovations conditional on pre-meeting controls (the substantive justification is in Section 3). Shock lags are set to zero, consistent with the AD/R&R narrative-instrument convention, and supported by the serial-correlation tests on $\hat{s}_t$ (Table 13, which finds no significant own-lag predictability at any horizon); robustness with $\hat{s}_{t-1}, \hat{s}_{t-2}$ as included exogenous controls leaves the IRFs qualitatively unchanged (Appendix D.1.2). Standard errors are Newey-West HAC with bandwidth $h + 1$ (Jordà, 2005), and all responses are normalised to a 25 bp narrative surprise (Jordà & Taylor, 2025; Ramey, 2016).[18]

The first stage includes 4 lags of the federal funds rate plus 4 lags of five macro controls (unemployment, log PCE Price Index, log industrial production, S&P 500, excess bond premium); the second stage includes 4 lags of the outcome and the same controls, with lagged FFR excluded. Macro outcomes use $L = 4$ lags following Aruoba and Drechsel (2024); financial variables use $L = 2$. The baseline asymmetric-controls levels specification departs from the symmetric LP-IV in Aruoba and Drechsel (2024): the federal funds rate is highly persistent ($\hat{\rho} \approx 0.99$) and the ZLB-excluded sample is small, a combination that Jordà and Taylor (2025) (§3) shows induces a small-sample bias of order $O(T^{-1})$ in levels LPs and makes the symmetric counterpart numerically fragile here. The long-difference variant of Jordà and Taylor (2025) replicates the headline magnitudes to within rounding, and the reduced-form LP on $\hat{s}_t$ recovers the same qualitative shapes (Appendices D.1.2, D.1.2).

The narrative surprise covers 272 FOMC meetings over 1996–2026; the IRFs that follow are estimated on the meeting-month subset surviving the ZLB exclusion. The communication-

[18] $\sigma(\hat{s}_t) = 14.7$ bp, so one-standard-deviation responses are roughly 59% of the reported magnitudes. HAC standard errors treat $\widehat{\mathrm{FFR}}_t$ as observed; Montiel Olea and Plagborg-Møller (2021) bound the resulting wedge under strong instruments, which the first-stage F-statistics (well above 10) support.

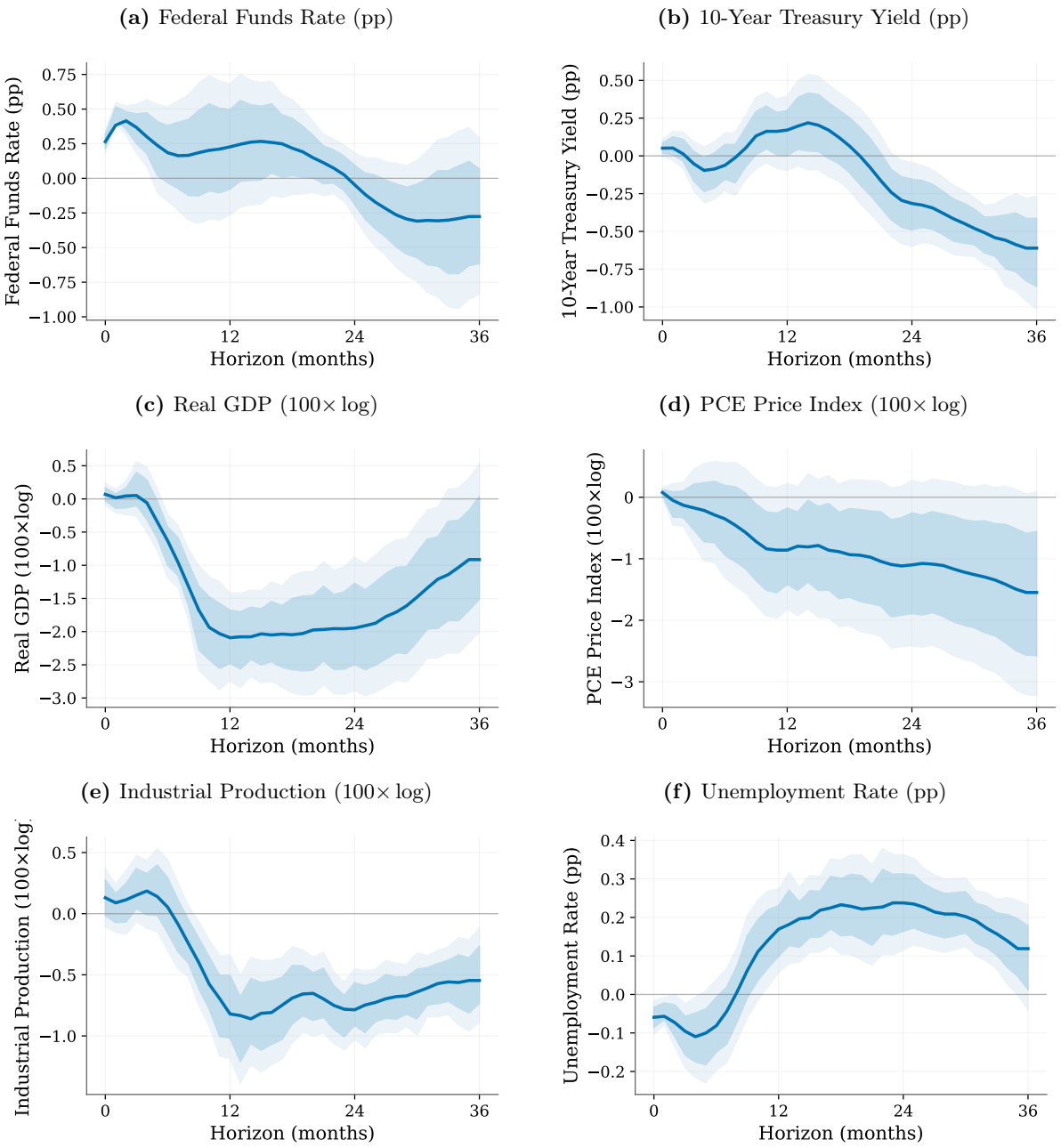
bundle scope is shared with Gertler and Karadi (2015), who instrument the two-year yield with medium-horizon futures, though the two strategies differ on identification (GK exploits announcement-window timing; the present instrument relies on the temporal cutoff between document release and the policy decision, Appendix B); cleaned high-frequency shocks (Bauer & Swanson, 2023b; Miranda-Agrippino & Ricco, 2021) instead target a narrower object by purging information and rule-revision components. Pre-crisis comparisons with these benchmarks and with narrative methods (Aruoba & Drechsel, 2024; Romer & Romer, 2004) on the common 1996:01–2008:10 window confirm directional consistency (Appendix D.1.5).

Figure 10 presents the baseline 2SLP-IV results, instrumenting the federal funds rate with the narrative surprise and excluding meetings in the conventional ZLB episode (2008M12–2015M12), where the federal funds rate is mechanically pinned and provides essentially no identifying variation as the instrumented variable. COVID-era meetings (2020M03–2022M03) are retained because they contain genuine FFR variation (the 50bp and 100bp emergency cuts of March 2020 and the March 2022 liftoff sequence) that the FFR-instrumenting strategy can exploit. Including the conventional-ZLB sample attenuates the FFR-instrumented macro responses substantially without flipping signs, reflecting the FFR's pinned variation rather than a methodological failure; Appendix D.1.3 reports the full-sample variant alongside a Gertler–Karadi-style specification that instruments the two-year Treasury yield rather than the policy rate. The 2Y-yield instrument is invariant to ZLB inclusion (first-stage $F = 47.53$ full sample vs $F = 28.41$ ZLB-excluded; first-stage coefficient stable to within one percent) and is the methodologically appropriate full-sample object; the FFR-instrumented headline corresponds to the conventional-policy estimand.

Following a 25bp contractionary surprise, the fitted federal funds rate rises 0.25pp on impact by construction and mean-reverts. Macroeconomic variables respond significantly only after month five. The PCE price index turns negative within months and reaches -0.7 to -0.9% by months 10–15, avoiding the "price puzzle" common in market-based identifications (Bauer & Swanson, 2023b; Miranda-Agrippino & Ricco, 2021). Real GDP declines -1.5 to -1.9% by months 10–15, industrial production -0.7 to -1.1%, and unemployment peaks around $+0.24$pp by month 15.

The identification evidence in Section 5.4 is consistent with this reading: the narrative surprise is not rejected against the joint $(\beta_{\text{MP}}, \beta_{\text{CBI}}) = (1, 0)$ benchmark, though with wide standard errors and a non-zero CBI point estimate that the data do not pin down precisely, so

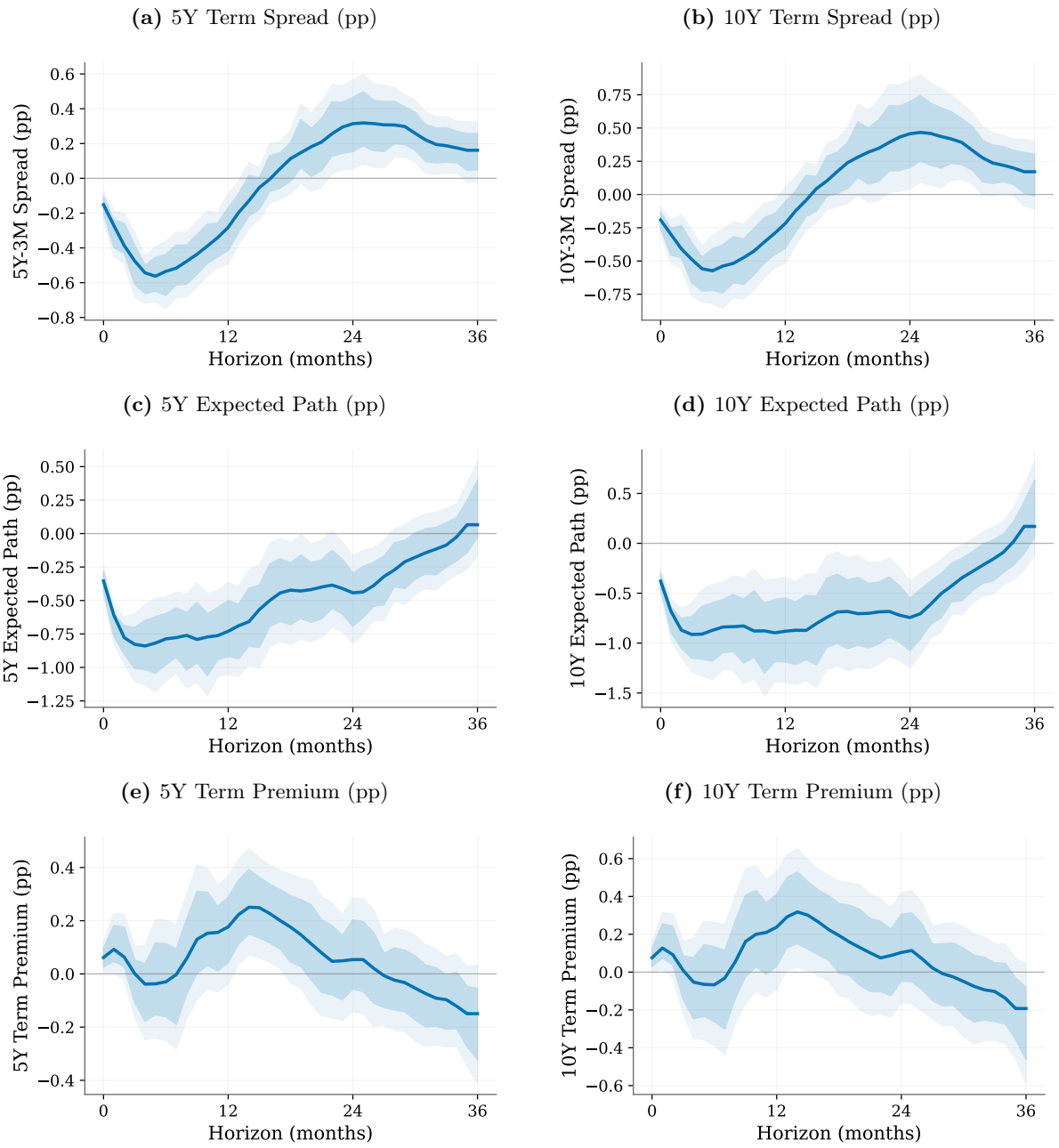
Figure 10: Macroeconomic Responses to a 25 bp Narrative Surprise



Note: Impulse response functions to a 25 bp document-conditioned narrative surprise. Two-stage local projections (2SLP-IV) with federal funds rate instrumented by the narrative surprise. Specification: 4 lags, 0 shock lags. First-stage controls: federal funds rate, unemployment, log PCE, log IP, SP500, EBP (4 lags each). Second-stage controls: unemployment, log PCE, log IP, SP500, EBP (lagged FFR excluded from second stage). Newey-West HAC standard errors with bandwidth $h + 1$ (Jordà, 2005). Solid line: 3-period moving average of the raw point estimates as a visual smoother (Jordà and Taylor, 2025, §5; bands are unaffected). Shaded areas: 68% (inner, darker) and 90% (outer, lighter) pointwise HAC confidence bands of the unsmoothed coefficients. Sample: 171 meeting-month observations (1996–2025), ZLB excluded, COVID included.

a CBI admixture cannot be ruled out from the J&K projection alone; the Bauer and Swanson (2023a) purging exercise leaves the responses unchanged (Appendix D.1.4). Appendix D.1.3 reports robustness to instrumenting the two-year Treasury yield in the spirit of Gertler and

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



Note: Impulse response functions to a 25 bp document-conditioned narrative surprise. 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: 2 lags for outcomes and controls (financial variables), 0 shock lags. Controls: same as Figure 10. Newey-West HAC standard errors with bandwidth $h+1$ (Jordà, 2005). Solid line: 3-period moving average of the raw point estimates as a visual smoother (Jordà and Taylor, 2025, §5; bands are unaffected). Shaded areas: 68% (inner, darker) and 90% (outer, lighter) pointwise HAC confidence bands of the unsmoothed coefficients. Sample: 171 meeting-month observations (1996–2025), ZLB excluded, COVID included.

Karadi (2015) and to including the ZLB period.

Turning to financial transmission, I decompose yield-curve spreads into expected short rates

and term premia:

$$r_t^{(n)} - r_t^{(1)} = \sum_{i=1}^{n-1} \left(1 - \frac{i}{n}\right) \mathbb{E}_t \Delta r_{t+i}^{(1)} + \theta_t^{(n)}, \tag{19}$$

where n is maturity expressed in months ($n = 60$ for the 5-year and $n = 120$ for the 10-year yield), $r_t^{(1)}$ is the 1-month yield, $\mathbb{E}_t \Delta r_{t+i}^{(1)}$ is the expected one-month change in the 1-month yield i months ahead, and $\theta_t^{(n)}$ is the term premium. The attribution between components is model-dependent; I use Favero and Fernández-Fuertes (2025)'s data-congruent specification, which imposes stationarity of term premia, a property standard affine models (e.g., Adrian et al., 2013) do not enforce. Figure 11 presents the complete decomposition for 5- and 10-year maturities.

Term spreads compress 0.4–0.5pp by month 5 as the front end rises, then recover and steepen past zero around month 13–15. Expected policy paths drive the initial compression: both 5Y and 10Y paths drop to about −0.4pp on impact and deepen to roughly −0.6pp by month 5 before recovering as the FFR response mean-reverts. Term premia trace a small impact uptick, a mid-horizon hump of roughly +0.17pp by month 15, and a persistent decline to −0.3 to −0.5pp by year three; spread recovery from month 15 onward is therefore driven primarily by the expected path normalising rather than by sustained increases in term premia.

The IRFs scale a communication-rich event into rate units rather than identifying the effect of an isolated rate change. The next subsection turns to asset-pricing evidence and asks whether the signal carries content that announcement-window derivatives do not already span.

6.2 Economic Validation Through Yield Curve Trading

This subsection is asset-pricing validation that the document-conditioned residual $\hat{s}_t$ carries directional, policy-path-relevant information not spanned by the standard announcement-window basis, primarily in active rate-cycle episodes where the Fed signals state-contingent decisions; it is not a tradable alpha and not an identified wedge. The span test below identifies a linear-orthogonal share — an *upper bound* on a documentary-vs-market wedge, since the residual $\hat{u}_t$ also absorbs LLM extraction noise and finite-basis approximation error. The investigation proceeds in two steps: a span projection plus yield-curve local projection localises where the orthogonal content appears across maturities, motivating the portfolio design that follows.[19]

[19]Sample sizes vary across exercises for data-availability reasons (span test, yield LP, baseline portfolio, sign-disagreement, paired-horizon difference each use different LLM/HF/GSS intersections); precise N is reported in each table, and cross-sample qualitative agreement is the integrity check.

To gauge how much of the LLM surprise is new relative to standard high-frequency measures, I project $\hat{s}_t^{\text{LLM}}$ onto the four Kuttner (2001)–Gürkaynak et al. (2005) announcement-window surprises:

$$\hat{s}_t^{\text{LLM}} = \boldsymbol{\beta}^\top \mathbf{f}_t + u_t, \qquad \mathbf{f}_t = (\text{FF1}_t,\, \text{FF4}_t,\, \text{ED1}_t,\, \text{ED4}_t)^\top. \tag{20}$$

The spanning null $\text{Var}(u_t) = 0$ holds if and only if $\hat{s}_t^{\text{LLM}} \in \text{span}(\mathbf{f}_t)$; the sample analogue is $1 - \widehat{R}^2$. On 217 meetings (1996–2024), the four factors are jointly significant yet explain only 18.5% of the variance: 81.5% of the narrative surprise lies outside the linear span of standard announcement-window derivatives.[20]

To trace where that orthogonal content propagates across maturities, I estimate

$$\Delta y_{n,t+h} = \alpha_h + \beta_h \hat{s}_t + \gamma_h \,\text{target}_t + \delta_h \,\text{path}_t + \text{controls} + \varepsilon_{t+h}, \tag{21}$$

where $\Delta y_{n,t+h}$ is the cumulative yield change from the previous meeting to h *meetings* ahead and target_t, path_t are the standardised Gürkaynak et al. (2005) announcement-window factors. Conditioning on these is the identifying step: β_h measures the response to the component of $\hat{s}_t$ orthogonal to announcement-window pricing. Figure 12 shows a sign change across the curve: the 1-month yield peaks above 30 bp around h=4, the 2-year crosses zero by h=4 and bottoms near −16 bp at h=6, and the 10-year is flat throughout. A 1m/2y flattener captures both sides of that sign change; cross-pair variants are in Appendix Table 40.

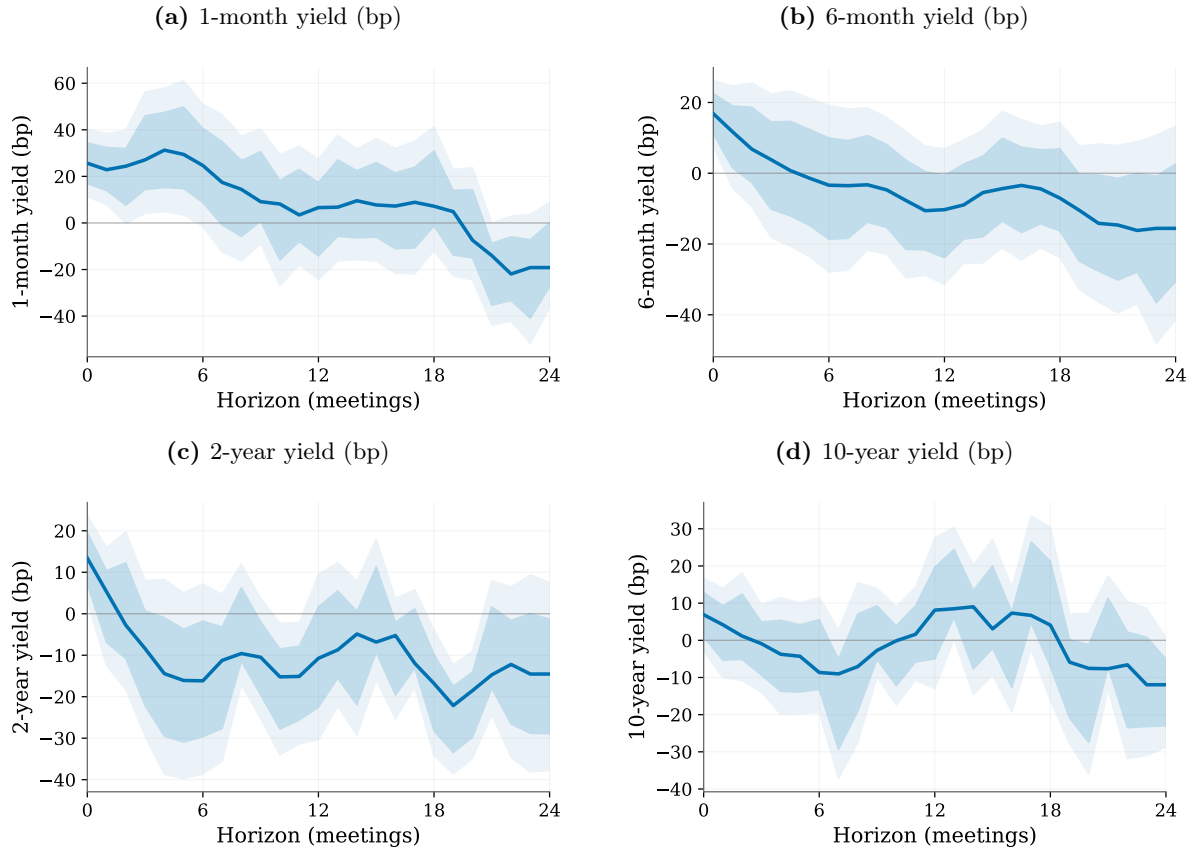
The cross-maturity sign change maps into an event-study portfolio that operationalises a directional test of $\hat{s}_t$. The position opens at meeting t when $|\hat{s}_t|$ exceeds its expanding-window two-thirds quantile (computed strictly on $\tau < t$): a positive surprise opens an equal-notional 1m/2y flattener, a negative surprise the mirror-image steepener; entry is the business day after the announcement, exit at least $H = 180$ calendar days later. Per-trade returns are in yield-spread basis points, gross of frictions; the exercise is a diagnostic of directional information, not a tradable strategy.[21]

Table 14 evaluates the curve response at $H = 180$ days on the eligible sample of 254 FOMC meetings with valid 180-day entry and exit yields (1996–2024); the top-tercile filter on $|\hat{s}_t|$ selects 70 events. On high-signal meetings the front-end curve moves in the direction predicted by $\hat{s}_t$

[20] The residual $\hat{u}_t$ mixes the document-vs-market wedge ξ_t^{doc} (the gap between the market's information set and what the documents reveal, defined formally in Appendix E.2) with finite-basis approximation error and LLM extraction noise; it is not a clean estimate of the wedge.

[21] Equal-notional weighting preserves the carry component; the duration-neutral alternative is in Appendix Table 39 and lifts per-trade returns from +11.4 bp to +34.0 bp.

Figure 12: Yield Curve Response to Narrative Surprise (Local Projections)



Note: Local projection estimates of equation (21). Dependent variable: cumulative yield change at maturity n from the previous meeting's yield to h *meetings* ahead (in basis points; horizon axis is therefore in meetings, not months). Shock: narrative surprise $\hat{s}_t$, normalized to 25 bp. Controls: 4 lags of the outcome variable and 4 lags of six macro controls (federal funds rate, unemployment rate, log PCE Price Index, log industrial production, log S&P 500, excess bond premium). HAC standard errors with bandwidth $h + 1$ (Jordà, 2005). Solid line: 3-period moving average of the raw point estimates as a visual smoother (Jordà and Taylor, 2025, §5; bands are unaffected). Shaded bands: 68% (inner, darker) and 90% (outer, lighter) pointwise HAC confidence intervals. Sample: 136 FOMC meetings (1999–2024), ZLB excluded. Yields: Gürkaynak–Sack–Wright (2007) zero-coupon curve.

at 62.9% of meetings, with a mean response of +11.4 bp per meeting, significant at the 5% level under the bootstrap. The unconditional and matched-N random-sign benchmarks average −1.46 bp and +0.76 bp respectively, and Panel B confirms that no entry-time term-structure factor absorbs the response ($R^2 = 0.027$).

The cleanest test of the not-spanned claim runs the same portfolio on the orthogonal residual $\hat{u}_t$ from (20), on the restricted sample where that residual is defined. Table 15 reports it: the raw LLM signal yields +15.3 bp per meeting and the orthogonal residual a statistically indistinguishable +13.6 bp, with duration-neutral weighting lifting the response to +34.0 bp at the 1% level. Because $\hat{u}_t$ is by construction orthogonal to (FF1, FF4, ED1, ED4), this is the direct evidence that the payoff is not a re-projection of the linear high-frequency span. The

Table 14: Validation of $\hat{s}_t$: Front-End Curve Response on High-Signal Meetings

PANEL A: Front-end curve response, high-signal meetings vs. unconditional and matched-N random benchmarks.

	N	Hit (%)	Mean (bp)	Sharpe	p
High-signal meetings (top tercile $\lvert\hat{s}_t\rvert$, signed by $\hat{s}_t$)	70	62.9	+11.37	+0.38	0.035
All meetings (unconditional, sign $\equiv +1$)	254	44.5	-1.46	-0.06	0.640
Random meetings + random sign (matched $N = 70$, 100 sims)	70	51.1	+0.76	+0.02	0.168

PANEL B: Orthogonality to entry-time term-structure factors (OLS).

	$\lvert\hat{s}_t\rvert$	Slope (2y−1m)	Level (1m)	Slope mom. (90d)	FFR mom. (90d)	Constant
Coefficient	+23.11	−0.01	−0.55	+13.88	+3.98	+7.16
	(51.90)	(18.72)	(4.39)	(20.39)	(13.36)	(21.59)
N	70					
R^2	0.027					

Note: 1m/2y equal-notional flattener, 180-day hold, per-meeting yield-spread change in basis points signed by the convention indicated. Sharpe annualised by $\sqrt{365.25/H}$. Two-sided p-values from a stationary block bootstrap (Politis & Romano, 1994): 5,000 resamples, mean block length $L = 4$ events; the random-meeting p averages over 100 simulations. Panel B regresses the per-meeting response (bp) on entry-time term-structure factors (pp of yield); HC3 robust standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

result is stable across holding windows from $H = 90$ to $H = 360$ days, with point estimates rising from +10.4 to +16.4 bp; cross-pair robustness is in Appendix Table 40 and a horizon sweep against ED4/ED1 in Appendix D.2.

Figure 13 traces cumulative returns by holding horizon for the LLM signal alongside ED4 and ED1: the LLM-signed flattener accumulates returns earlier in the holding period (roughly 8% at three months versus 2% for ED4), with the gap closing only after twelve months; the paired-meeting LLM–ED4 difference is positive in point estimate at every horizon, but the short-horizon gap is not statistically significant even before any multiple-testing correction on the small paired sample ($N_h \in [21, 25]$), and the only horizon clearing raw 5% does not survive Holm adjustment (Appendix D.2).

The result is not specific to one extraction model: at a common prompt version v30.2 with all six LLMs run on the full sample, the four frontier-class models (DeepSeek-V3.1, GPT-5-mini, Qwen-3.6, GPT-4.1-mini) all deliver Sharpe ratios in $[0.33, 0.47]$ at the 5% level (Appendix D.2); the two smaller models are directionally consistent but statistically insignificant, suggesting that extraction quality, not idiosyncratic model choice, is the binding constraint on shock quality.

The baseline result is asymmetric across the sign of the surprise, and that asymmetry is the central scope condition of the trading evidence (Table 16, Panel B). Dovish surprises ($s_t < 0$,

Table 15: Per-Trade Returns: Raw, Orthogonal-Component, and Duration-Neutral Variants

Variant	N	Per-trade (bp, yield-spread)	Sharpe	p
Baseline: raw $\hat{s}_t$, equal-notional, full sample	70	$+11.37^{**}$	+0.38	0.035
Restricted sample, raw $\hat{s}_t$	59	$+15.31^{***}$	+0.52	0.009
Restricted sample, orthogonal residual $\hat{u}_t$	56	$+13.61^{**}$	+0.46	0.034
Baseline pair, duration-neutral weighting	70	$+34.01^{***}$	+0.45	0.010

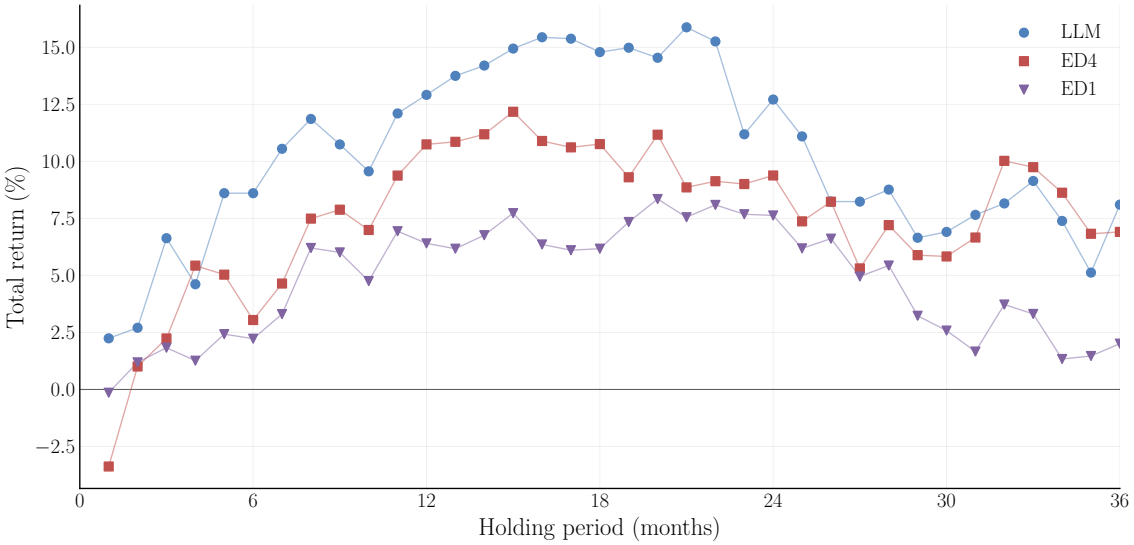
Note: Per-trade returns in yield-spread basis points. The first row is the baseline (Table 14). Rows 2–3 use the LLM ∩ JK common sample on which the orthogonal residual $\hat{u}_t$ is defined; the raw-minus-residual paired difference on those meetings is +0.98 bp ($p = 0.261$), so the two are statistically indistinguishable. The duration-neutral row (Table 39) weights legs so a parallel shift earns zero. The paper retains 1m/2y as the baseline because it is the maturity pair most tightly linked to the LP sign change in Figure 12. Two-sided p-values from a stationary block bootstrap (Politis and Romano, 1994; 5,000 resamples, $L = 4$). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

opening a steepener) earn +16.1 bp per trade at the 1% level, with a Sharpe of 0.61 and a 75.7% hit rate; hawkish surprises ($s_t > 0$, opening a flattener) earn only +6.1 bp with a Sharpe of 0.18 and a 48.5% hit rate, statistically indistinguishable from random sign. The directional skill the baseline reports is therefore a dovish-side effect; the hawkish leg, on which the strategy is named, does not carry the result. The result is also regime-dependent (Panel A). Excluding the 2022–2024 tightening cycle (entry date < 2022-01-01) drops the per-trade return from +11.4 bp to +5.9 bp, the Sharpe from 0.38 to 0.22, and the result is no longer statistically significant. The natural reading is that the baseline is driven by the recent cycle; a sample-composition reading qualifies it. Roughly a third of the pre-2022 window is at the zero lower bound (2008-12 through 2015-12 and 2020-03 through 2021-12), when the front of the curve is mechanically pinned and a 1m/2y flattener has little to capture; the strategy's economic content is directional skill on an *active* rate cycle, and the pre-2022 subsample is by composition the period with the least cycle to read. Panel C reports the disagreement-subsample economic test against ED4: the LLM-direction magnitude-weighted return is +8.5 bp, significant at the 5% level; the full sign-disagreement decomposition, unweighted paired-direction diagnostics, and joint horse race are in Subsections D.2 and D.2.

Figure 14 visualises the regime breakdown: the 2022–2024 cycle accounts for 57% of cumulative response, the pre-crisis tightening (1996–2007) contributes 24%, the COVID window 18%, and the post-2015 normalisation is roughly flat. Three cycles contribute positively; only the quiet post-2015 period is a net drag.

Two robustness checks rule out alternative readings of the residual (Appendix Table 47, Panels A–B): the per-trade return is uncorrelated with entry-month macro variables ($R^2 = 0.15$, no factor significant at 5%), ruling out a time-varying risk-premium interpretation, and with

Figure 13: Cumulative portfolio returns by holding period: LLM, ED4, ED1



Note: Geometric cumulative returns from compounding all top-tercile events on the 1m/2y equal-notional flattener, by holding period in months, applied separately to the LLM surprise, ED4, and ED1 signals on the common HF ∩ LLM sample (1996–2024). The figure displays only the point estimates: the relevant inference object is the *paired* difference between signals at short horizons (LLM minus ED4 at $h = 3, 6$ months), not the marginal sampling uncertainty of each signal at each horizon. Per-trade significance at the baseline 180-day horizon is reported in Table 14.

entry-time fixed-income factors (level, slope, curvature, carry, momentum; $R^2 = 0.03$), ruling out a disguised generic exposure. The substantive decomposition (Panel C) projects $\hat{s}_t$ on the Gürkaynak et al. (2005) target and path factors directly (rather than the four-factor (FF1, FF4, ED1, ED4) basis used in equation (20)) and yields a target-factor loading of $\hat{\beta}_{\text{target}} = +0.047$, significant at the 1% level, and a path-factor loading $\hat{\beta}_{\text{path}} \approx 0$ statistically indistinguishable from zero, with $R^2 = 0.124$. The LLM signal therefore correlates linearly with current-meeting target news but *not* with the linear announcement-window path factor; the 88% orthogonal residual may reflect a non-linear function of the path factor, path-relevant content uncorrelated with the linear GSS basis, finite-basis approximation error, or LLM extraction noise. The diagnostics that follow probe *when* that orthogonal content is informative rather than attribute specific language to it: $\hat{s}_t$ is by construction the announcement information *not* anticipated by the pre-meeting documents, so any text feature defined on those documents is most naturally read as a moderator of the residual, not as content the residual carries.

A complementary observation is that the surprise's profitability concentrates in dovish episodes. Splitting the post-2008 sample into surprise terciles *within* each Fed communication regime (pre-FG, ZLB/FG, post-liftoff, COVID/ZLB, 2022+) and counting conditional constructions ('if', 'until', 'should', 'provided', 'unless', and the modal verbs 'would', 'could', 'might')

Table 16: Subsample and Asymmetry Breakdowns for the Baseline Strategy

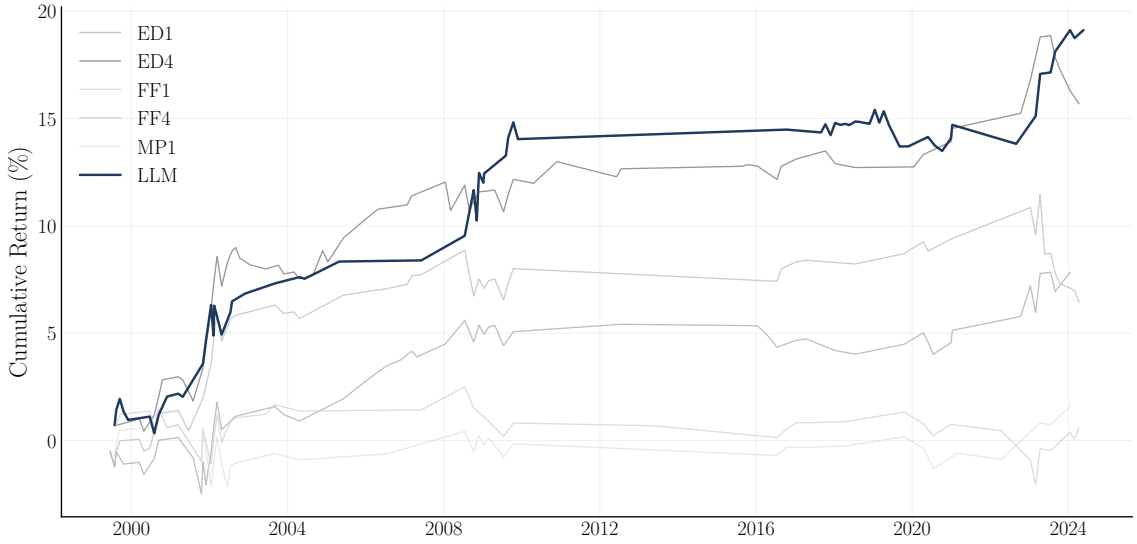
Subsample / split	N	Per-trade (bp)	Sharpe	Hit (%)	p
Panel A: subsample stability					
Baseline (full sample)	70	+11.37	+0.38	62.9	0.035**
Ex-2022 (entry < 2022-01-01)	62	+5.88	+0.22	61.3	0.123
Panel B: hawkish vs dovish asymmetry					
Hawkish surprises (s>0, flattener)	33	+6.07	+0.18	48.5	0.534
Dovish surprises (s<0, steepener)	37	+16.10	+0.61	75.7	0.009***
Panel C: disagreement subsample (LLM vs ED4)					
LLM direction, magnitude-weighted	76	+8.49	+0.40	57.9	0.035**

Note: Per-meeting unsigned 1m/2y flattener payoff over a 180-day hold (matching Table 14). Panel A varies the sample window: "Ex-2022" restricts entries to before 2022-01-01, isolating the result from the 2022–2024 tightening cycle. Panel B splits the baseline sample by the sign of the original surprise: hawkish events ($s_t > 0$) open as flatteners, dovish events ($s_t < 0$) as steepeners. Panel C reports the 76-meeting disagreement subsample where $\text{sign}(\hat{s}_t^{\text{LLM}}) \neq \text{sign}(\text{ED4}_t)$, signed by the LLM direction and weighted by realised payoff; the ED4 row is its exact mirror by construction. Full sign-disagreement decomposition (Table 36) and unweighted paired-direction diagnostics in Subsection D.2; the formal information-distinctness test (joint horse race) is in Subsection D.2. p-values from a stationary block bootstrap (Politis and Romano, 1994; 5,000 resamples, $L = 4$ events). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

over the prior statement and prior minutes shows that documents preceding dovish surprises use roughly 1.5 to 2 more conditional markers per 1,000 words than documents preceding hawkish surprises; the gap survives in the prose direction in 20 of 23 extraction runs spanning six LLM families, is significant at the 5% level in 13, and never significantly reverses (Appendix D.2). Dovish episodes are typically signalled through state-contingent easing paths ("rates will remain at the lower bound *until*..."); hawkish episodes lean on unconditional commitments. Asymmetric monetary transmission compounds this, easing surprises move the front of the curve more, and more persistently, than tightening surprises of equal size, but the 75.7% vs. 48.5% hit-rate gap is too wide to be explained by payoff convexity alone, so episode selection is doing real work.

Two scope conditions delimit the result. First, the LLM signal and ED4 share the same rate channel ($R^2 = 0.155$ vs. 0.154): the LLM advantage is cross-sectional directional content where the two disagree, not a new transmission mechanism. Second, the effect concentrates in active rate cycles; ex-2022 the per-trade return is no longer significant and the paired LLM–ED4 difference is statistically weak on the small paired sample (Appendix D.2). The exercise is therefore measurement validation rather than an alpha claim: on this delimited subset of meetings, the document-conditioned residual carries directional content orthogonal to the linear announcement-window basis, which we read as forward-guidance information given where it loads on the curve. Accordingly, the trading evidence should not be read as establishing a

Figure 14: Cumulative Front-End Response on High-Signal Meetings, 1996–2025



Note: Cumulative front-end curve response on top-tercile high-signal meetings (Section 6.2), signed by $\hat{s}_t$ and compounded across meetings as a regime-stability check on the directional content of the signal. The visualization uses a 300-day window (rather than the baseline 180-day window) so that overlapping cycles are visually separable; the per-meeting statistics, placebo battery, and inference in Table 14 all use the 180-day window. Sample: 1996–2025 FOMC meetings, ZLB excluded. Reported in yield-spread basis points captured over the window (the per-meeting yield-spread change between the two legs).

symmetric or full-sample implementable strategy; it shows that, when Fed documents contain a large LLM-measured surprise, the component unexplained by standard high-frequency factors forecasts the relevant curve direction, with the strongest pricing content on the dovish leg in active rate-cycle episodes.

7 Conclusion

Measuring monetary policy surprises requires first measuring expectations and then stating clearly which information set those expectations condition on. This paper extracts those expectations directly from the sequential release of central bank communications via multi-agent LLM synthesis on a known, time-stamped documentary record. The *New Hope* is methodological: as extraction methods become more sophisticated, identification quality improves at the source rather than through *ex post* adjustments. A four-stage filtration over public Fed communications produces a well-calibrated prior whose residual is a disciplined forecast error against an explicit conditioning set. Three exercises validate the measure. First, the surprise is moderately predictable from public non-document predictors, which quantifies the gap between the documentary record and a richer information set; a news-augmented stage ($\mathcal{P}_5$) closes that

gap. Second, the measure produces theoretically coherent impulse responses with persistent contractionary transmission. Third, an asset-pricing test shows the residual carries directional, policy-path-relevant information orthogonal to the linear announcement-window basis, loading where forward-guidance theory predicts (the front of the curve), with channel diagnostics that scope to active rate-cycle episodes whose pre-meeting documents flag state-contingent policy decisions rather than to a sentiment-dictionary baseline. As an auxiliary result, splitting the surprise by documentary timing (an announcement-day residual and an expectation revision over the pre-FOMC blackout window) recovers a decomposition along the lines of Jarociński and Karadi (2020), separating information from monetary policy without relying on asset-price sign restrictions.

Methodologically, the paper delivers an auditable communication-based innovation with an explicit conditioning set: the surprise conditions on a known, time-stamped documentary record. This structure makes the remaining wedges between the documentary record, the full public information set, and private Fed forecasts interpretable, rather than hiding them inside announcement-window prices or *ex post* cleaning regressions.

The framework extends naturally beyond this proof of concept: central banks could incorporate internal forecasts alongside public communications, market participants could integrate real-time news flows arriving during blackout periods, and researchers could apply the architecture to other domains where institutional communications precede market-moving events, including earnings announcements, regulatory decisions, and geopolitical developments.

The results carry implications beyond monetary economics. For policymakers, the findings validate the Fed's communication strategy: public documents explain over half of rate-decision variance, consistent with communication serving as a substantive policy instrument rather than a cosmetic one. For researchers, the framework offers a scalable alternative to both hand-coded narrative measures and cleaned high-frequency identification, while keeping the estimand explicit. The genuine policy surprises, what remains unpredictable from the public record, can now be measured in real time, opening new questions about monetary transmission and the role of conditional language in central bank communication.

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. https://www.ijcb.org/journal/v19n3/new-measure-central-bank-transparency-and-implications-effectiveness-monetary-policy

Adrian, T., Crump, R. K., & Moench, E. (2013). Pricing the term structure with linear regressions. *Journal of Financial Economics*, *110*(1), 110–138. https://doi.org/10.1016/j.jfineco.2013.04.009

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. https://doi.org/10.1016/j.jeconom.2024.105921

Ahrens, M., & McMahon, M. (2021). Extracting economic signals from central bank speeches. *Proceedings of the Third Workshop on Economics and Natural Language Processing*. https://aclanthology.org/2021.econlp-1.12/

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.

Apel, M., & Blix Grimaldi, M. (2014). How Informative Are Central Bank Minutes? *Review of Economics*, *65*(1), 53–76. https://doi.org/10.1515/roe-2014-0104

Aruoba, S. B., & Drechsel, T. (2024). *Identifying Monetary Policy Shocks: A Natural Language Approach* (tech. rep.). National Bureau of Economic Research. https://doi.org/10.3386/w32417

Bauer, M. D., & Swanson, E. T. (2023a). An Alternative Explanation for the "Fed Information Effect". *American Economic Review*, *113*(3), 664–700. https://doi.org/10.1257/aer.20201220

Bauer, M. D., & Swanson, E. T. (2023b). A reassessment of monetary policy surprises and high-frequency identification. *NBER Macroeconomics Annual*, *37*(1), 87–155. https://doi.org/10.1086/723574

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

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 Communication 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.

Bügel, D., Hidalgo, A., & Luetticke, R. (2026). *Unconventional but different after all? A unified series of narrative monetary policy shocks* [CEPR Discussion Paper No. 19163, R&R at Journal of Money, Credit and Banking]. https://www.ralphluetticke.com/

Bybee, J. L. (2023a). The Ghost in the Machine: Generating Beliefs with Large Language Models. *Working paper, Yale School of Management*.

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. https://doi.org/10.1353/eca.2012.0004

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

Christiano, L. J., Eichenbaum, M., & Evans, C. L. (1999). Monetary policy shocks: What have we learned and to what end? *Handbook of macroeconomics*, *1*, 65–148. https://doi.org/10.1016/S1574-0048(99)01005-8

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. (2024). *Did I Make Myself Clear? The Fed and the Market under the 2020 Monetary Policy Framework* (CEPR Discussion Paper No. 19360). Centre for Economic Policy Research.

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. https://doi.org/10.1257/jel.37.4.1661

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

Cochrane, J. H. (2011). Presidential address: Discount rates. *The Journal of finance*, *66*(4), 1047–1108. https://doi.org/10.1111/j.1540-6261.2011.01671.x

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. https://www.bis.org/publ/work1231.htm

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263. https://doi.org/10.1080/07350015.1995.10524599

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*.

Feng, T., Trinh, T. H., Bingham, G., Hwang, D., Chervonyi, Y., Jung, J., Lee, J., Pagano, C., Kim, S.-h., Pasqualotto, F., et al. (2026). Towards Autonomous Mathematics Research. *arXiv preprint arXiv:2602.10177*.

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-LMs: Language Models for Central Banking*. Bank for International Settlements, Monetary; Economic Department. https://www.bis.org/publ/work1215.htm

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. https://arxiv.org/abs/2309.17322

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., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., & Zhang, X. (2024). Large Language Model based Multi-Agents: A Survey of Progress and Challenges. *International Joint Conference on Artificial Intelligence (IJCAI)*. https://dl.acm.org/doi/10.24963/ijcai.2024/890

Gürkaynak, R. S., Sack, B., & Swanson, E. (2005). The Sensitivity of Long-Term Interest Rates to Economic News: Evidence and Implications for Macroeconomic Models. *American Economic Review*, *95*(1), 425–436. https://doi.org/10.1257/0002828053828443

Hack, L., Istrefi, K., & Meier, M. (2024). *The Systematic Origins of Monetary Policy Shocks* [CEPR Discussion Paper No. 19063; CRC TR 224 DP 2024/557; Banque de France Working Paper No. 1021].

Hansen, A. L., & Kazinnik, S. (2023). Can ChatGPT Decipher Fedspeak. *Available at SSRN*. https://doi.org/10.2139/ssrn.4399406

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

Hanson, S. G., & Stein, J. C. (2012). Monetary Policy and Long-Term Real Rates. *Finance and Economics Discussion Series*, (2012-46). https://doi.org/10.17016/FEDS.2012.46

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70.

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* (CEPR Discussion Paper No. 19923) (This version: September 11, 2025; First version: February 3, 2025). Centre for Economic Policy Research.

Jiang, H. (2023). A Latent Space Theory for Emergent Abilities in Large Language Models. *arXiv preprint arXiv:2304.09960*.

Jordà, Ò. (2005). Estimation and inference of impulse responses by local projections. *American economic review*, *95*(1), 161–182. https://doi.org/10.1257/0002828053828518

Jordà, Ò., & Taylor, A. M. (2025). Local projections. *Journal of Economic Literature*, *63*(1), 59–110. https://doi.org/10.1257/jel.20241521

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

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, 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. https://doi.org/10.1016/S0304-3932(01)00055-1

Leeper, E. M. (1997). Narrative and VAR Approaches to Monetary Policy: Common Identification Problems. *Journal of Monetary Economics*, *40*(3), 641–657. https://doi.org/10.1016/S0304-3932(97)00051-2

Li, J., Zhang, Q., Yu, Y., Fu, Q., & Ye, D. (2024). More Agents Is All You Need. *arXiv preprint arXiv:2402.05120*. https://arxiv.org/abs/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 Forecast 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. https://doi.org/10.1111/jofi.12196

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

Mincer, J. A., & Zarnowitz, V. (1969). The Evaluation of Economic Forecasts. In J. A. Mincer (Ed.), *Economic forecasts and expectations: Analysis of forecasting behavior and performance* (pp. 3–46). National Bureau of Economic Research. https://www.nber.org/system/files/chapters/c1214/c1214.pdf

Miranda-Agrippino, S., & Ricco, G. (2021). The transmission of monetary policy shocks. *American Economic Journal: Macroeconomics*, *13*(3), 74–107. https://doi.org/10.1257/mac.20180124

Montiel Olea, J. L., & Pflueger, C. (2013). A robust test for weak instruments. *Journal of Business & Economic Statistics*, *31*(3), 358–369. https://doi.org/10.1080/00401706.2013.806694

Montiel Olea, J. L., & Plagborg-Møller, M. (2021). Local projection inference is simpler and more robust than you think. *Econometrica*, *89*(4), 1789–1823. https://doi.org/10.3982/ECTA18756

Nakamura, E., & Steinsson, J. (2018). High-Frequency Identification of Monetary Non-Neutrality: The Information Effect. *The Quarterly Journal of Economics*, *133*(3), 1283–1330. https://doi.org/10.1093/qje/qjy004

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 in 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

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*.

Plagborg-Møller, M., & Wolf, C. K. (2021). Local projections and VARs estimate the same impulse responses. *Econometrica*, *89*(2), 955–980. https://doi.org/10.3982/ECTA17813

Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, *89*(428), 1303–1313. https://doi.org/10.1080/01621459.1994.10476870

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. https://doi.org/10.1016/bs.hesmac.2016.03.003

Ricco, G., & Savini, E. (2025, April). *Decomposing Monetary Policy Surprises: Shock, Information, and Policy Rule Revision* (CEPR Discussion Paper No. 20166). Centre for Economic Policy Research. https://cepr.org/publications/dp20166

Romer, C. D., & Romer, D. H. (1989). Does Monetary Policy Matter? A New Test in the Spirit of Friedman and Schwartz. *NBER Macroeconomics Annual*, *4*, 121–184. https://doi.org/10.1086/654119

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

Romer, C. D., & Romer, D. H. (2004). A New Measure of Monetary Shocks: Derivation and Implications. *American Economic Review*, *94*(4), 1055–1084. https://doi.org/10.1257/0002828042002651

Rudebusch, G. D. (2002). Term structure evidence on interest rate smoothing and monetary policy inertia. *Journal of Monetary Economics*, *49*(6), 1161–1187. https://doi.org/10.1016/S0304-3932(02)00149-6

Sarkar, S. K., & Vafa, K. (2024, March). *Lookahead Bias in Pretrained Language Models* (Working Paper) (SSRN 4754678). Harvard University. https://doi.org/10.2139/ssrn.4754678

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. *arXiv preprint arXiv:2411.16569*.

Sims, C. A. (1980). Macroeconomics and Reality. *Econometrica: Journal of the Econometric Society*, 1–48. https://doi.org/10.2307/1912017

Sims, C. A. (1992). Interpreting the Macroeconomic Time Series Facts: The Effects of Monetary Policy. *European Economic Review*, *36*(5), 975–1000. https://doi.org/10.1016/0014-2921(92)90041-T

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. https://doi.org/10.1257/jep.15.4.101

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. https://doi.org/10.1111/ecoj.12593

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- and Longer-Term Interest Rates. *American Economic Review*, *104*(10), 3154–3185. https://doi.org/10.1257/aer.104.10.3154

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 [arXiv preprint arXiv:2508.01370]. *LLM4ECommerce Workshop, KDD 2025*. 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

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, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., & Zhou, D. (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., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022b). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica*, *64*(5), 1067–1084. https://doi.org/10.2307/2171956

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.

Wu, J. C., & Xia, F. D. (2016). Measuring the macroeconomic impact of monetary policy at the zero lower bound. *Journal of Money, Credit and Banking*, *48*(2–3), 253–291. https://doi.org/10.1111/jmcb.12300

Wu, Z., Bai, H., Zhang, A., Gu, J., Vydiswaran, V., Jaitly, N., & Zhang, Y. (2024). Divide-or-Conquer? Which Part Should You Distill Your LLM? *arXiv preprint arXiv:2402.15000*.

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

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

Zhang, A. L., Kraska, T., & Khattab, O. (2025). Recursive Language Models. *arXiv preprint arXiv:2512.24601*.

Zhu, K., Du, H., Hong, Z., Yang, X., Guo, S., Wang, Z., Wang, Z., Qian, C., Tang, X., 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

I present representative outputs from the six-agent pipeline (v30.5, `deepseek-v3.1:671b-cloud`). The pipeline processes four document types in a fixed sequence for each FOMC meeting t: the Statement from meeting $t-1$, the Chair's press conference from the same day (available from April 2011), the Minutes from meeting $t-1$ (released approximately three weeks later), and the Beige Book released approximately two weeks before meeting t. Each document feeds one extraction agent; the Expectation Engine then updates the probability distribution sequentially through stages $\mathcal{P}_1 \to \mathcal{P}_4$, with an optional $\mathcal{P}_5$ incorporating inter-meeting news. All examples use two contrasting episodes: December 2008 (financial crisis, ZLB transition) and March 2022 (tightening cycle liftoff).

A.1 Statement Summarizer

The March 2022 example illustrates a strong tightening signal with unconditional path language and an 8-1 vote. The Statement Summarizer extracts the guidance signal from each FOMC Statement: the policy direction, commitment strength (0 = none, 1 = conditional, 2 = unconditional), guidance horizon, conditionality structure, and any balance sheet language. The output seeds the $\mathcal{P}_1$ update in the Expectation Engine for the following meeting.

Statement Summarizer: March 16 2022 FOMC Statement


```
{
  "guidance_signal": {
    "meeting_date": "2022-03-16",
    "vote_tally": "8-1",
    "dissents": ["preferred to raise the target range by 0.5 percentage points"],
    "direction": "tightening",
    "commitment": 2,
    "horizon": "next_few",
    "conditionality": "explicit_data_dependent",
    "balance_sheet_guidance": "The Committee expects to begin reducing its holdings
      of Treasury securities and agency debt and agency MBS at a coming meeting."
  }
}
```


A.2 Policy Extractor (Minutes)

The Policy Extractor processes the FOMC Minutes to recover the Committee's internal deliberations: committee dynamics (dissent structure, coalition), mandate signals (inflation and labor market pressure scores from -1 to $+1$), forward guidance classification, and a calibrated policy

stance score anchored to historical episodes. The output informs the $\mathcal{P}_3$ update.

A.2.1 Example 1: Financial Crisis Period (December 2008)

The October 2008 Minutes informed expectations for the December 16, 2008 meeting. The Policy Extractor identifies unanimous agreement on a large cut, sharply subdued inflation (-0.70) and labor market conditions (-0.80), and an "outlook_conditional" forward guidance type reflecting the new ZLB language. The calibrated stance score of -0.90 is near the dovish anchor (March 2020 $= -1.0$).

Policy Extractor: October 2008 Minutes (informing December 2008 meeting)


```
{
  "decision_explanation": {
    "main_rationale": "The Committee established a target range of 0 to 1/4 percent
      due to significant economic downturn, deteriorating labor market conditions,
      diminished inflation pressures, and strained financial markets.",
    "key_drivers": ["inflation", "labor_market", "financial_conditions", "economic_activity"]
  },
  "committee_dynamics": {
    "num_dissents": 0,
    "dissent_balance": "none",
    "coalition_structure": "broad_consensus",
    "internal_debate_summary": "Unanimous agreement on near-zero target range;
      discussion centered on whether to announce an explicit rate target rather
      than on the direction of policy."
  },
  "forward_guidance": {
    "guidance_type": "outlook_conditional",
    "guidance_strength_score": 0.7,
    "explicit_guidance_text": "The Committee anticipates that weak economic conditions
      are likely to warrant exceptionally low federal funds rates for some time,
      conditional on economic outlook."
  },
  "mandate_signals": {
    "inflation_pressure": { "score": -0.7, "direction": "subdued", "confidence": "high" },
    "labor_market_tightness": { "score": -0.8, "direction": "loose", "confidence": "high" }
  },
  "narrative_diagnostics": {
    "policy_stance_score": {
      "score": -0.9,
      "nearest_historical_anchor_dovish": "September 2007 crisis response",
      "nearest_historical_anchor_hawkish": "January 2020 steady state"
    }
  }
}
```

A.2.2 Example 2: Tightening Cycle Liftoff (March 2022)

The January 2022 Minutes informed expectations for the March 16, 2022 meeting. The Policy Extractor identifies one dissent (Bullard, preferring a 50bp hike), elevated inflation (+0.80) and tight labor markets, and outlook-conditional guidance. The calibrated stance score of +0.70 is anchored to the early-2022 inflation response.

Policy Extractor: January 2022 Minutes (informing March 2022 meeting)


```
{
  "decision_explanation": {
    "main_rationale": "The Committee raised rates due to elevated inflation pressures,
      a very tight labor market, and the need to begin removing policy accommodation.",
    "key_drivers": ["inflation", "labor_market", "financial_conditions", "economic_activity"]
  },
  "committee_dynamics": {
    "num_dissents": 1,
    "dissent_balance": "more_hawkish",
    "coalition_structure": "broad_consensus",
    "internal_debate_summary": "Most participants agreed on a 25bp hike and near-term
      balance sheet runoff; one member (Bullard) preferred 50bp to address inflation
      more aggressively."
  },
  "forward_guidance": {
    "guidance_type": "outlook_conditional",
    "guidance_strength_score": 0.6,
    "explicit_guidance_text": "The Committee anticipates that ongoing increases in
      the target range will be appropriate and will monitor incoming information
      to adjust policy as needed."
  },
  "mandate_signals": {
    "inflation_pressure": { "score": 0.8, "direction": "elevated", "confidence": "high" },
    "labor_market_tightness": { "score": 0.7, "direction": "tight", "confidence": "high" }
  },
  "narrative_diagnostics": {
    "policy_stance_score": {
      "score": 0.7,
      "nearest_historical_anchor_dovish": "January 2020 steady state (0.0)",
      "nearest_historical_anchor_hawkish": "early 2022 inflation response (0.7)"
    }
  }
}
```


A.3 Beige Book Decoder

The Beige Book Decoder aggregates regional economic narratives into mandate-level signals. For each mandate dimension (inflation pressure, labor market tightness, and auxiliary topics), it derives a policy-direction probability $(p_{\text{tighten}}, p_{\text{neutral}}, p_{\text{ease}})$ from each district and aggregates using GDP weights. The resulting mandate signals feed the $\mathcal{P}_4$ update.

The March 2, 2022 Beige Book shows broadly elevated inflation with 82% probability of tightening and near-zero easing probability. The twelve districts contribute probabilities that are consistently hawkish on inflation, with only minor variation across regions.

Beige Book Decoder: March 2 2022 Beige Book


```
{
  "meta": { "reference_date": "2022-03-16", "districts_analyzed": 12 },
  "mandate_signals": {
    "inflation_pressure": {
      "narrative": "Across all districts, inflation pressures are broadly elevated
        and persistent, driven by widespread input cost increases, strong wage growth,
        and ongoing supply chain disruptions.",
      "policy_probabilities": { "tighten": 0.82, "neutral": 0.16, "ease": 0.02 },
      "district_contributions": {
        "NYC": { "p_tighten": 0.85, "p_ease": 0.02, "gdp_weight": 0.15 },
        "SFR": { "p_tighten": 0.85, "p_ease": 0.00, "gdp_weight": 0.14 },
        "CHI": { "p_tighten": 0.85, "p_ease": 0.02, "gdp_weight": 0.10 },
        "DAL": { "p_tighten": 0.85, "p_ease": 0.00, "gdp_weight": 0.09 }
      }
    },
    "labor_market_tightness": {
      "policy_probabilities": { "tighten": 0.79, "neutral": 0.19, "ease": 0.02 }
    }
  }
}
```


A.4 Press Conference Analyzer

The Press Conference Analyzer processes the Chair's Q&A transcripts (available from April 2011) to extract forward guidance signals, commitment strength, and the composition of guidance between outlook-based and commitment-based language. It provides the $\mathcal{P}_2$ update. In March 2022, Chair Powell delivered unconditional path language committing to ongoing increases, with an equal split between outlook-based and commitment-based guidance.

Press Conference Analyzer: March 16 2022 Press Conference


```
{
  "signal": {
    "forward_guidance_signals": {
      "rate_path": {
        "direction": "hawkish",
        "key_quotes": [
          "the Committee anticipates that ongoing increases in the target range
            for the federal funds rate will be appropriate",
          "if we conclude that it would be appropriate to move more quickly to
            remove accommodation, then we'll do so"
        ]
      },
      "commitment_strength": {
```

Table 17: Sequential filtration: probability distributions and expected rate changes

Stage	Document	Probability mass		$\hat{\mathbb{E}}[\Delta i]$
		Primary action	Hold	(bp)
December 16, 2008 (ZLB transition; no press conference available)				
$\mathcal{P}_1$	Statement, Oct. 2008	−25bp: 25%, −50bp: 70%	5%	−41.2
$\mathcal{P}_3$	Minutes, Oct. 2008	−25bp: 25%, −50bp: 73%	2%	−42.8
$\mathcal{P}_4$	Beige Book	−25bp: 24%, −50bp: 75%	1%	−43.5
Realized: −87.5bp (new range: 0–25bp)				
Surprise $\hat{s}_t$ (using $\mathcal{P}_4$)				−44.0
March 16, 2022 (tightening liftoff)				
$\mathcal{P}_1$	Statement, Jan. 2022	+25bp: 90%	10%	+22.5
$\mathcal{P}_2$	Press conference, Jan. 2022	+25bp: 90%	10%	+22.5
$\mathcal{P}_3$	Minutes, Jan. 2022	+25bp: 90%	10%	+22.5
$\mathcal{P}_4$	Beige Book	+25bp: 85%	15%	+21.2
Realized: +25bp				
Surprise $\hat{s}_t$ (using $\mathcal{P}_4$)				+3.8

Note: Each stage conditions on the documents available up to and including that point in the communication timeline. $\mathcal{P}_1$ is seeded by the prior FOMC Statement; $\mathcal{P}_2$ updates on the Chair's press conference (unavailable before April 2011, so $\mathcal{P}_2 = \mathcal{P}_1$ for December 2008); $\mathcal{P}_3$ on the prior Minutes; $\mathcal{P}_4$ on the Beige Book (final prior used for the main results); $\mathcal{P}_5$ on inter-meeting news (available from mid-2003; used for the $P_4 \to P_5$ comparison in Appendix C.2). Surprise $\hat{s}_t = \Delta i_t - \hat{\mathbb{E}}[\Delta i_t \mid \mathcal{P}_5]$. Probabilities are rounded to the nearest integer.


```
        "assessment": "unconditional",
        "score": 2,
        "justification": "The phrase 'anticipates that ongoing increases will be
          appropriate' is binding path language committing to future action."
      },
      "guidance_composition": {
        "outlook_based": 0.5,
        "commitment_based": 0.5,
        "explanation": "The Chair balanced forward-looking economic expectations
          with strong commitments to ongoing rate increases and balance sheet reduction."
      }
    }
  }
}
```


A.5 Sequential Bayesian Filtration

The Expectation Engine processes each document in sequence and updates the probability distribution over potential rate actions. Table 17 shows how the distribution evolves across stages for both example meetings. For December 2008, press conferences were not yet available, so the sequence runs $\mathcal{P}_1 \to \mathcal{P}_3 \to \mathcal{P}_4 \to \mathcal{P}_5$. For March 2022, all four documents are available and the full $\mathcal{P}_1 \to \mathcal{P}_4$ sequence runs before the optional $\mathcal{P}_5$ news update.

A.6 Surprise Quantification

The Surprise Quantification agent computes $\hat{s}_t = \Delta i_t - \hat{\mathbb{E}}[\Delta i_t \mid \mathcal{P}_5]$ and records the surprise direction, a contextual surprise score (0–1), and a narrative justification. The December 2008 meeting produces a large dovish surprise (−46.2bp): the Fed cut to a 0–25bp range when the prior concentrated 65% probability on a 50bp cut and 35% on a 25bp cut, far short of the 75bp effective cut to the range midpoint. The March 2022 meeting produces a small hawkish surprise (+6.2bp): a 25bp hike was well-anticipated, with the prior assigning 75% probability to exactly that outcome.

Surprise Quantification: December 16 2008


```
{
  "meeting_date": "2008-12-16",
  "expected_rate_change": -0.4125,
  "realized_rate_change": -0.875,
  "surprise_rate": -0.4625,
  "surprise_score": 0.85,
  "surprise_direction": "dovish",
  "justification": "The realized rate change of -0.875% versus an expected -0.4125%
    yields a surprise of -0.4625%. The prior concentrated 65% probability on a 50bp
    cut and 35% on a 25bp cut; the 75bp effective cut to the near-zero range midpoint
    was a strong dovish surprise, consistent with the unprecedented ZLB transition."
}
```


Surprise Quantification: March 16 2022


```
{
  "meeting_date": "2022-03-16",
  "expected_rate_change": 0.1875,
  "realized_rate_change": 0.25,
  "surprise_rate": 0.0625,
  "surprise_score": 0.5,
  "surprise_direction": "hawkish",
  "justification": "The Fed raised rates by 25bp versus an expected 18.75bp,
    resulting in a 6.25bp hawkish surprise. The prior assigned 75% probability
    to exactly a 25bp hike, so this is a modest surprise: the direction was
    correctly anticipated but the magnitude slightly exceeded expectations."
}
```

B Pipeline Validation

This appendix validates the pipeline along three targets: internal integrity of the measurement instrument (Section B.1), external construct validity against market and survey expectations (Section B.2), and architectural validation of the one decoder whose design involves a non-trivial choice (Section B.3).

B.1 Internal Integrity Checks

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. Both are addressed below.

B.1.1 Look-Ahead Bias: Design and Empirical Test

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). Glasserman and Lin (2024) document analogous indirect leakage: references to "pandemic," "disease outbreak," or "supply chain" were 3.6 times more common in LLM-generated 2020 risk assessments than 2019 ones. The pipeline addresses this threat through both architectural design and an empirical test.

Architectural controls. Four complementary safeguards constrain what each agent can see and how it is asked to reason. *(i) Document-level predetermined cutoffs.* Each agent processes only publicly available information with strict temporal ordering: the Beige Book Decoder analyzes Beige Books released at least two weeks before each FOMC meeting; the Policy Extractor processes Minutes from the *previous* meeting, ensuring no overlap with the current decision; the Expectation Engine constructs probabilistic expectations using only information available before the blackout period begins. *(ii) 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..."), following the recommendation of Sarkar and Vafa (2024). *(iii) 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. *(iv) Out-of-knowledge-cutoff validation.* The sample extends beyond DeepSeek-v3.1's training cutoff (July 2024); meetings after this date were not in the model's training data, enabling a direct comparison across the boundary.

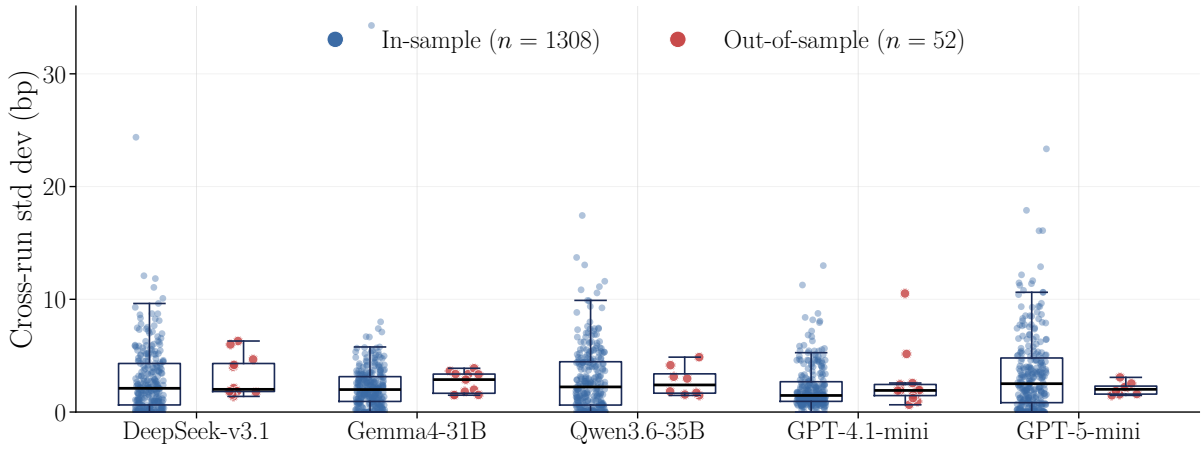
These controls reduce but cannot eliminate look-ahead bias. Sarkar and Vafa (2024) show that LLMs can infer censored temporal information even from indirect cues (correlation of 0.79 between predicted and actual years from date-censored earnings calls), and this study provides explicit temporal context rather than removing it. The empirical test below is therefore the load-bearing diagnostic.

Empirical test. Under the memorization hypothesis, meetings beyond a model's training cutoff should produce *higher* cross-run dispersion, since the model can no longer fall back on memorized outcomes and must reason stochastically from the documents alone. Figure 15 reports the test for five model families spanning three open-weight architectures (DeepSeek-v3.1 671B, Gemma4-31B, Qwen3.6-35B) and two proprietary models (GPT-4.1-mini, GPT-5-mini). Each family is split at its own training cutoff (July 2024 to January 2025) and the cross-run standard deviation is computed per meeting from independent pipeline executions at temperature 0. Across all five families the in-sample and out-of-sample distributions overlap heavily and the medians sit on top of each other; none of the five t-tests is significant. Table 18 reports the DeepSeek-v3.1 summary as a baseline: 259 in-sample meetings have median cross-run standard deviations of 2.12 bp (expected rate change) and 2.05 bp (surprise rate); 13 out-of-sample meetings show 1.90 bp and 1.82 bp respectively, with neither difference rejecting at any conventional level. Out-of-sample dispersion is therefore not elevated relative to in-sample, rejecting the look-ahead hypothesis.

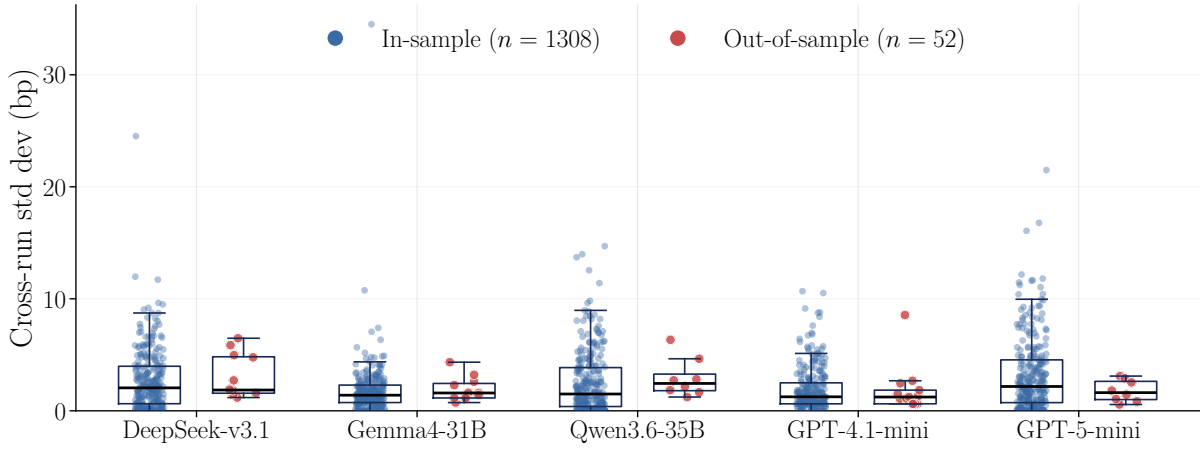
B.1.2 Cross-Model Consistency

The multi-run stability exercise quantifies within-model stochasticity. A complementary question is whether the results are specific to the DeepSeek-v3.1 (671B) model used in the main analysis, or whether independently trained models converge to similar surprises when reading the same documents.

Figure 15: Look-ahead bias test: in-sample vs. out-of-sample cross-run stability across model families



(a) Expected Rate Change



(b) Surprise Rate

Note: Cross-run standard deviation per FOMC meeting, restricted to meetings with at least 4 independent runs per model family. Color encodes model family; marker shape encodes in-sample versus out-of-sample relative to each model's training cutoff (vertical dashed lines, color-matched). Distributional summaries are reported in Table 18 for the deepseek baseline. None of the five model families shows out-of-sample dispersion significantly elevated relative to in-sample.

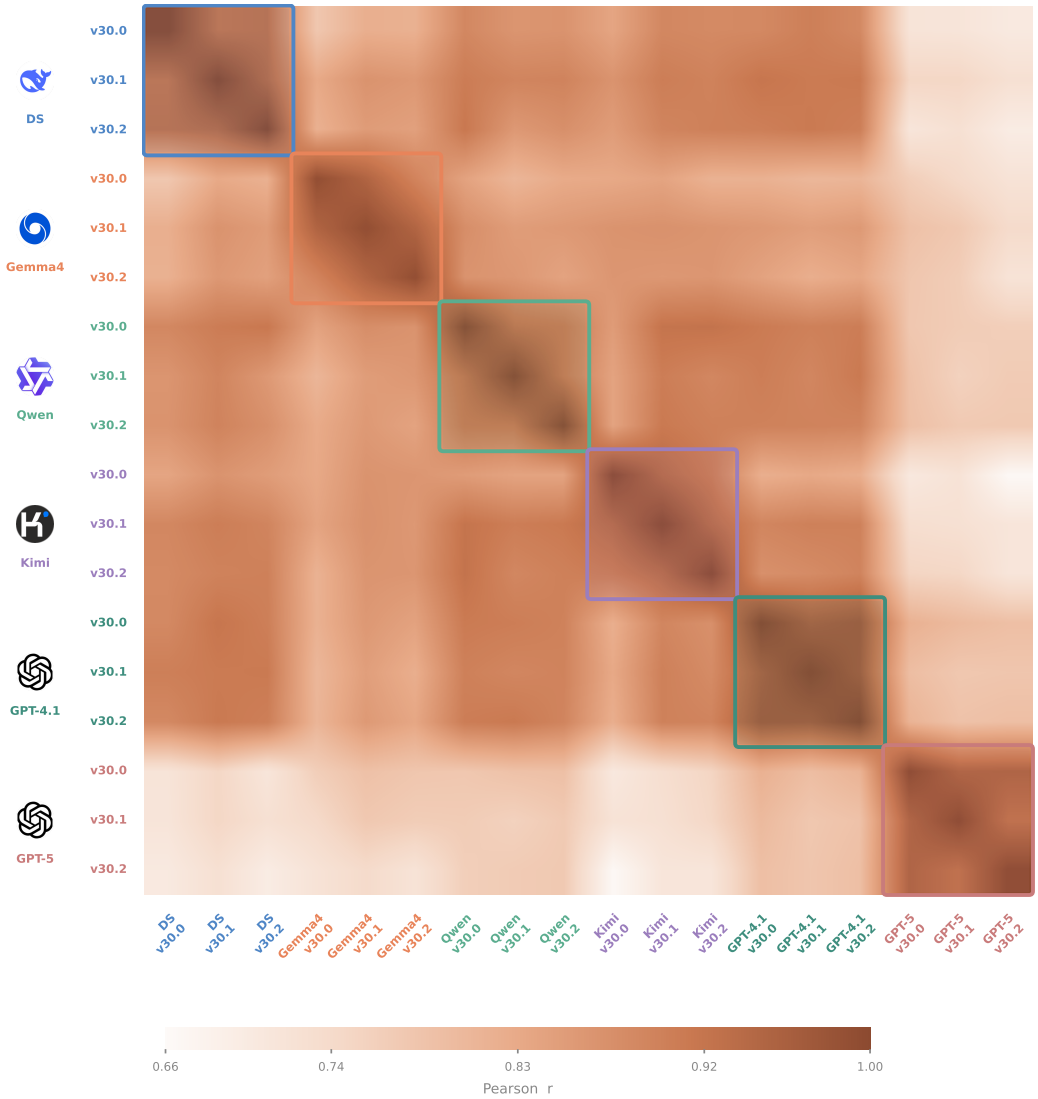
Table 18: Cross-run stability: in-sample vs. out-of-sample meetings

	In-sample ($N = 259$)		Out-of-sample ($N = 13$)		Test	
Measure	Mean	Median	Mean	Median	t-stat	p-value
Expected rate change (bp)	2.83	2.12	2.96	1.90	−0.25	0.81
Surprise rate (bp)	2.75	2.05	2.86	1.82	−0.19	0.85

Note: Cross-run standard deviations (in basis points) across 6 independent pipeline executions, restricted to meetings with ≥ 4 runs. In-sample: 259 meetings before July 2024 training cutoff. Out-of-sample: 13 meetings after July 2024 (beyond training cutoff). p-values from two-sided Welch t-test. Neither difference is significant at conventional levels.

Figure 16 reports the full 9×9 Pearson correlation matrix of surprise series across three model families (DeepSeek-v3.1 671B, Gemma4-31B, Qwen3.6-35B) and three pipeline versions

Figure 16: Cross-model and cross-version surprise correlations



Note: Pearson correlations of monetary policy surprise series ($\hat{s}_t$, in basis points) across 18 independent pipeline runs: DeepSeek-v3.1, Gemma4-31B, Qwen3.6-35B, Kimi-K2.6, GPT-4.1-mini, and GPT-5-mini, each run at three pipeline versions (v30.0–v30.2). Rounded borders delineate the six 3×3 within-family blocks; color scale runs from 0.66 (cream, sample minimum) to 1.0 (dark sienna). Within-family averages: DeepSeek 0.927, Gemma4 0.934, Qwen 0.915, Kimi 0.927, GPT-4.1-mini 0.961, GPT-5-mini 0.943. Cross-family averages range from 0.72 (GPT-5-mini paired with DeepSeek or Kimi) to 0.90 (DeepSeek×GPT-4.1-mini); all 15 cross-family pairs lie in $[0.72, 0.90]$. Pairwise correlations use all available common meetings; Kimi runs are restricted to 235–249 meetings (subset of the full 272-meeting sample) due to API quota limits at processing time.

(v30.0–v30.2), yielding nine independent runs in total. The matrix has a natural block structure: diagonal 3×3 blocks capture within-family stability; off-diagonal blocks capture cross-family agreement.

Within-family correlations are uniformly high, ranging from 0.915 (Qwen) to 0.961 (GPT-4.1-mini), confirming the same high stability documented in Section B.1.1. Cross-family correlations are systematically lower but remain substantial: the 15 cross-family pair-averages span

$[0.72, 0.90]$, with the strongest agreement among the open-weight families (DeepSeek, Gemma4, Qwen, Kimi: 0.83–0.89) and slightly weaker agreement when GPT-5-mini is involved (0.72–0.79). The within-to-cross gap is 0.04–0.20, comparable to or modestly larger than the gap between adjacent and non-adjacent pipeline versions within a single family.

This pattern has a direct bearing on the memorization concern. The six families are trained by different organizations (DeepSeek, Google, Alibaba, Moonshot, OpenAI) on different corpora and have different knowledge cutoffs. If the pipeline's accuracy stemmed from models recalling memorized FOMC outcomes, cross-family correlations would either be near-perfect (all models recalling identical outcomes) or near-zero (models memorizing conflicting information). Instead, cross-family agreement lies below within-family stability but far above zero, consistent with all six models performing the same document-reading task and disagreeing primarily on ambiguous passages rather than on the direction or magnitude of well-identified surprises.

B.1.3 Idiosyncratic variance: information or noise?

A high but imperfect cross-family correlation can reflect two very different things. The slice of variance one family does not share with the others may be genuine information about FOMC documents that the family recovers and the rest of the panel misses, or it may be model-specific miscalibration noise. The first case is also where look-ahead leakage would manifest: a family that has partially memorized realized rate decisions during training would show the same statistical fingerprint as a family that simply reads the documents better. The exercises below quantify how big this idiosyncratic slice is, whether it predicts what markets do, and whether its structure is compatible with memorization.

The starting point is a leave-one-out projection of each family's surprise on the consensus of the rest,

$$\hat{s}_{i,t} \;=\; \alpha_i \;+\; \beta_i\, \bar{s}_{-i,t} \;+\; \hat{\varepsilon}_{i,t}, \qquad \bar{s}_{-i,t} \;=\; \frac{1}{N-1} \sum_{j \neq i} \hat{s}_{j,t}, \tag{22}$$

where $\bar{s}_{-i,t}$ is the leave-one-out mean across the other five families. I exclude Kimi-K2.6 so that the common sample retains the full $T = 272$ FOMC meetings between 1996 and 2026, including the post-2022 tightening and easing cycle that drives most of the variance in the surprise series. The constant α_i absorbs any level differences across families, including the modest hawkish bias of the OpenAI variants ($\bar{\hat{s}}_i \approx -4$ basis points for GPT-4.1-mini and GPT-5-mini, against ≈ -0.8

Table 19: Cross-model variance decomposition and yield-information test

	DeepSeek-v3.1	Gemma4-31B	Qwen3.6-35B	GPT-4.1-mini	GPT-5-mini
$\hat{\sigma}(\hat{s}_i)$ (bp)	14.51	13.78	14.85	15.46	19.76
$\hat{\sigma}(\hat{\varepsilon}_i)$ (bp)	5.91	6.19	6.27	6.13	11.81
Residual share	0.166	0.202	0.179	0.157	0.357
non-OpenAI consensus	0.163	0.219	0.185	0.147	0.369
ΔR^2 on Δy_{3m}	0.0056	0.0176	0.0080	0.0376*	0.0150

Notes: Estimates of equations (22) and (23) on $T = 272$ FOMC meetings (1996–2026), pipeline version v30.1. Rows report the standard deviation of family i's surprise series, the standard deviation of its residual from (22), the residual variance share $\hat{\sigma}^2(\hat{\varepsilon}_i)/\hat{\sigma}^2(\hat{s}_i)$ under the baseline pool, the same share when the consensus is built only from the four non-OpenAI families, and the incremental R^2 from adding $\hat{\varepsilon}_{i,t}$ to the consensus-only specification of (23) at the 3-month maturity. HC3 heteroskedasticity-robust standard errors. Stars on ΔR^2 denote rejection of $H_0 : \gamma_{i,3m} = 0$ at $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

for the other four families).[22] I then test whether the residual $\hat{\varepsilon}_{i,t}$ has incremental predictive content for FOMC-day yield changes,

$$\Delta y_{m,t} \;=\; \delta_{i,m} \;+\; \theta_{i,m}\, \bar{s}_{-i,t} \;+\; \gamma_{i,m}\, \hat{\varepsilon}_{i,t} \;+\; u_{i,m,t}, \tag{23}$$

where $\Delta y_{m,t}$ is the one-day change, on the FOMC announcement date, in the daily Treasury constant-maturity yield at maturity m from the Federal Reserve Economic Database. This is a market-validation test, not a truth test: a coefficient $\hat{\gamma}_{i,m}$ statistically distinct from zero with a positive incremental R^2 says the residual carries content markets price; a coefficient indistinguishable from zero says it is uncorrelated with FOMC-day market reactions, which I treat as a working definition of noise.

Table 19 delivers a clean first reading. Five of the six families have residual shares between 0.16 and 0.20: roughly four-fifths of each family's surprise variance is already explained by the others' consensus, and the result is essentially unchanged when OpenAI models are removed from the consensus pool. GPT-5-mini stands apart, with a residual share of 0.36 and a total surprise standard deviation about thirty percent wider than the others. The yield-information test in the last row takes the question of whether that dispersion is signal or noise to the data: across all six families, the incremental R^2 on the 3-month yield change is small ($\Delta R^2 \leq 0.04$), and only GPT-4.1-mini reaches even marginal significance under HC3 standard errors. GPT-5-mini's wider residual produces no detectable incremental content ($\hat{\gamma}$ insignificant, $\Delta R^2 \approx 0.015$), which

[22] To check that the projection is not silently absorbing OpenAI-family clustering into the consensus, I reconstruct $\bar{s}_{-i,t}$ from the four non-OpenAI families only; the resulting residual variance shares are reported alongside the baseline in Table 19 and move by less than one percentage point in every case.

Table 20: Direct leakage diagnostic: residual on realized rate change

	DeepSeek-v3.1	Gemma4-31B	Qwen3.6-35B	GPT-4.1-mini	GPT-5-mini
$\hat{\beta}$ on realized	0.060***	-0.099***	0.089***	0.108***	-0.042
	(0.015)	(0.019)	(0.025)	(0.026)	(0.037)
R^2	0.051	0.123	0.098	0.152	0.006
T	272	272	272	272	272

Notes: Univariate regressions of each family's residual $\hat{\varepsilon}_{i,t}$ from (22) on the realized rate change at t, with HC3 standard errors. A residual that reflects pure miscalibration noise should be uncorrelated with the realized outcome; a residual that reflects memorization should load on it with the same sign across families and with magnitudes that increase in training-data exposure. Stars: $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

on this metric is the noise interpretation; the other five families show at most limited evidence of incremental signal at the short end of the curve.

The yield-information test can rule the residuals in as informative for what markets price, but it does not separate genuine reading skill from partial leakage of training-set FOMC outcomes. To probe the leakage channel, I run a direct diagnostic: regress each family's residual on the realized rate change at t.[23] The results are in Table 20. Five of the six families show small but precisely estimated loadings (R^2 between 0.05 and 0.15); only GPT-5-mini's residual is uncorrelated with the realized outcome. The signs of the significant loadings are decisive. DeepSeek-v3.1, Qwen3.6-35B, and GPT-4.1-mini all have positive loadings, which (by the identity in the previous footnote) implies their priors are *less* accurate than the consensus and rules out memorization for those three families. Gemma4-31B has a negative loading, indicating a prior that is more accurate than the consensus average; this is the one family whose loading is consistent with either better reading or memorization, but Gemma4's mid-2024 training cutoff and broad multilingual training mix give no special prior reason to expect FOMC-specific memorization. GPT-5-mini's loading is statistically zero. The two model variants with the latest cutoffs and the heaviest English-press exposure, where leakage would be most plausible *a priori*, therefore land on opposite sides of the test: GPT-5-mini is uncorrelated with the realized rate, and GPT-4.1-mini's positive loading actively rules out memorization. The pattern lines up with heterogeneous prior accuracy across families, not with a memorization gradient.

[23] Working out the projection in (22) gives the exact identity $\hat{\beta}_{\hat{\varepsilon}_i,r} = (1-\beta_i) - \rho_i + \beta_i\,\rho_{-i} = (\rho_{-i} - \rho_i) + (1-\beta_i)(1-\rho_{-i})$, where $\rho_i = \mathrm{cov}(\mathrm{prior}_i, r_t)/\mathrm{var}(r_t)$ measures how predictive family i's prior is of the realized rate change. The first term is the differential prior accuracy of i versus the consensus; the second is a scale-mismatch term that vanishes when $\beta_i = 1$ or when the consensus is a perfect linear predictor of r_t. In the data $\beta_i \in [0.82, 1.18]$ and $\rho_{-i} \approx 0.43$, so both terms can contribute. The substantive direction nevertheless follows from inverting the identity for each family. Memorization can only push ρ_i upward (a memorizing prior is more accurate). A loading $\hat{\beta}_{\hat{\varepsilon}_i,r}$ that is positive therefore implies $\rho_i < \rho_{-i} + (1-\beta_i)(1-\rho_{-i})/\beta_i$, that is, family i's prior is *less* accurate than a regime-adjusted version of the consensus, which rules out memorization for that family. A negative loading is consistent with either better reading or memorization.

Table 21: Pooled training-cutoff interaction with placebo benchmark

	DeepSeek-v3.1	Gemma4-31B	Qwen3.6-35B	GPT-4.1-mini	GPT-5-mini
$\hat{\gamma}_{\text{pre}}$	-0.048	-0.116	0.014	0.096	0.054
	(0.075)	(0.077)	(0.085)	(0.094)	(0.048)
$\hat{\zeta}$ (interaction)	0.272	0.027	0.130	0.073	-0.124
	(0.195)	(0.254)	(0.244)	(0.190)	(0.149)
Placebo max $\lvert\hat{\zeta}\rvert$	0.133	0.315	0.228	0.188	0.150
Cutoff date	2024-07-01	2024-08-01	2024-09-01	2024-06-01	2024-09-30
T_{pre}	259	260	260	258	261
T_{post}	13	12	12	14	11

Notes: Estimates of equation (24) for each family. $\hat{\gamma}_{\text{pre}}$ is the residual coefficient before the family's training cutoff and $\hat{\zeta}$ is the change in that coefficient after the cutoff. Pure leakage would produce a negative $\hat{\zeta}$ statistically distinct from zero. The placebo row reports the largest absolute value of $\hat{\zeta}$ obtained when the cutoff is replaced with one of four pre-2024 placebo dates (2008, 2012, 2016, 2020); for every family the actual $\lvert\hat{\zeta}\rvert$ is smaller than the placebo maximum. HC3 standard errors; stars at $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

A complementary test exploits the training-cutoff structure directly. For each family I pool pre- and post-cutoff observations and estimate

$$\Delta y_{3m,t} \;=\; \delta + \theta\,\bar{s}_{-i,t} + \gamma\,\hat{\varepsilon}_{i,t} + \rho\,\mathbf{1}\{t \geq c_i\} + \zeta\,\hat{\varepsilon}_{i,t}\cdot\mathbf{1}\{t \geq c_i\} + u_t, \tag{24}$$

where c_i is family i's published training cutoff and $\hat{\zeta}$ is the change in residual informativeness after the cutoff. Pure leakage would produce $\hat{\zeta} < 0$, statistically distinct from zero: information that markets price should be present in the residual when the model could have memorized it and absent when it could not. To benchmark the test against ordinary sample-period heterogeneity, I re-estimate (24) at four placebo cutoffs (2008, 2012, 2016, 2020), all of which pre-date every family's training cutoff and so are null by construction. Table 21 reports the result. None of the five actual interactions is statistically distinct from zero, and in every case the actual $\lvert\hat{\zeta}\rvert$ is smaller than the largest placebo interaction obtained on the same family. The test thus sees no leakage signature, while the placebo distribution warns that interactions of the magnitude one would need to detect can be generated by ordinary heterogeneity over time without any role for training data.

A final sanity check bounds how much memorization the residual-on-realized loadings could in principle accommodate, even setting aside the sign argument that already rules it out for three families. For each family, the share of total surprise variance that could be attributed to a linear leakage channel is at most the residual's variance share times its R^2 on the realized rate. Per-family upper bounds are 0.85 percent (DeepSeek-v3.1), 2.48 percent (Gemma4-31B),

1.76 percent (Qwen3.6-35B), 2.38 percent (GPT-4.1-mini), and 0.22 percent (GPT-5-mini); the average is 1.5 percent and no family reaches three percent. The tightest bound, notably, sits on the family with the widest residual variance: GPT-5-mini's wide dispersion does not buy it any room to be hiding leaked information, because its residual is uncorrelated with realized rate changes. Taken together, the four exercises bound the idiosyncratic slice of variance, find that slice not informative for short-end yield reactions, find that the sign and recency pattern across families is incompatible with a memorization gradient, find no leakage signature at the actual training cutoffs once those are benchmarked against placebos, and place a hard upper bound below three percent on the share of total variance that any family's residual could attribute to leakage even under maximally adversarial assumptions. The cross-family disagreement is mostly noise around a shared signal, with no detectable memorization signature in the data.

Main econometric results use the full sample. Narrative-based results are compared against market-based benchmarks in Section 5.

B.2 External Construct Validation

B.2.1 Real-Time Information Set and Transparency Regimes

The sequential filtration pipeline uses exactly one document of each type per meeting: the most recently *released* statement, press conference, minutes, and Beige Book strictly before the meeting date. For statements and press conferences this is unambiguous, since both are published on the day of the previous meeting. For minutes, it is not: prior to February 2005 the FOMC released minutes with a lag of approximately 50 days, meaning that the minutes of meeting $M-1$ typically arrived one to three days *after* meeting M. Including those minutes in $\mathcal{P}_3$ for meeting M would thus incorporate information that was unavailable to market participants.

I correct for this by maintaining a real-time release calendar (`fomc_calendar.csv`) that records the actual publication date of every document. At stage $\mathcal{P}_3$, the pipeline queries this calendar for the most recently released minutes strictly before each meeting date. For all meetings from January 1996 through November 2004 (76 meetings in total), this returns the minutes of meeting $M-2$ rather than $M-1$; from February 2005 onward, when the Fed shortened the release lag to approximately 21 days, the pipeline correctly uses $M-1$ minutes throughout. The correction has no effect on any meeting after November 2004.

Direction of the bias. The uncorrected pipeline gave the LLM *more* information pre-2005 than was actually available, by passing it minutes that had not yet been published. Correcting this makes the prior $\mathcal{P}_3$ less informed in the early sample, widening the distribution and increasing surprise magnitudes. This is conservative relative to the main results: the corrected early-era surprises are noisier, not cleaner. The finding that identification quality improves with Fed transparency is therefore not an artefact of the correction but survives it.

B.2.2 Validation Against Market Expectations (Fed Funds Futures)

I first validate the LLM's pre-meeting expectation against the contemporaneous market expectation embedded in Fed Funds Futures. Using the high-frequency surprise series of Jarociński and Karadi (2020), I back out the market-implied expected rate change as the realized decision minus the front-month futures (FF1) surprise:

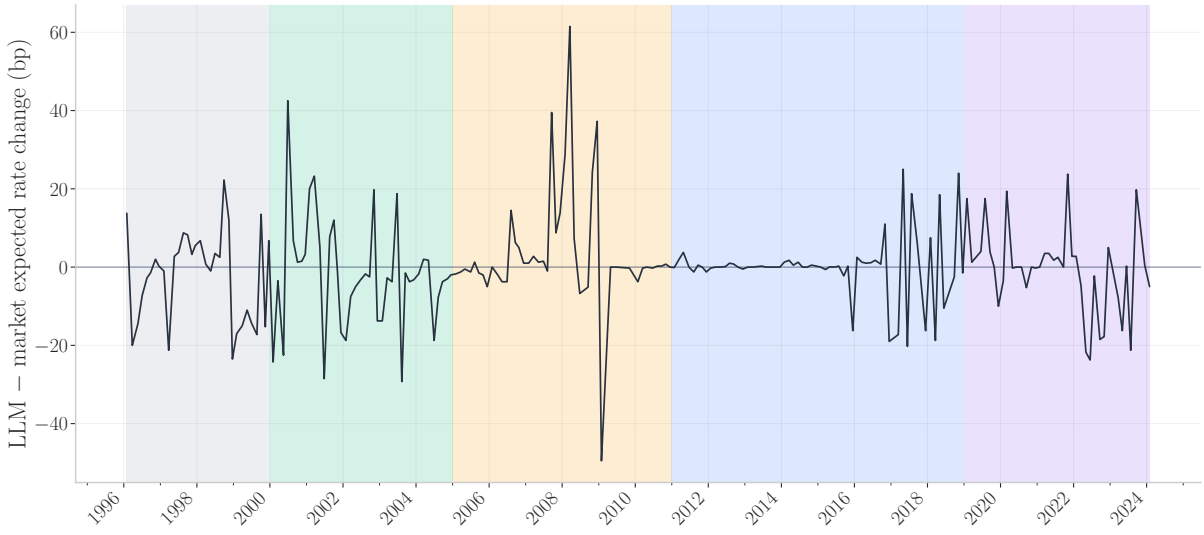
$$\mathbb{E}^{\text{market}}[\Delta i_t] \;=\; \Delta i_t - s_t^{\text{FF1}}. \tag{25}$$

This benchmark is independent of the LLM and of survey measurement: it is a price-based expectation, sampled in the announcement window, for the same policy decision that the LLM forecasts from public Fed documents.

Figures 17 and 18 show close agreement between the document-based LLM expectation and the market-implied expectation. The full-sample correlation is $r = 0.80$ across 214 meetings, and the strongest validation comes from the post-2019 regime, where every-meeting press conferences make the public information set richest and the correlation rises to $r = 0.92$. The 2011–18 correlation is lower ($r = 0.45$), but this period is mechanically compressed by the zero lower bound: when the funds rate is pinned near zero, both expected rate changes have little variance, so small idiosyncratic differences dominate the correlation.

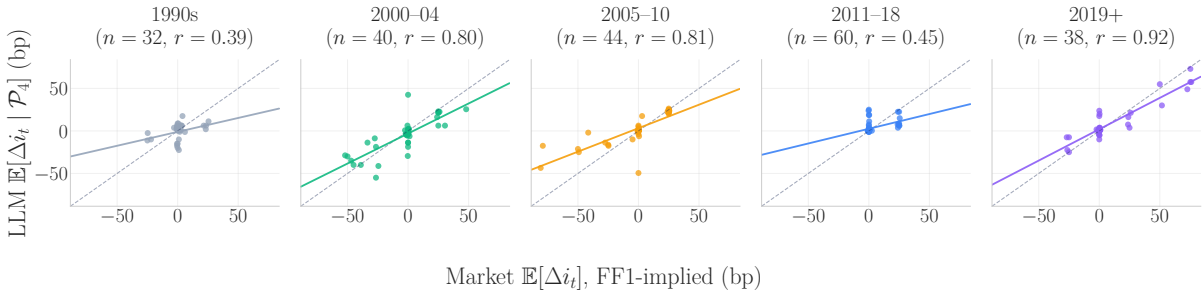
This price-based comparison establishes that the LLM is not producing arbitrary numerical forecasts. It does not, however, test the main object of interest: whether the residual surprise is larger when the policy decision is genuinely harder to infer from public information. For that, I turn from market prices to the cross-section of professional forecasts.

Figure 17: Gap between LLM and market expected rate change, FOMC meetings 1996–2024



Note: Difference between the LLM's documents-only expectation $\mathbb{E}[\Delta i_t \mid \mathcal{P}_4]$ and the market expectation $\mathbb{E}^{\text{market}}[\Delta i_t] = \Delta i_t - s_t^{\text{FF1}}$ backed out from Jarociński and Karadi (2020) front-month Fed Funds Futures surprises, in basis points. Background bands mark Fed transparency regimes: *grey* (1990s, opaque), *green* (2000–04), *amber* (2005–10, fast minutes), *blue* (2011–18, selected press conferences), and *purple* (2019+, every-meeting press conferences). Full sample $N = 214$.

Figure 18: LLM vs. market expected rate change, scatter by transparency regime



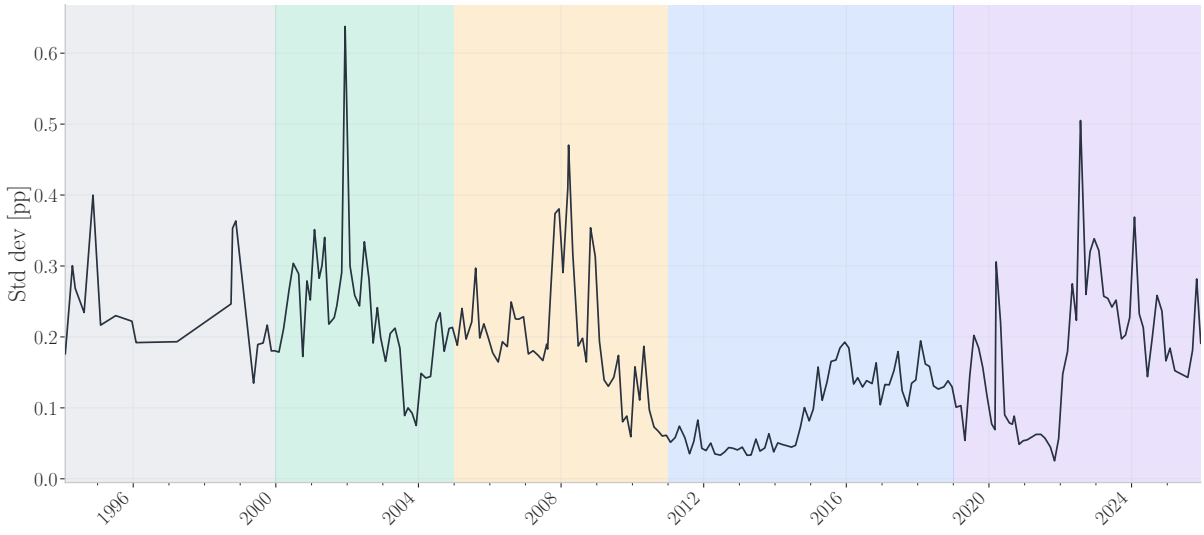
Note: Each point is one FOMC meeting. Horizontal axis: market expectation in basis points. Vertical axis: LLM expectation in basis points. Colours denote transparency regime. The dashed line is the 45-degree reference; the solid line is an OLS fit. Pearson correlations by regime: full sample $r = 0.80$ ($N = 214$); 1990s $r = 0.39$ ($N = 32$); 2000–04 $r = 0.80$ ($N = 40$); 2005–10 $r = 0.81$ ($N = 44$); 2011–18 $r = 0.45$ ($N = 60$); 2019+ $r = 0.92$ ($N = 38$).

B.2.3 Forecaster Disagreement and Policy Uncertainty

Consensus Economics (CE) provides a complementary validation source. The survey has polled approximately 30 professional forecasters on US macroeconomic outcomes since 1990, including the expected 3-month interest rate. I use the cross-forecaster standard deviation of that forecast, aligned to the next FOMC meeting, as an external measure of how difficult the near-term policy path was to predict from public information.

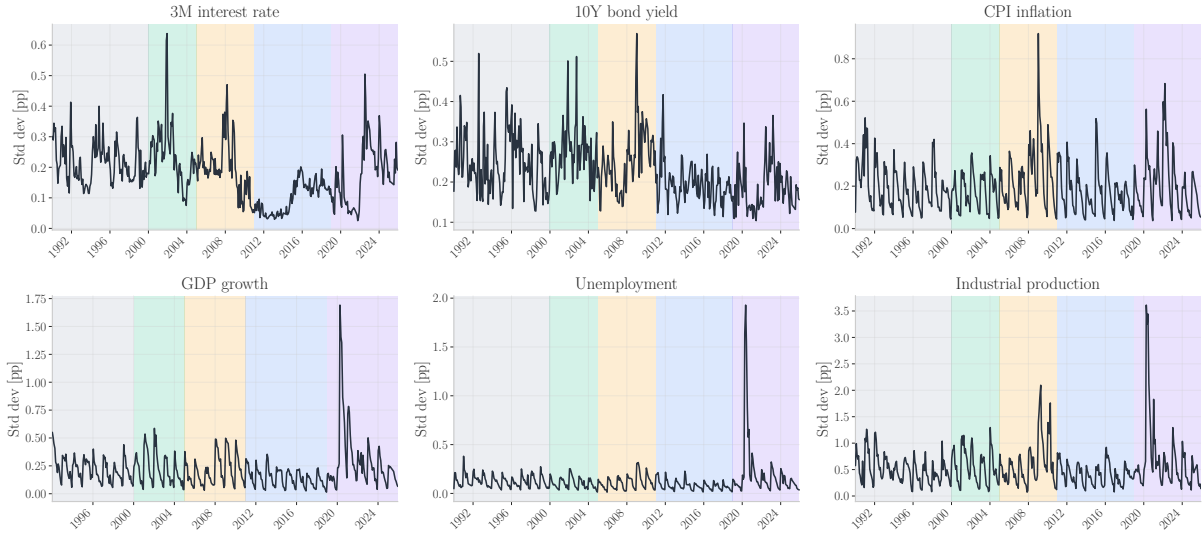
Figures 19 and 20 show that raw forecast disagreement is strongly cyclical. Rate disagree-

Figure 19: Cross-forecaster disagreement on the 3-month interest rate, 1994–2025



Note: Cross-forecaster standard deviation of the 3-month interest rate forecast from the Consensus Economics USA monthly survey, aligned to the next FOMC meeting. Background bands mark the same Fed transparency regimes used in Figure 17: *grey* (1990s, opaque), *green* (2000–04), *amber* (2005–10, fast minutes), *blue* (2011–18, selected press conferences), and *purple* (2019+, every-meeting press conferences).

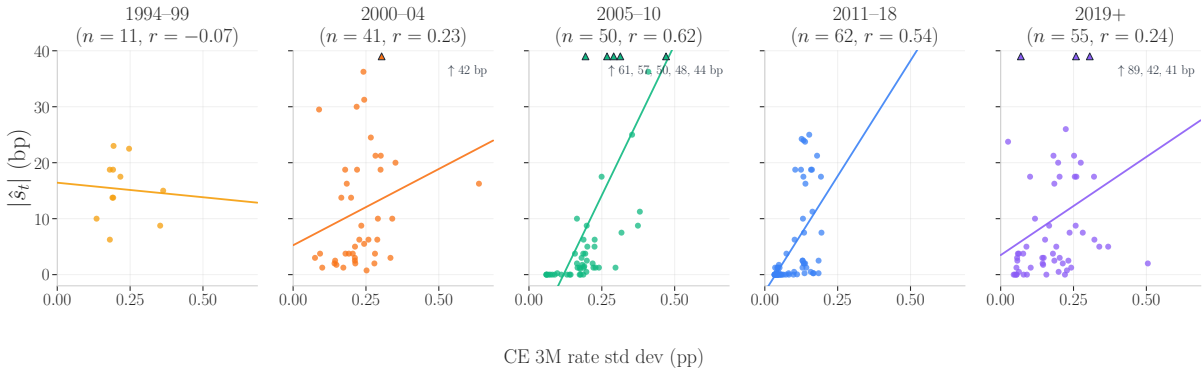
Figure 20: Cross-forecaster disagreement across mandate variables, 1994–2025



Note: Cross-forecaster standard deviation, in percentage points, for six variables surveyed by Consensus Economics: the 3-month interest rate, the 10-year Treasury yield, CPI inflation, GDP growth, the unemployment rate, and industrial production. Each panel shows the monthly raw standard deviation indexed by survey date. Background bands mark the same Fed transparency regimes used in Figure 17 (grey, green, amber, blue, purple from the 1990s to the post-2019 era).

ment spikes around the 2001 easing cycle and the 2007–2008 crisis, then compresses during the zero-lower-bound years. The same pattern appears across CPI inflation, unemployment, GDP growth, industrial production, and long rates, indicating that the CE series captures broad

Figure 21: Forecaster disagreement and LLM surprise magnitude



Note: Each point is one FOMC meeting. Horizontal axis: CE cross-forecaster standard deviation of the 3-month interest rate forecast (pp). Vertical axis: absolute LLM surprise $|\hat{s}_t|$ (bp). One panel per Fed transparency regime; per-panel Pearson r shown in titles. Pooled relationship is positive and significant ($r = 0.43$, $p < 0.001$, $N = 221$).

macroeconomic uncertainty rather than a narrow measurement artifact in the 3-month rate forecast.

The first link to the LLM measure is descriptive. Figure 21 plots CE disagreement against the absolute LLM surprise $|\hat{s}_t|$. Meetings with greater professional disagreement also tend to produce larger narrative surprises, with a pooled correlation of $r = 0.43$ over 221 meetings, significant at the 1% level. This relationship is not causal in either direction: forecasters do not observe the LLM surprise, and the LLM does not observe the survey cross-section. Both variables respond to the same underlying policy uncertainty.

The sharper test is predictive. I estimate whether a large LLM surprise at meeting $t-1$ predicts higher professional disagreement before meeting t, while controlling for transparency regimes, financial volatility, and recessions:

$$\sigma_t = \alpha + \sum_k \beta_k \, \mathbb{1}_{t \in \mathcal{R}_k} + \gamma \, \mathrm{VIX}_t + \delta \, \mathrm{Rec}_t + \zeta \, |\hat{s}_{t-1}| + \varepsilon_t, \tag{26}$$

where σ_t is the CE cross-forecaster standard deviation of the 3-month rate forecast at meeting t, and $\{\mathcal{R}_k\}$ are transparency regime intervals (2000–04, 2005–10, 2011–18, 2019+) defined by FOMC communication breakpoints, with 1994–99 as the omitted baseline. Controls are added progressively across columns in Table 22.

Column (4) contains the main cross-validation result. The lagged absolute LLM surprise enters with a coefficient of 0.349, significant at the 1% level, and raises R^2 from 0.316 to 0.426 relative to the specification with regimes, VIX, and recessions. A measure constructed only from

Table 22: CE cross-forecaster disagreement and transparency regimes

	(1) Regimes	(2) + VIX	(3) + Recession	(4) + $\lvert\hat{s}_{t-1}\rvert$
2000–04	−0.005	−0.007	−0.015	0.003
	(0.029)	(0.030)	(0.028)	(0.037)
2005–10	−0.036	−0.038	−0.056**	−0.012
	(0.029)	(0.029)	(0.024)	(0.033)
2011–18	−0.138***	−0.135***	−0.140***	−0.101***
	(0.021)	(0.022)	(0.020)	(0.030)
2019+	−0.063**	−0.064**	−0.065**	−0.035
	(0.030)	(0.031)	(0.030)	(0.037)
VIX		0.001	−0.000	−0.000
		(0.001)	(0.001)	(0.001)
Recession			0.079**	0.027
			(0.034)	(0.029)
$\lvert\hat{s}_{t-1}\rvert$				0.349***
				(0.057)
Intercept	0.237***	0.220***	0.247***	0.184***
	(0.013)	(0.021)	(0.022)	(0.034)
R^2	0.276	0.281	0.316	0.426
N	229	229	229	218

Note: OLS with Newey-West HAC standard errors (6 lags) in parentheses; dependent variable is the CE cross-forecaster standard deviation of the 3-month interest rate forecast (pp). Regime coefficients are deviations from the 1994–99 baseline (the intercept reports the baseline mean). VIX (CBOE Volatility Index, monthly mean) is statistically indistinguishable from zero across specifications, indicating the regime effect is not a residual macro-uncertainty story; the 2008 recession indicator enters positively but loses significance once the lagged surprise is included. $\lvert\hat{s}_{t-1}\rvert$, the absolute LLM narrative surprise at the previous FOMC meeting (bp), is the headline cross-validation result: meetings following a large LLM surprise see significantly higher CE disagreement, raising R^2 from 0.316 to 0.426. The 2011–18 coefficient should be interpreted with caution because the federal funds rate was near the zero lower bound for most of this period, mechanically compressing forecast disagreement; the 0.005pp 2000–04 result is the cleanest test outside the ZLB confound. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

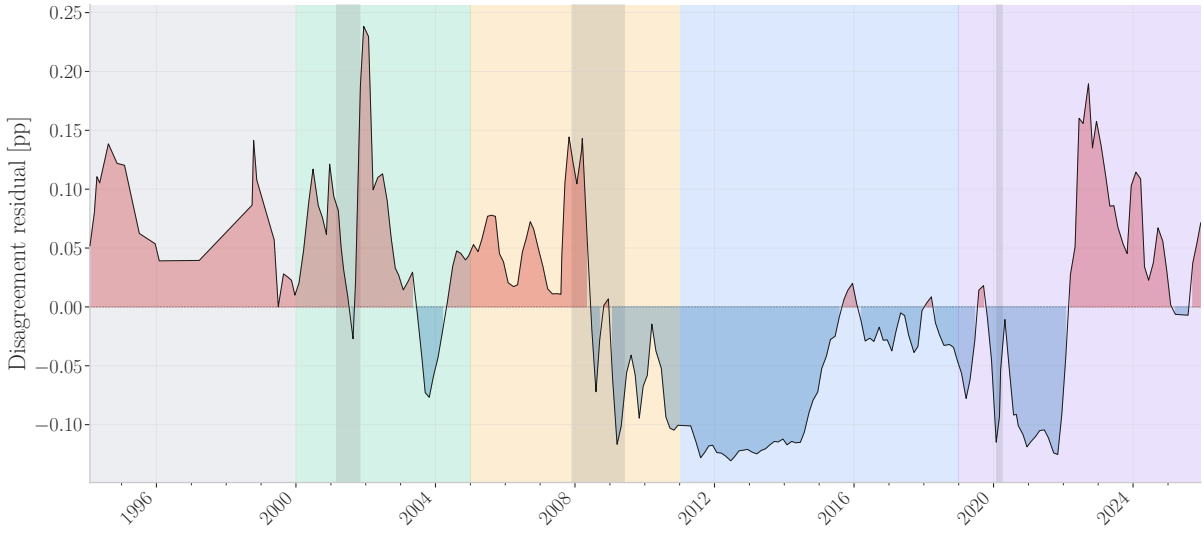
Fed documents therefore predicts an independent survey-based measure of policy uncertainty one meeting ahead. This is the load-bearing validation result: when the LLM finds a decision surprising, professional forecasters subsequently disagree more about the near-term rate path.

The regime coefficients are more suggestive. The 2000–04 coefficient, the cleanest post-1990s test outside the zero-lower-bound confound, is statistically indistinguishable from zero. The 2011–18 coefficient is large and negative across specifications, but much of that period coincides with a pinned policy rate, which mechanically compresses disagreement about short rates. Accordingly, the transparency-regime evidence should be read as consistent with the interpretation, not as the main identification result.

B.2.4 Decoder-Output Sanity Checks (Beige Book Scores)

A minimal validation question for any decoder is whether its output behaves like a real economic signal: does it co-move sensibly across dimensions, and does it align with conventional

Figure 22: CE disagreement residualised on VIX and recessions

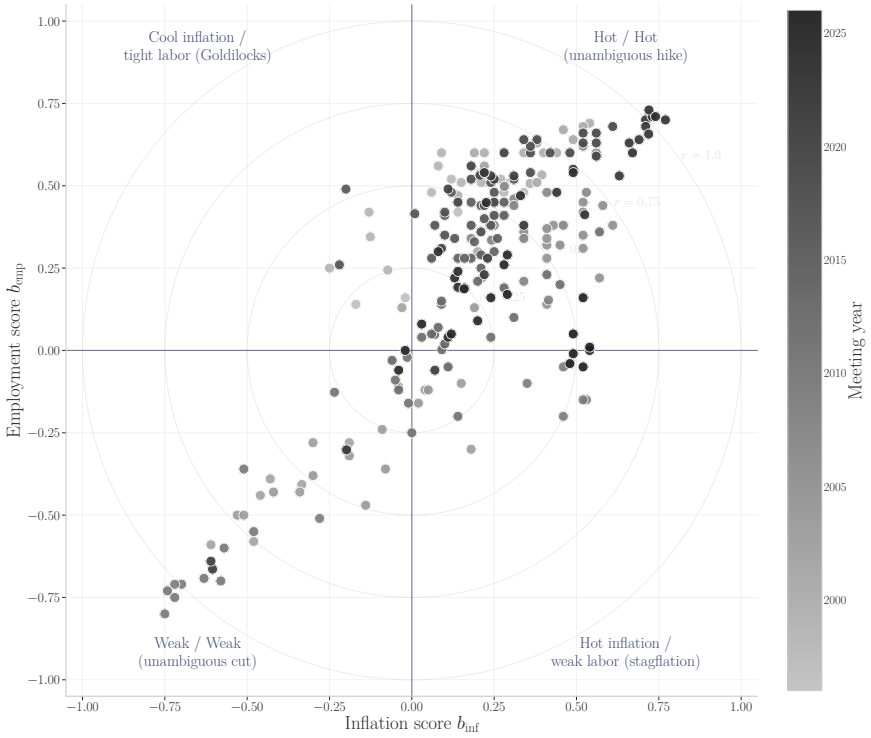


Note: CE 3M rate cross-forecaster standard deviation after partialing out the CBOE VIX and an NBER recession indicator (3-meeting centred rolling mean for readability). Positive values indicate above-trend disagreement; negative values, below-trend. Background bands mark the same Fed transparency regimes used in Figure 17 (grey 1990s, green 2000–04, amber 2005–10, blue 2011–18, purple 2019+); the residual is systematically positive in the early-transparency eras and negative through the 2011–18 ZLB period. Grey vertical bands indicate NBER recessions. The 2011–18 below-trend cluster reflects mechanical compression of forecast disagreement at the zero lower bound and should not be read as a clean transparency effect.

macroeconomic indicators? Among the four decoders, only the Beige Book Decoder produces continuous, real-valued scores (b_{inf}, b_{emp}, b_{agg}) amenable to such tests; the Statement, Press Conference, and Minutes Decoders produce categorical or sparse outputs validated through the look-ahead and cross-model checks of Section B.1. This subsubsection reports three checks on the Beige Book scores: internal co-movement of the two mandate scores, external linkage of the aggregate score to standard macro indicators, and a sharper external test on the auxiliary topics the decoder discovers unsupervised.

Dual-mandate co-movement. The mandate scores b_{inf} and b_{emp} are extracted independently but should not be informationally independent: most of the time the U.S. economy is uniformly hot or uniformly cold, and the Beige Book narratives reflect that. Figure 23 plots one point per meeting in the $(b_{\text{inf}}, b_{\text{emp}})$ plane. The two scores co-move strongly (corr $= 0.79$, $N = 241$), and 88% of meetings lie on the unambiguous-hike (NE, 72%) or unambiguous-cut (SW, 16%) diagonal. The remaining 12% are economically interpretable: the small SE cluster (hot inflation, weak labor) coincides with late-cycle stagflation-flavored episodes, and the NW points (cool inflation, tight labor) with late-1990s low-inflation expansions and the post-pandemic reopening

Figure 23: Dual-mandate plane: Beige Book inflation vs. employment scores



Note: One point per meeting in the $(b_{\text{inf}}, b_{\text{emp}}) \in [-1, 1]^2$ plane, color-coded by year. Concentric circles mark constant total mandate pressure $r = \sqrt{b_{\text{inf}}^2 + b_{\text{emp}}^2}$ at $r \in \{0.25, 0.5, 0.75, 1.0\}$. Quadrant labels indicate the implied policy regime. Sample: 241 meetings (1996–2026), v30.1 deepseek-v3.1.

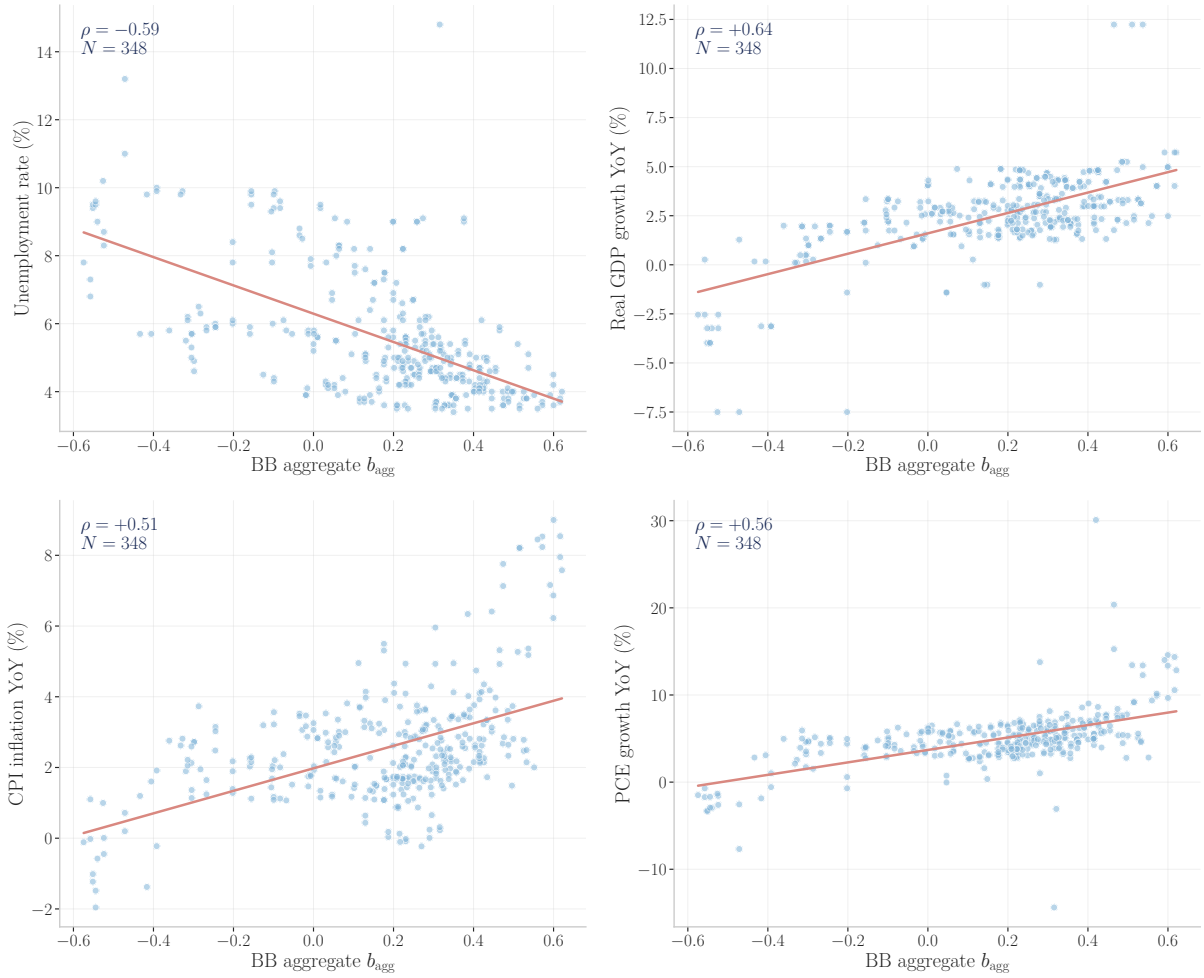
before 2022. This co-movement is the methodological reason the weighted aggregate b_{agg} is a sufficient summary in most predictive regressions, while the off-diagonal points are precisely those that load on the inflation-employment interaction in Panel B of Table 2.

Macro linkage of the aggregate score. The aggregate score correlates strongly and with the expected sign with conventional macroeconomic indicators (Figure 24): forward-filled to monthly frequency, it moves positively with real GDP growth ($\rho = +0.65$) and PCE growth ($\rho = +0.56$), positively with CPI inflation ($\rho = +0.49$), and negatively with the unemployment rate ($\rho = -0.58$).[24] The CPI link is somewhat weaker than the others, as expected: the decoder measures inflation *pressures* from regional narratives rather than realized headline price changes.

Auxiliary topics track their natural macro counterparts. The linkages above test the dimensions the LLM was *prompted* to extract. A sharper test is whether the auxiliary topics, which the decoder *discovers* unsupervised, line up with the macro indicators a human would

[24] Beige Book scores are released eight times per year; I forward-fill each score until the next release to align with monthly macro data.

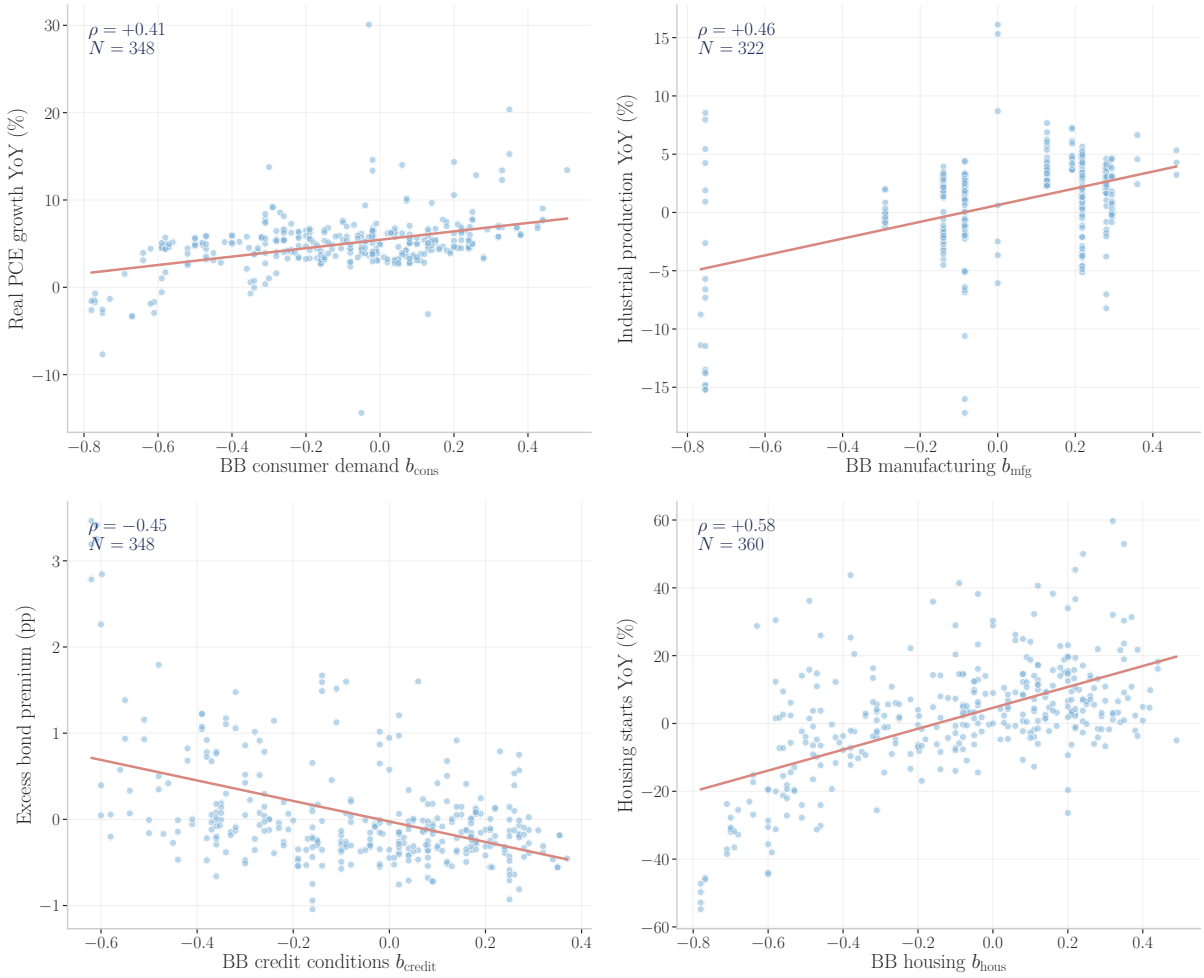
Figure 24: Macro linkage of the Beige Book aggregate score



Note: Each panel pairs the Beige Book aggregate score b_{agg} with one canonical macroeconomic indicator at monthly frequency. BB scores are forward-filled between releases. Correlations are reported in the body; sample is 1996–2026, $N = 348$ monthly observations, v30.1 deepseek-v3.1.

naturally pair them with. Figure 25 pairs the four most-frequent auxiliary topics with one canonical counterpart each: consumer demand with real PCE growth, manufacturing with industrial production, credit conditions with the Gilchrist–Zakrajšek excess bond premium, and housing activity with housing starts. All four correlations have the expected sign and magnitudes in the same range as the dual-mandate linkages above ($|\rho|$ between 0.38 and 0.56): the credit-conditions correlation is negative because high BB credit-conditions scores indicate healthy credit (pro-tightening), whereas a high excess bond premium indicates credit stress (pro-easing). The decoder is therefore well-calibrated not only on the dimensions the prompt requested, but also on topics the LLM identifies as policy-relevant from the Beige Book text alone.

Figure 25: Auxiliary topic scores and their canonical macro counterparts

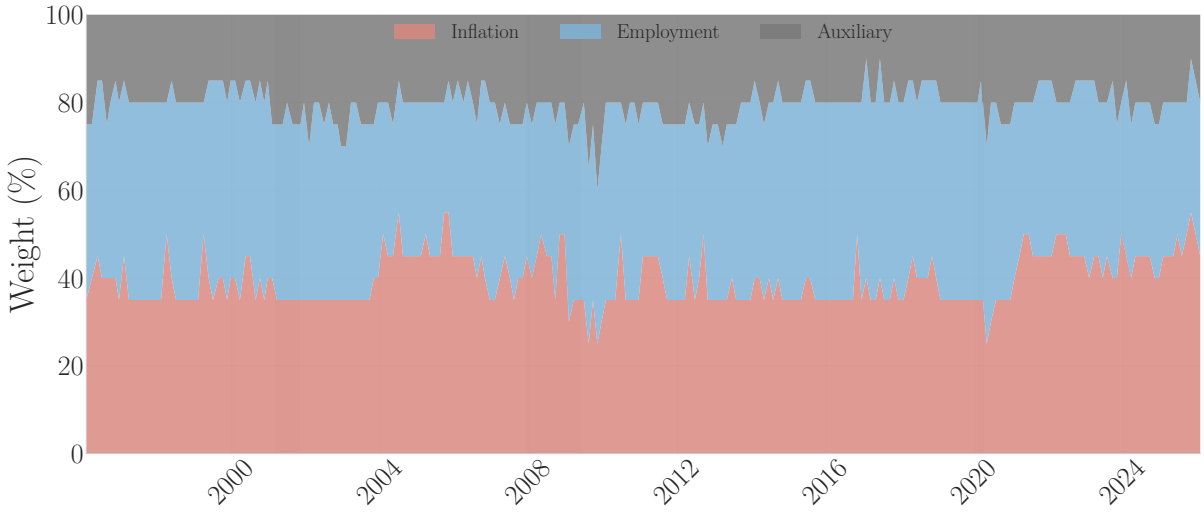


Note: Each panel pairs one LLM-discovered auxiliary topic score with one canonical macro counterpart at monthly frequency. Correlations: consumer demand vs. real PCE growth $\rho = +0.41$ ($N = 348$); manufacturing vs. industrial production YoY $\rho = +0.38$ ($N = 293$); credit conditions vs. excess bond premium $\rho = -0.47$ ($N = 348$); housing activity vs. housing starts YoY $\rho = +0.56$ ($N = 362$). Sample: 1996–2026, v30.1 deepseek-v3.1.

B.3 Architectural Validation: Beige Book Regional Aggregation

Of the four document decoders, only the Beige Book introduces a non-trivial architectural choice: the raw document is split along its twelve Federal Reserve district sections, each section is processed independently by a district-level decoder, and the twelve outputs are recombined by a national orchestrator using GDP-share weights and time-varying salience weights. This subsection verifies that those design choices do not introduce identification problems beyond what the GDP-weighted national aggregate captures. Section B.3.1 reports the weight dynamics and auxiliary topic scores. Sections B.3.2–B.3.5 run formal identification checks on the regional decomposition.

Figure 26: Beige Book Weight Dynamics



Note: Stacked area showing the decoder's weight allocation across economic variables at each FOMC meeting date (Equation 3). Inflation and employment jointly account for 85–95% of the total weight, reflecting the Federal Reserve's dual mandate. The remaining weight is distributed across auxiliary topics discovered from each Beige Book's content (consumer demand, housing, manufacturing, credit conditions). All weights sum to 1.0 at each meeting.

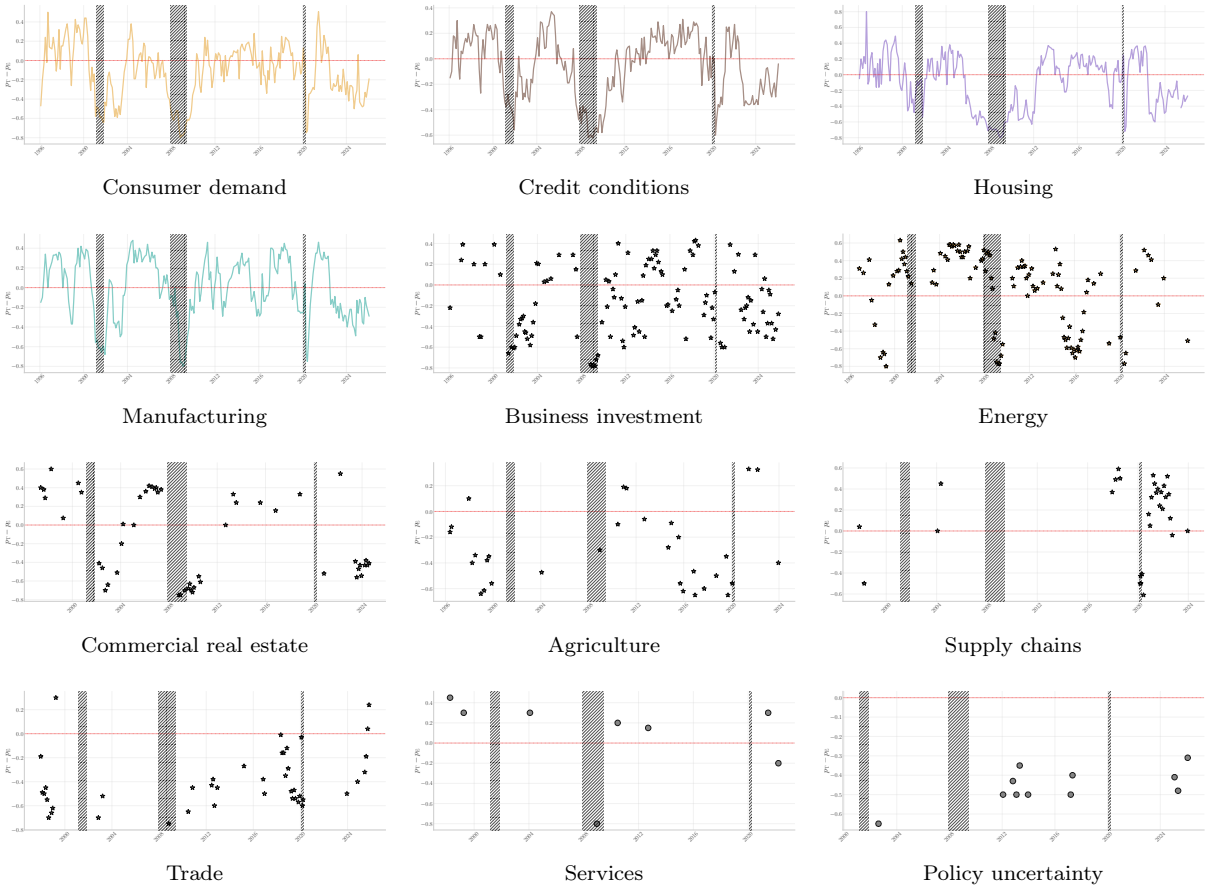
B.3.1 Weight Dynamics and Auxiliary Topic Scores

The decoder reweights its inputs meeting by meeting and discovers auxiliary topics beyond the dual mandate. This subsection documents both diagnostic outputs of that architecture: the weight allocation across topics over time, and the scored series for every auxiliary topic the decoder surfaced from the corpus.

Inflation and employment jointly absorb 85–95% of the decoder's weight in every meeting, consistent with the dual mandate (Figure 26). The remainder reallocates with the macro state: inflation weight rises through 2021–2022 as price pressures dominate the narrative, and credit conditions and employment gain weight during the 2008 crisis. The reallocation is informative for prediction: in an expanding-window forecasting exercise, the LLM-assigned weights ω_t^j outperform equal weights by 4.9 percentage points in R^2.

The auxiliary topics fall cleanly into two regimes (Figure 27). Four — consumer demand, credit conditions, housing, manufacturing — appear in nearly every meeting and track the dual-mandate cycle. The remainder surface only when the underlying narrative warrants: agriculture during commodity-price episodes, supply chains during 2021–2022, trade during tariff episodes. The decoder is therefore not a fixed taxonomy applied uniformly but a sparse, narrative-conditioned vocabulary.

Figure 27: Beige Book Auxiliary Topic Scores



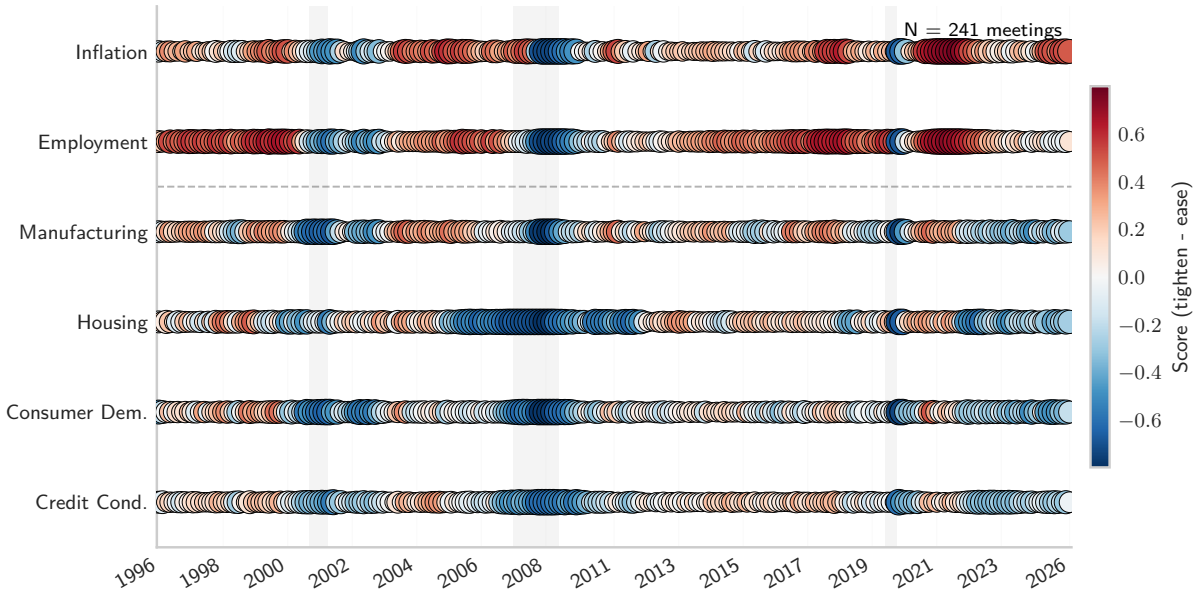
Note: Scores for all auxiliary topics discovered by the decoder, ordered by sample coverage. Positive values indicate hawkish-leaning conditions; negative values indicate dovish-leaning conditions. Consumer demand, credit conditions, housing, and manufacturing appear in ≥98% of meetings; business investment and energy in ≈50%; the remaining topics surface intermittently as economic narratives shift (commercial real estate, agriculture during commodity-price episodes, supply chains during 2021–2022, trade during tariff episodes, services and policy uncertainty during recent regimes). For sparsely sampled topics, observations are shown as discrete markers (no interpolating line). Recession periods are shaded.

B.3.2 District-Level Identification Checks

Two features of the regional architecture could in principle bias the aggregate: geographic dispersion across the twelve districts, and the rotating voting structure of the FOMC. I find no evidence that either does. The construction is the district-level analogue of the national score (Section 2.5): for each district d at meeting t, the decoder produces a policy probability simplex over (tighten, neutral, ease), and the net mandate signal $b_{d,t} = p_{\text{tighten},d,t} - p_{\text{ease},d,t} \in [-1, 1]$ is positive when local conditions lean hawkish; the GDP-weighted national aggregate is $\bar{b}_t = \sum_d \omega_d \, b_{d,t}$. I run two null tests, each exploiting institutional structure plausibly exogenous to any individual Beige Book.

The first asks whether geographic dispersion attenuates the aggregate's predictive content:

Figure 28: Topic Activation Over Time



Note: Per-meeting topic activation across the LLM's decoded narrative. Bubble size and ribbon thickness scale with the LLM-assigned weight ω_t^j; colour encodes the topic score (red: hawkish/tightening, blue: dovish/easing). Inflation and employment activate every meeting; auxiliary topics activate selectively (supply chains 2021–22, energy across the cycle), confirming the decoder's narrative-conditioned vocabulary. $N = 241$ meetings.

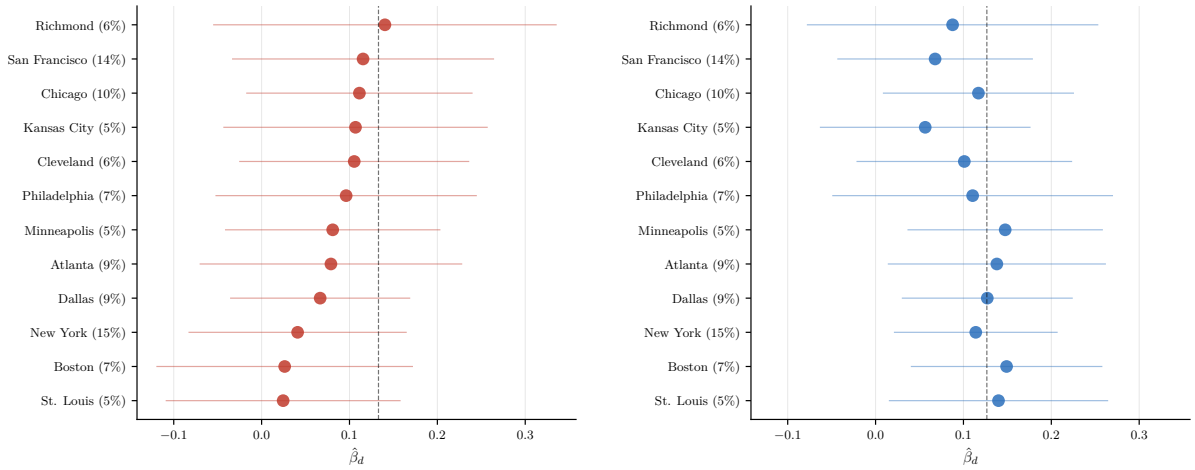
if the FOMC responds more strongly to broad-based than to concentrated signals, the national mean would understate predictive content in dispersed meetings. Define the district *coherence index* as the GDP-weighted cross-district standard deviation of net signals, $\sigma_t = (\sum_d \omega_d (b_{d,t} - \bar{b}_t)^2)^{1/2}$, and augment the baseline regression with σ_t and its interaction with $\bar{b}_t$:

$$\Delta i_t = \alpha + \beta\, \bar{b}_t + \gamma\, \sigma_t + \delta\, (\bar{b}_t \times \sigma_t) + \varepsilon_t. \tag{27}$$

Both $\hat{\gamma}$ and $\hat{\delta}$ are far from significance at any conventional level, and R^2 is unchanged: within this specification, I find no evidence that cross-district dispersion moderates the mapping from the national Beige Book aggregate to policy decisions.

The second asks whether the voting rotation reweights signals. Reserve Bank presidents rotate through four voting slots on a fixed annual schedule,[25] but all twelve presidents attend every meeting and participate in deliberations regardless of voting status. If voting-year signals carried more weight, a voter-minus-nonvoter gap should enter the regression. Constructing GDP-weighted average signals separately for the n_t^V voting and the $12 - n_t^V$ non-voting districts

[25]The Federal Reserve Bank of New York votes at every meeting. The remaining eleven Reserve Bank presidents share four rotating slots: Chicago and Cleveland alternate one slot annually; three groups of three (Boston, Philadelphia, Richmond; Atlanta, St. Louis, Dallas; Minneapolis, Kansas City, San Francisco) each contribute one representative per year on a three-year cycle.



(a) Inflation mandate ($\hat{\beta}_d^{\pi}$)

(b) Labor mandate ($\hat{\beta}_d^{\ell}$)

Figure 29: District-level slopes by mandate. Each marker is a Federal Reserve district (GDP share in parentheses on the y-axis); bars are 95% confidence intervals. District ordering is consistent across panels (sorted by inflation slope). Dashed vertical line: pooled OLS estimate. A Cochran Q test fails to reject homogeneity within either mandate ($Q^{\pi} = 2.80$, $Q^{\ell} = 2.88$, both $p > 0.99$ on 11 d.f.; $I^2 = 0\%$). Across mandates, however, the rank correlation between $\hat{\beta}_d^{\pi}$ and $\hat{\beta}_d^{\ell}$ is strongly negative (Spearman $\rho = -0.76$, $p = 0.005$): districts that are inflation-loaded tend to be relatively labor-light, and vice versa.

and adding the difference $\bar{b}_t^V - \bar{b}_t^{NV}$ to the baseline yields a coefficient of 0.006, statistically indistinguishable from zero, consistent with all districts being weighted equally in the Committee's deliberations. Together, these results suggest that the GDP-weighted aggregate summarizes district-level information efficiently for the purpose of predicting the policy decision; they do not, however, imply that the underlying signals are themselves homogeneous, a question taken up in Sections B.3.3 and B.3.4.

B.3.3 District Heterogeneous Slopes

The GDP-weighted aggregate treats every district's signal as equally informative per unit of economic size. Whether that restriction binds is an empirical question: I estimate, for each district d, a separate regression on the long panel with one observation per district-meeting pair,

$$\Delta i_t = \alpha_d + \beta_d^{\pi}\, b_{d,t}^{\pi} + \beta_d^{\ell}\, b_{d,t}^{\ell} + \varepsilon_{d,t}, \tag{28}$$

with district fixed effects α_d and meeting-clustered standard errors. Figures 29a and 29b display the estimated $\hat{\beta}_d^{\pi}$ and $\hat{\beta}_d^{\ell}$, ordered top-to-bottom by the inflation slope.

The visual scatter of point estimates is not statistically decisive. A Cochran Q test fails to reject homogeneity within either mandate at any conventional level, with $I^2 = 0\%$ for both,

but with confidence intervals this wide the appropriate conclusion is limited: the data do not support estimating district-specific slope magnitudes. The cross-mandate ordering, by contrast, is statistically meaningful: the Spearman rank correlation between $\hat{\beta}_d^\pi$ and $\hat{\beta}_d^\ell$ is strongly negative and statistically significant ($\rho = -0.76$), so districts that load relatively heavily on inflation load relatively lightly on labor. This pattern is consistent with regional economic specialization, but it lives in the relative composition of mandates within a district, not in absolute slope magnitudes; district slopes are themselves uncorrelated with GDP weights for both dimensions, ruling out the alternative that larger districts are systematically more informative per dollar of output. The composition is, in any case, not exploitable for prediction: estimating district-specific slopes on the first 137 meetings and applying them to construct a $\hat{\beta}$-weighted aggregate on the held-out 92 meetings (a 60/40 chronological split, $N = 229$) yields a hold-out R^2 of 0.449, statistically indistinguishable from 0.450 under GDP weighting. The pooled GDP-weighted aggregate is, in this sense, a well-specified aggregator.

B.3.4 Big-District Disagreement as a Signal-Quality Indicator

Even when the aggregate is well-specified on average, episodes in which large and small districts send conflicting signals might flag meetings where the national mean obscures more than it reveals. To test this, I split the twelve districts into a "big-6" group (top six by average GDP weight: New York, San Francisco, Chicago, Atlanta, Dallas, Boston, covering approximately 64% of sample GDP) and a "small-6" group (Philadelphia, Cleveland, Richmond, Kansas City, Minneapolis, St. Louis). At each meeting t, I take GDP-weighted averages within each group across both mandate dimensions and define the big-minus-small gap as

$$\Delta_t = \frac{1}{2}\Big[\big(\bar{b}^\pi_{\text{big},t} - \bar{b}^\pi_{\text{small},t}\big) + \big(\bar{b}^\ell_{\text{big},t} - \bar{b}^\ell_{\text{small},t}\big)\Big], \tag{29}$$

positive when large districts signal more tightening pressure than small ones. The gap is small in ordinary times — median absolute value roughly 0.06 on the $[-1, +1]$ scale — but spikes to $|\Delta_t| \geq 0.25$ at every macro turning point in the sample: the dot-com contraction (2001–2003), the financial crisis (2007–2008), the COVID shock (six consecutive 2020 meetings), and the 2023–2025 tightening cycle. Figure 30 plots the full series.

Splitting meetings at the median $|\Delta_t|$ reveals a modest predictive-accuracy gradient: the sample correlation between $\bar{b}_t$ and Δi_t is $r = 0.388$ in low-gap meetings and $r = 0.365$ in high-gap

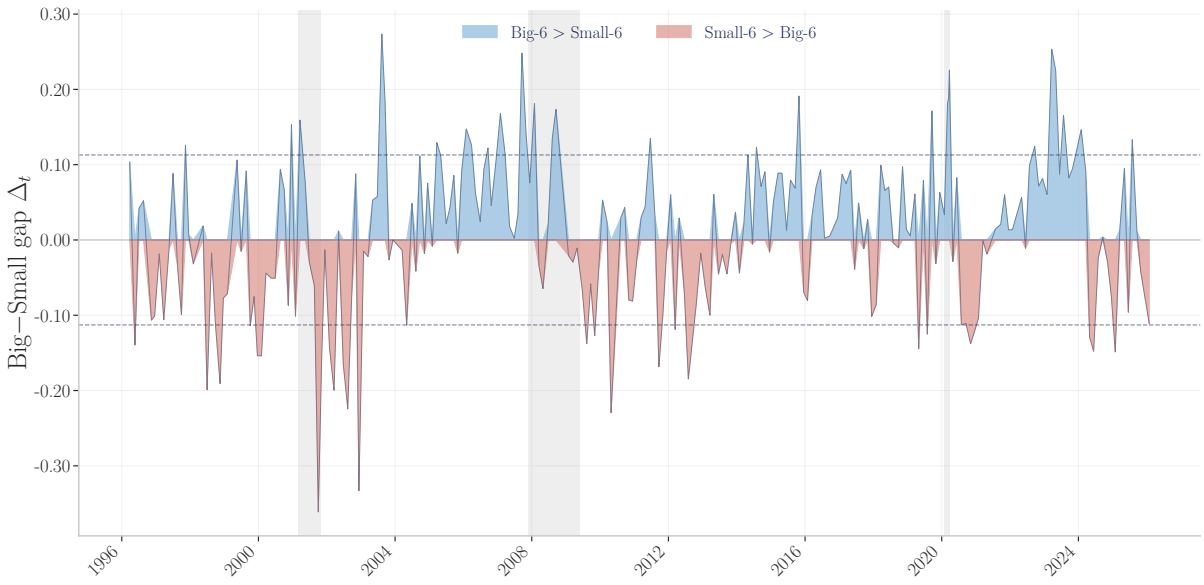


Figure 30: Big-minus-small district gap Δ_t over time. Positive values indicate large districts (≈64% of GDP: NYC, SFR, CHI, ATL, DAL, BOS) signal more policy tightening than small districts. Dashed lines mark the ±75th percentile. Gray shading: NBER recessions.

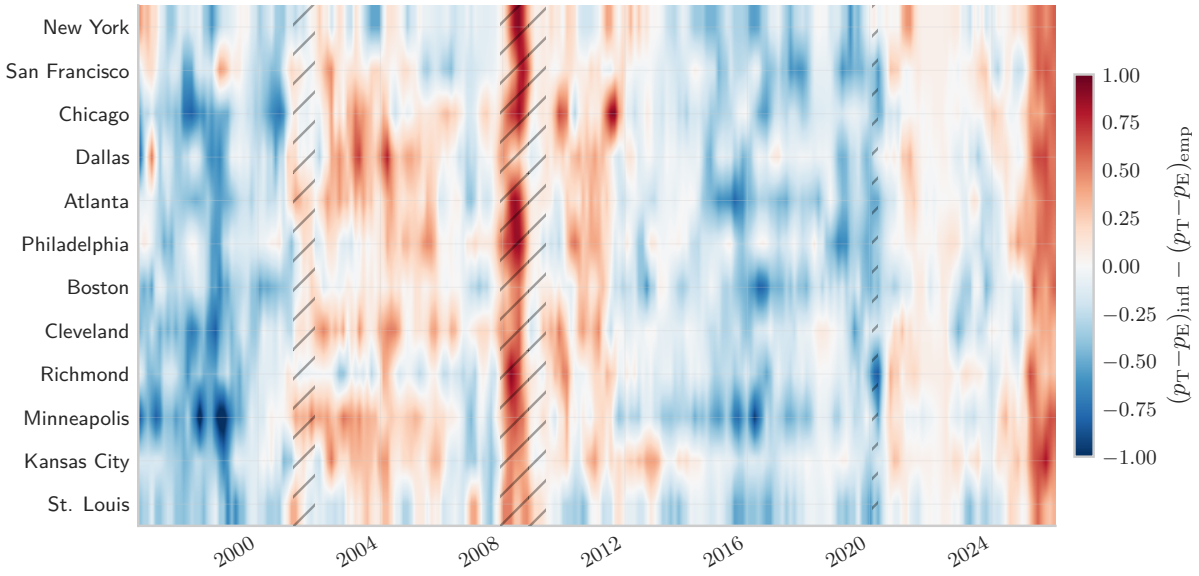
meetings. Table 23 formalizes the comparison with two interaction specifications, $|\Delta_t|$ directly (column 1) and a binary above-median indicator (column 2). Both interaction coefficients are positive and imprecisely estimated; with 229 meetings, the test lacks power against a noisy moderator. The regression does, however, identify a level effect: meetings with strong big–small disagreement see smaller average rate changes, suggesting the FOMC acts more cautiously when its regional intelligence is internally inconsistent. In high-gap meetings the small-6 aggregate is marginally more predictive ($r = 0.375$) than the big-6 ($r = 0.346$). One possible interpretation is compositional: the largest districts are more exposed to finance- and technology-sensitive conditions, whereas several smaller districts are closer to manufacturing, agriculture, and energy. Given the imprecision of the interaction estimates, however, that mechanism should be read as suggestive rather than established. The gap Δ_t is therefore not a standalone predictor, and the interaction tests are too imprecise to support a strong moderator claim: neither group dominates consistently enough to displace the national aggregate. The safer interpretation is descriptive rather than structural. A large $|\Delta_t|$ can be treated as a caution indicator, marking meetings in which the Beige Book aggregate is more likely to compress cross-regional disagreement and should be read alongside hard data or statement language.

Table 23: Big-District Gap as a Signal-Quality Indicator

	(1) Continuous	(2) Discrete
$\bar{s}_t$	0.254***	0.222***
	(0.090)	(0.078)
$\lvert\Delta_t\rvert$	-0.352	
	(0.294)	
$\bar{s}_t \times \lvert\Delta_t\rvert$	-0.090	
	(0.875)	
D_t^{big}		-0.044
		(0.028)
$\bar{s}_t \times D_t^{\text{big}}$		0.086
		(0.091)
Constant	-0.009	-0.018
	(0.032)	(0.026)
R^2 (IS)	0.148	0.145
R^2 (OOS)	—	—
Obs.	229	229

Note: 1996-03 to 2026-01. $\bar{s}_t = (\hat{s}_t^{\pi} + \hat{s}_t^{\ell})/2$ is the mean of the inflation and labor GDP-weighted aggregates. $|\Delta_t|$ is the absolute big-minus-small gap (GDP aggregate of the top-6 districts minus GDP aggregate of the bottom-6, averaged across dimensions). $D_t^{\text{big}} = \mathbf{1}[|\Delta_t| > \text{median}(|\Delta_t|)]$. OOS R^2 is not reported for these interaction regressions; correlations are sample statistics, not OOS estimates. HAC SEs (Newey-West). *, **, *** at 10%, 5%, 1%.

Figure 31: District Mandate Conflict over Time



Note: District-level mandate-priority heatmap: $(p_T - p_E)_{\text{infl}} - (p_T - p_E)_{\text{emp}}$ per district per meeting, 3-meeting rolling smooth. Red bands mark periods when the district's narrative is dominated by inflation pressure; blue bands mark periods when labor-market weakness dominates. Districts ordered by approximate GDP weight; NBER recessions hatched. The heatmap complements the big-minus-small aggregate (Figure 30) by showing *which* districts disagree at any given moment.

Table 24: State-Dependent Mandate Weights: Beige Book Signals in Macro Regimes

	(1) Baseline	(2) + Infl Regime	(3) + Labor Regime	(4) Full
s_t^π	0.133	-0.190	0.177	-0.119
	(0.124)	(0.119)	(0.120)	(0.110)
s_t^ℓ	0.127	0.171*	0.360**	0.342**
	(0.093)	(0.095)	(0.147)	(0.142)
D_t^π		-0.080**		-0.065*
		(0.036)		(0.036)
$s_t^\pi \times D_t^\pi$		0.507***		0.418***
		(0.133)		(0.130)
D_t^ℓ			0.136**	0.113**
			(0.057)	(0.055)
$s_t^\ell \times D_t^\ell$			-0.386**	-0.248*
			(0.155)	(0.141)
Constant	-0.058**	-0.044**	-0.170***	-0.134**
	(0.024)	(0.019)	(0.051)	(0.052)
R^2	0.108	0.186	0.174	0.220
Obs.	271	271	271	271

Note: 1996-03 to 2026-01. OLS with HAC standard errors (Newey-West). s_t^π and s_t^ℓ are GDP-weighted cross-district Beige Book signals for the inflation and labor mandate dimensions. $D_t^\pi = \mathbf{1}[\pi_t > 2\%]$ where π_t is CPI YoY. $D_t^\ell = \mathbf{1}[u_t > \tilde{u}_t]$ where $\tilde{u}_t$ is the HP-filtered trend. *, **, *** denote significance at 10%, 5%, 1%.

B.3.5 State-Dependent Mandate Weights

The baseline specifications impose constant coefficients on the inflation and labor signals. Clarida et al. (1999) predicts otherwise: the Fed should shift attention toward whichever objective is further from target. To test this, I define two binary indicators,

$$D_t^\pi = \mathbf{1}\big[\pi_t > 2\%\big], \quad \pi_t = \text{CPI (year-over-year)}$$
$$D_t^\ell = \mathbf{1}\big[U_t > \tilde{U}_t\big], \quad \tilde{U}_t = \text{HP-filtered unemployment trend},$$

and estimate the augmented regression

$$\Delta i_t = \alpha + \beta_\pi b_t^\pi + \beta_\ell b_t^\ell + \gamma_\pi D_t^\pi + \varphi_\pi\big(b_t^\pi \times D_t^\pi\big) + \gamma_\ell D_t^\ell + \varphi_\ell\big(b_t^\ell \times D_t^\ell\big) + \varepsilon_t, \tag{30}$$

where b_t^π and b_t^ℓ are the GDP-weighted cross-district Beige Book signals.[26] The effective weight on the inflation signal in regime $D_t^\pi = 1$ is $\beta_\pi + \varphi_\pi$, and analogously for labor.

The full model (Table 24, column 4) yields two findings of opposite sign. The inflation

[26] Both regime series are from FRED, merged to meeting dates using the most recently available release before each meeting. The HP trend is estimated on the full sample, so D_t^ℓ is not a real-time indicator; it uses post-t data to define what "above trend" means. Because the HP trend mostly tracks visible business-cycle phases, this approximation is unlikely to materially affect the results.

interaction $\hat{\varphi}_\pi = 0.418$, significant at the 1% level, implies the effective weight on b_t^π rises from 0.171 when CPI is below 2% to 0.589 when it exceeds 2%, a 3.4-fold increase consistent with a convex-loss reaction function: each unit of inflationary signal maps to a larger policy response when price stability is already stressed. This pattern is not confined to 2021–2022: CPI has been above 2% for most of the post-1996 sample outside 2008–2015, and the high-inflation regime covers 65% of all meetings. The labor interaction cuts in the opposite direction. With $\hat{\varphi}_\ell = -0.248$, marginally significant at the 10% level, b_t^ℓ *loses* predictive content when unemployment runs above trend — the opposite of the symmetric mandate-switching prediction.

One possible explanation is mechanical: when labor-market slack is greatest, the zero lower bound may compress observed rate changes and mechanically weaken the labor interaction. The robustness check does not support that explanation. Re-estimating column 4 on the sample that excludes both ZLB episodes (December 2008–December 2015 and March 2020–March 2022, $N = 188$) leaves the labor interaction essentially unchanged at $\hat{\varphi}_\ell = -0.292$, if anything slightly more negative than the full-sample estimate. The inflation interaction survives the same exclusion ($\hat{\varphi}_\pi = +0.341$, less precise on the smaller sample) and attenuates to zero within the ZLB sample itself ($\hat{\varphi}_\pi = +0.064$, $N = 83$), where rates cannot move regardless of the inflation signal. The labor interaction behaves differently: it is driven by non-ZLB labor-slack episodes, chiefly the 2001–2003 recovery and the 2024–2025 cooling. The implication is straightforward. The Fed's reaction function appears asymmetric in inflation, but the symmetric mandate-switching prediction does not extend to labor: when employment is below target, the marginal Beige Book labor signal carries less weight, not more, in the rate decision. Together the two interactions raise R^2 from 0.189 to 0.333 (+14.4 pp), with the aggregate Beige Book signal significant in every specification.

B.3.6 Aggregation Scheme and the Case for the Dual-Mandate Specification

The decoder assigns LLM-estimated policy-salience weights ω_t^j to each topic at each meeting (Section 2.5). Table 25 compares this time-varying scheme to two simpler alternatives — equal weights across all topics, and constant OLS coefficients on inflation and employment alone — and provides the empirical justification for the dual-mandate specification used throughout the main text.

All three specifications are significant and carry near-identical in-sample fit ($R^2 \approx 0.40$), so the Beige Book signal is not an artifact of the specific weighting choice. The LLM-estimated

Table 25: Beige Book Weighting Scheme Robustness: IS and OOS Fit

	Federal Funds Rate Change (Δi_t)		
	(1) LLM weights	(2) Equal weight	(3) Dual mandate
Constant	-0.014	-0.008	-0.022
	(0.032)	(0.031)	(0.034)
Δi_{t-1}	0.526***	0.535***	0.540***
	(0.089)	(0.104)	(0.084)
d_{BB}	0.017	0.011	0.026
	(0.031)	(0.031)	(0.033)
BB^{agg} (LLM)	0.133***		
	(0.040)		
BB^{equal}		0.153***	
		(0.048)	
BB^{infl}			0.065
			(0.055)
BB^{empl}			0.038
			(0.046)
R^2 (IS)	0.403	0.403	0.395
R^2 (OOS)	0.454	0.408	0.463
Obs.	271	271	271

Note: 1996-03 to 2026-01. Comparison of three aggregation schemes for the Beige Book signal. *LLM weights*: $\bar{s}_t = \sum_j \omega_t^j s_t^j$ using LLM-assigned policy-salience weights updated each meeting. *Equal weight*: uniform $\omega_t^j = 1/J$ across topics. *Dual mandate*: separate OLS-estimated coefficients on the inflation and employment components only. All specifications include Δi_{t-1} (lagged rate change) and d_{BB}. OOS R^2 from a chronological 60/40 split (train $N = 162$, test $N = 109$); the benchmark is the prevailing training-sample mean (not the test-period mean), so the denominator is fixed before the hold-out begins. Newey-West HAC standard errors. ***, **, *: 1%, 5%, 10%.

salience weights outperform equal weights out of sample by 4.6 percentage points (0.454 vs. 0.408), confirming that meeting-by-meeting reallocation toward the most policy-relevant topics carries genuine information rather than in-sample noise. The dual-mandate specification — dropping auxiliary topics and estimating constant OLS coefficients on inflation and employment alone — nonetheless achieves the highest OOS fit (0.463), edging out the LLM-weighted aggregate. The auxiliary-topic layer is therefore useful primarily as measurement and audit infrastructure: it shows what additional narratives the decoder detects and how salience shifts across meetings (Figure 27), even though those extra topics do not improve hold-out prediction once the final aggregate is estimated. This is the empirical reason the main text restricts attention to the dual-mandate specification: for inference, the parsimonious predictor is the more stable one.

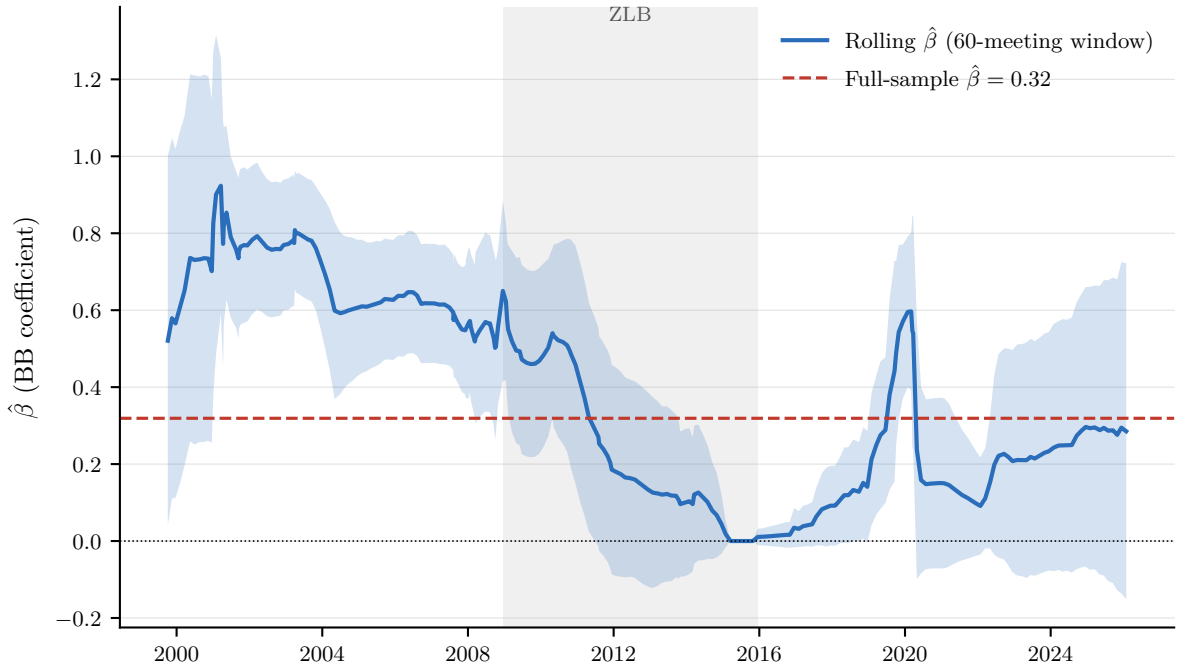


Figure 32: Rolling Beige Book coefficient $\hat{\beta}_t$. Each point is the OLS slope on the GDP-weighted Beige Book aggregate from a 60-meeting rolling window, with ±2 standard-error bands. The dashed horizontal line is the full-sample estimate ($\hat{\beta} = 0.32$).

B.3.7 Temporal Stability

The full-sample coefficient $\hat{\beta} = 0.32$ (SE = 0.08) masks substantial time variation. Figure 32 reports rolling 60-meeting OLS estimates of $\hat{\beta}_t$ from the baseline specification $\Delta i_t = \alpha + \rho\, i_{t-1} + \beta\, \bar{b}_t + \varepsilon_t$, with ±2 standard-error bands.

The coefficient tracks macro regimes. During the zero-lower-bound period (roughly 2009–2015), $\hat{\beta}_t$ compresses toward zero — when the policy rate is constrained, no Beige Book signal can predict rate changes — mirroring the labor-coefficient finding in Section B.3.5. As the Fed exits the lower bound and especially through the 2022–2023 tightening cycle, $\hat{\beta}_t$ rises sharply, peaking near 0.92, almost three times the full-sample mean, as the Beige Book's regional evidence on labor shortages and price pressures becomes the proximate driver of meeting-to-meeting decisions. The full-sample estimate of 0.32 is therefore best read as the time-average of a state-dependent relationship, not as a structural constant.

C Filtration and Surprise-Measure Properties

This appendix characterises four properties of the filtration's output. The document-only prior $\mathcal{P}_4$ is broadly well calibrated, with the realised outcome falling at or near the modal forecast in roughly 70% of meetings and accuracy concentrated in steady-state regimes (Section C.1). The residual surprise inherits the equity-return loading of standard market-based shocks rather than being diffuse across predictors, and the LLM's self-reported salience score adds no independent dimension beyond the prior's concentration at the realised outcome. An optional news stage (Section C.2) refines the prior in identifiable inter-meeting episodes without weakening forward-rate content.

C.1 Distributional Analysis of the Filtration

I present the full distributional analysis of the $\mathcal{P}_1 \to \mathcal{P}_4$ Bayesian filtration summarized in Section 5. While the first moment of the distribution is a sufficient statistic for predicting rate changes (Table 3), the full probability distributions reveal which documents contribute what information and when the system succeeds or fails.

C.1.1 $\mathcal{P}_4$ Probability Mass Over Time

Figure 33 reports the full $\mathcal{P}_4$ distribution at every FOMC meeting in the sample. The stacked-area chart compresses the rate-change support into five buckets and traces how probability mass migrates across regimes, providing a visual complement to the moment-based analyses elsewhere in this appendix.

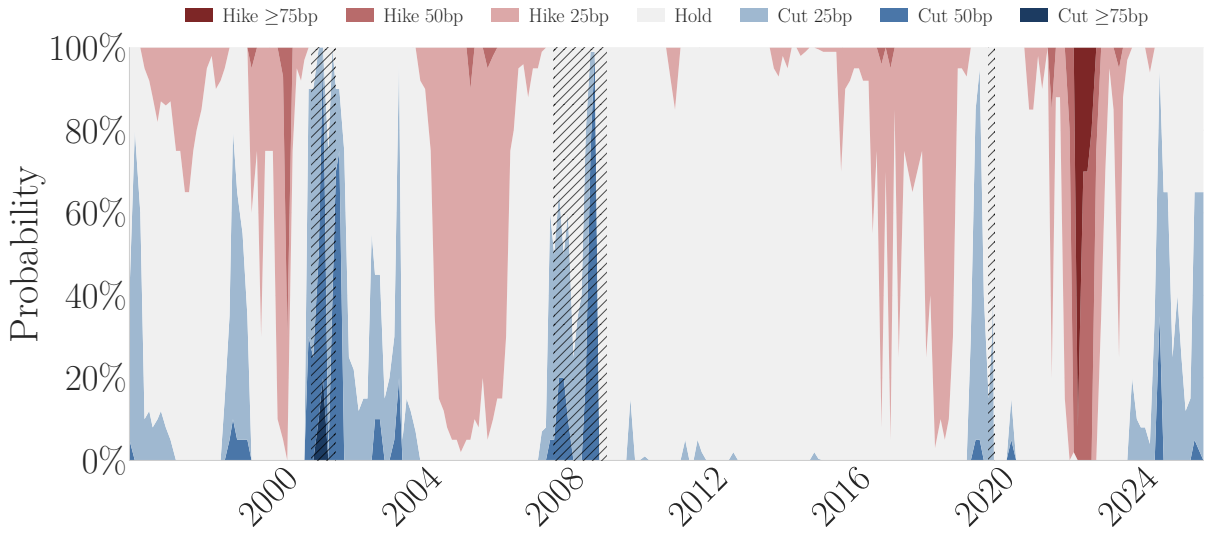
A complementary view of the same filtration is the distribution of update magnitudes $|E[\Delta i_t \mid \mathcal{P}_j] - E[\Delta i_t \mid \mathcal{P}_k]|$ at all stage pairs (j, k) with $j < k$ (Figure 34), separating marginal (adjacent-stage) from cumulative updates.

C.1.2 Filtration Accuracy

For each meeting t and filtration stage k, I compute $P_k(\Delta i_t^{\text{actual}})$: the probability the Forecaster assigned to the outcome that actually occurred. This measures how inferrable the decision was from public documents available at stage k, not Fed transparency per se, since the Forecaster observes only Fed publications and not market data or private briefings.

Table 26 reports summary statistics by stage. Mean accuracy is approximately 70% across

Figure 33: $\mathcal{P}_4$ Probability Mass Across FOMC Meetings



Note: Full probability distribution $\mathcal{P}_4$ over rate-change scenarios at each FOMC meeting date. The stacked areas show the probability mass assigned to each action: cuts (≥50 bp in dark blue, 25 bp in light blue), hold (gray), and hikes (25 bp in light red, ≥50 bp in dark red). During the zero lower bound period (2008–2015), hold probability approaches 100%. The 2022–2023 tightening cycle shows dominant hike probabilities, with large hikes (≥50 bp) concentrated in mid-2022. Regime transitions (2001 recession, 2007–08 crisis, 2015 liftoff, 2020 pandemic) exhibit rapid mass reallocation across scenarios.

Table 26: Filtration Accuracy: P_k(Realized Outcome) by Stage

Stage	N	Mean	Median	Std	$P > 0.50$	$P > 0.80$
$\mathcal{P}_1$ (Statement$_{t-1}$)	211	0.719	0.900	0.327	75.4%	59.2%
$\mathcal{P}_2$ (Press conf.$_{t-1}$)	88	0.712	0.900	0.345	73.9%	58.0%
$\mathcal{P}_3$ (Minutes$_{t-1}$)	269	0.701	0.850	0.339	73.6%	56.5%
$\mathcal{P}_4$ (Beige Book$_t$)	242	0.705	0.850	0.323	73.6%	54.5%

Note: Filtration accuracy is the probability the Forecaster's distribution assigned to the realized FOMC decision at each stage. Higher values indicate decisions more inferrable from public documents alone. $P > 0.50$: fraction of meetings where the realized outcome was the mode. $P > 0.80$: fraction with near-certainty.

all stages (range 0.701–0.719), with the realized outcome matching the modal forecast in 73–75% of meetings ($P > 0.50$ column). These headline numbers should be interpreted against the base rate: 70% of FOMC meetings result in no change, and an always-hold rule would achieve 70% modal accuracy. The informative comparisons are therefore conditional on the type of action.

Holds are highly predictable from public text (71% at $\mathcal{P}_4$), while large cuts (≥50bp) are nearly impossible to anticipate (7%). This reflects the nature of large easing actions: they are overwhelmingly emergency or intermeeting responses to sudden shocks (January 2001, March 2020) that no pre-meeting document could foresee. Standard hikes (25bp) are well-predicted (65%), consistent with the Fed's practice of telegraphing tightening cycles through forward guidance.

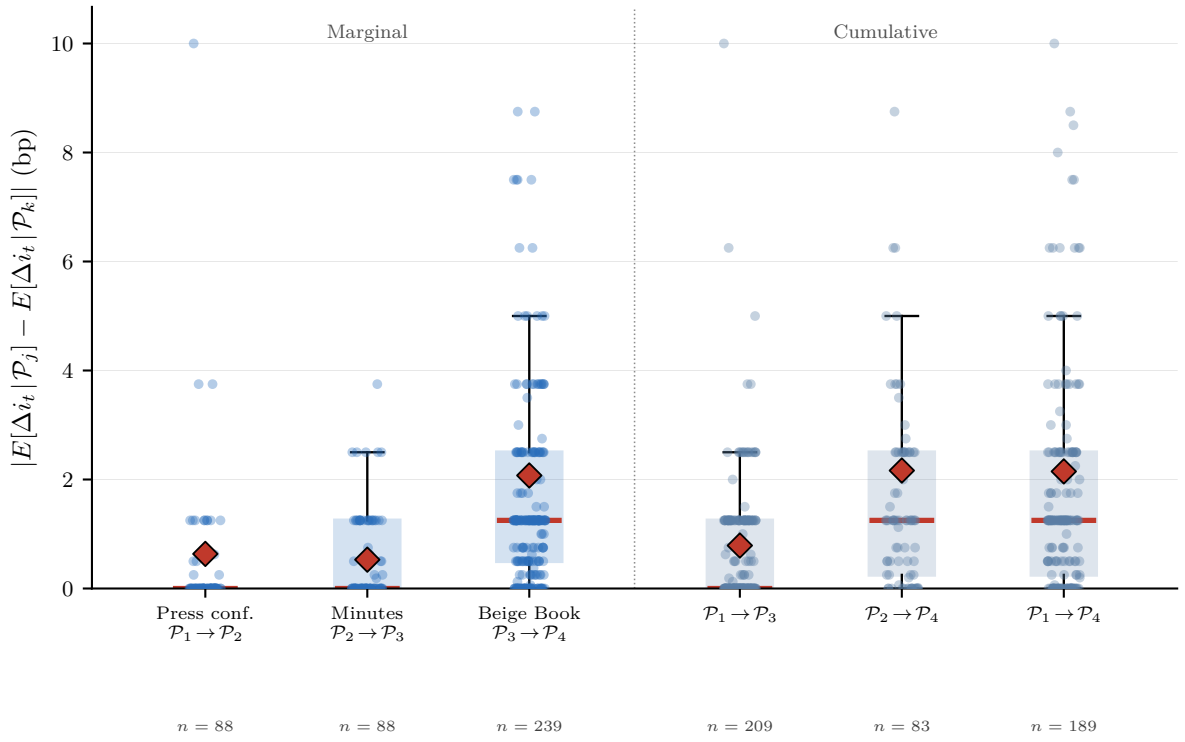
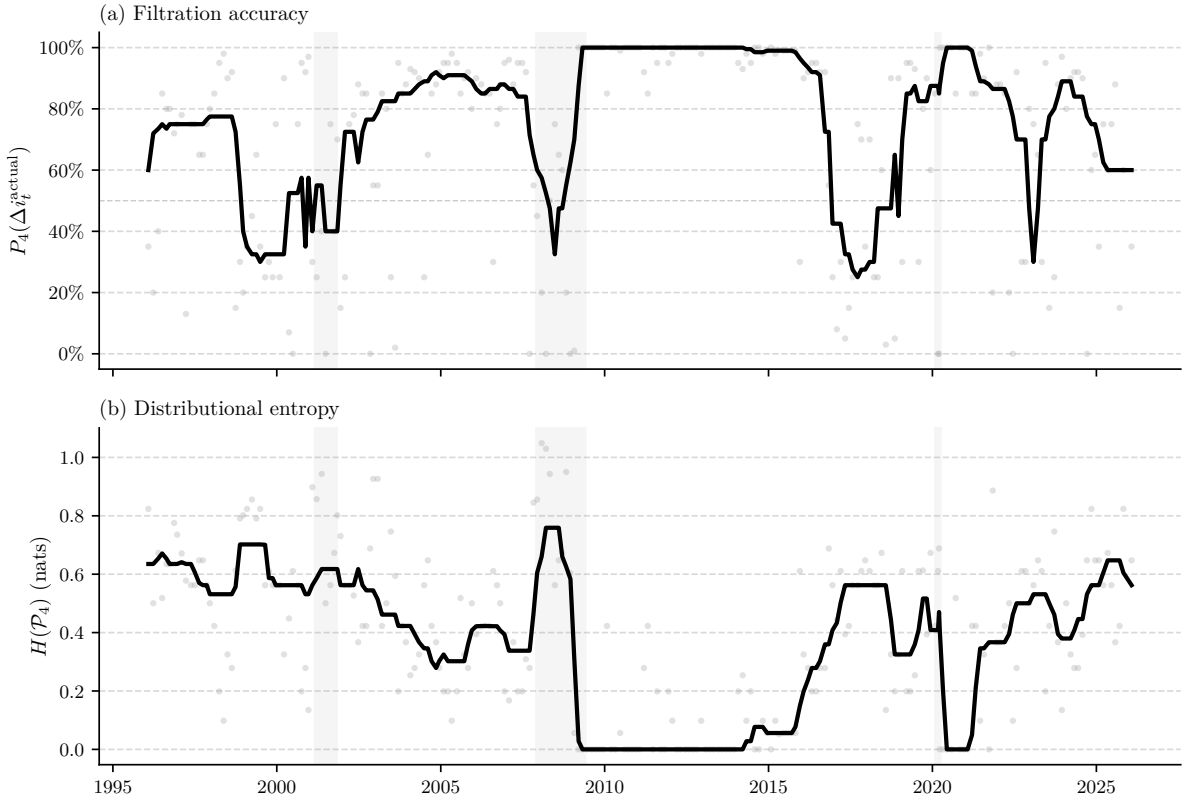


Figure 34: Prior update magnitudes $|E[\Delta i_t \mid \mathcal{P}_j] - E[\Delta i_t \mid \mathcal{P}_k]|$ at all stage pairs (j, k), $j < k$. The left group shows marginal (adjacent-stage) updates; the right group shows cumulative updates spanning two or three stages. Each box uses the pairwise-complete sample; n is reported below each box. Diamonds mark the mean; dots are individual meetings. Among marginal updates the Beige Book produces the largest average revision, reflecting fresh real-economy information not present in the policy-focused earlier documents; the Press Conference produces the smallest. Cumulative updates are larger by construction, with $\mathcal{P}_1 \to \mathcal{P}_4$ capturing the total pre-meeting information content.

The most interesting pattern emerges from the stage comparison on the common sample ($n = 182$ meetings with both $\mathcal{P}_1$ and $\mathcal{P}_4$). The Beige Book raises accuracy for hikes by 7.4 percentage points ($\mathcal{P}_4 > \mathcal{P}_1$ in 84% of hike meetings) but slightly reduces it for holds (−3.4pp). The Beige Book's district-level evidence on labor tightness and price pressures provides the confirming evidence needed to shift expectations toward tightening. For hold meetings, the same regional detail introduces two-sided risks that pull probability from an already-confident prediction.

Figure 35 plots $\mathcal{P}_4(\Delta i_t^{\text{actual}})$ and distributional entropy over time. Accuracy is highest during steady-state regimes: the zero lower bound period (2009–2015) approaches 95% as the system confidently predicts holds, and the measured tightening cycle (2004–2006) reaches 90% as the Fed's "measured pace" language made 25bp hikes nearly certain. Accuracy is lowest during transitions: the 2001 easing cycle, the 2015–2019 normalization, and the 2022 rapid tightening all show sharp declines as the policy reaction function became harder to infer from text alone. Panel (b) shows a broadly mirror pattern in entropy, though magnitudes should be read with

Figure 35: Filtration Accuracy and Distributional Entropy Over Time



Note: (a) $\mathcal{P}_4(\Delta i_t^{\text{actual}})$ for each FOMC meeting (gray dots) with 12-meeting centered rolling median (black line). (b) Shannon entropy $H(\mathcal{P}_4)$ at each meeting. High accuracy periods (ZLB, measured tightening) coincide with low entropy; transition periods show rising entropy as the distribution broadens to accommodate uncertainty about the direction and magnitude of policy changes. Gray shading: NBER recessions. Sample: 242 meetings (1996–2026).

caution: $\mathcal{P}_4$ distributions often have only two or three support points, so $H(\mathcal{P}_4)$ is mechanically sensitive to support size and is best read as a directional indicator of uncertainty rather than a precise information-theoretic quantity.

C.1.3 Residual Predictability of the Surprise

Table 5 identifies which macro-financial variables predict each surprise but not how much each contributes. I compute partial R^2 for each predictor as $R^2_{\text{full}} - R^2_{\text{full minus that predictor}}$.

The LLM surprise ($R^2 = 0.166$) and the market-based measures load materially on the same predictor: equity-return information. The S&P 500 contribution is 35.1% for the LLM, 34.6% for FF1, and 26.2% for both FF4 and ED1, declining to 17.4% only at the longer-dated ED4. Romer and Romer (2004) is the outlier: its predictability concentrates almost entirely in two yield-curve variables, with term spread (51.4%) and Treasury skewness (39.3%) together exceeding 90% and

Table 27: Variance Decomposition of Surprise Predictability

	R&R	LLM	FF1	FF4	ED1	ED4
NFP Surprise	1.8%	0.5%	1.8%	2.2%	3.5%	4.3%
NFP (12m)	1.1%	0.4%	0.9%	2.5%	6.5%	13.4%
S&P 500	2.8%	35.1%	34.6%	26.2%	26.2%	17.4%
Term Spread	51.4%	7.3%	8.6%	22.1%	14.9%	9.8%
Commodity	0.5%	0.2%	0.1%	2.1%	3.1%	8.8%
Treasury Skew	39.3%	17.5%	25.5%	11.2%	11.6%	12.7%
Total R^2	0.203	0.166	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.

S&P 500 contributing only 2.8%. The narrative pipeline therefore behaves like a market-based measure in the predictor space it occupies, even though it never observes price data directly: Beige Books, Minutes, and Statements discuss equity and financial conditions qualitatively, and the system internalises those signals in a way that mirrors how market-based shocks load on realised returns.

Rolling window analysis across 176 sixty-meeting windows confirms this structure is stable over time, with all measures showing a gradual declining trend in overall predictability.[27]

C.1.4 Salience and Prior Concentration

The salience score $\mu_t \in [0, 1]$, which the Surprise Extractor assigns alongside the point forecast, is almost entirely explained by the prior's concentration at the realised outcome: a single regressor, $p(\text{actual} \mid \mathcal{P}_4)$, accounts for 77% of μ_t's variance with $\text{corr}(\mu_t, p(\text{actual})) = -0.89$. Salience is therefore not an independent dimension of the monetary policy event; it is a near-monotone restatement of how much prior probability the document-only filtration assigned to what actually happened. The pipeline uses only $\hat{s}_t$ for identification, so this is a property of μ_t rather than a constraint on the analysis, but it affects how the score should be interpreted.

Table 28 runs the full set of regressions of μ_t on properties of the $\mathcal{P}_4$ prior distribution, organised by conceptually distinct dimensions. Adding the surprise magnitude $|\hat{s}_t|$ raises R^2 only marginally to 0.78 (column 2), and $|\hat{s}_t|$ itself is not significant once $p(\text{actual})$ is included.

[27] Pairwise bootstrap tests (H_0: $R^2_{\text{LLM}} = R^2_{\text{other}}$, 1,000 block-bootstrap samples, block length 4) yield no significant difference for any measure (smallest $p = 0.226$, FF1). The 5–20% R^2 range is a common feature of monetary surprise measurement broadly, not specific to the narrative approach.

Table 28: Salience score and prior distribution properties

	(1)	(2)	(3)	(4)	(5)	(6)
$p(\text{actual} \mid \mathcal{P}_4)$	-0.887^{***}	-0.806^{***}	-0.809^{***}			-0.735^{***}
	(0.031)	(0.090)	(0.090)			(0.116)
$\lvert\hat{s}_t\rvert$		0.104	0.104			0.108
		(0.089)	(0.089)			(0.091)
$\text{Skew}(\mathcal{P}_4)$			0.011			0.008
			(0.026)			(0.026)
$H(\mathcal{P}_4)$				0.632^{***}		0.209
				(0.044)		(0.187)
$\text{Var}(\mathcal{P}_4)$					0.606^{***}	-0.106
					(0.053)	(0.192)
R^2	0.773	0.776	0.776	0.400	0.368	0.784
N	242	242	242	242	242	242

Note: Dependent variable: LLM salience score $\mu_t \in [0, 1]$. All variables standardised to zero mean and unit variance; coefficients are in standard-deviation units. Columns (1)–(3) progressively add conceptually distinct dimensions: prior concentration, surprise magnitude, and asymmetry. Columns (4)–(5) show alternative concentration measures individually. Column (6) includes all five. HC3 robust standard errors. $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.10$.

Prior skewness contributes negligibly in column (3).

Columns (4) and (5) show that entropy and variance, alternative measures of distributional concentration, explain 40% and 37% individually. Both are significant univariate predictors but substantially weaker than p(actual), which directly captures how much probability mass the prior placed on what actually happened. When all five properties enter together (column 6), only p(actual) remains significant; the remaining predictors add less than one percentage point to R^2 beyond the parsimonious specification in column (1). The salience score is thus almost entirely a mechanical function of one quantity: the prior's concentration at the realized outcome. It does not capture an independent dimension of the monetary policy event.

C.2 Document-Only vs. Document-Plus-News Priors

Adding inter-meeting news reduces residual predictability without weakening forward-rate content. The $\mathcal{P}_4 \to \mathcal{P}_5$ comparison tests whether a broader information set improves expectation quality ex ante.

News extraction architecture. The baseline prior $\mathcal{P}_4$ conditions on the four Fed documents described in Section 2: Statement, press conference, Minutes, and Beige Book. The extended prior $\mathcal{P}_5$ adds inter-meeting news arriving between the Beige Book release and the FOMC decision, typically over the next 10–14 days. FactSet StreetAccount coverage begins in mid-2003;

Table 29: News-Stage ($\mathcal{P}_5$) Accuracy by Update Magnitude

		MAE (bp)		
$\mathcal{P}_5$ Update	N	$\mathcal{P}_4$	$\mathcal{P}_5$	Dir. Accuracy
No update ($\Delta = 0$)	57	1.0	1.0	—
Small ($0 < \Delta \leq 5$ bp)	77	3.9	3.5	67.5%
Medium ($5 < \Delta \leq 10$ bp)	9	13.4	9.2	77.8%
Large ($10 < \Delta \leq 20$ bp)	22	15.2	6.7	77.3%
Very large ($\Delta > 20$ bp)	13	31.0	11.2	92.3%
All meetings	178	6.8	3.9	72.5%

Note: Post-2004 sample ($N = 178$), restricted to meetings with FactSet StreetAccount coverage. $\Delta = |\mathbb{E}[\Delta i_t \mid \mathcal{P}_5] - \mathbb{E}[\Delta i_t \mid \mathcal{P}_4]|$ is the magnitude of the news-stage update. MAE is mean absolute error relative to the realized decision, in basis points. Directional accuracy is the fraction of updated meetings where $\mathcal{P}_5$ moves closer to the realized outcome. Filtration-quality diagnostics (AR(1) coefficients, variance reduction, and calibration regressions) are reported in Table 30 below.

for earlier meetings, $\mathcal{P}_5 = \mathcal{P}_4$.

I process articles through five domain modules in parallel. The *data releases* module records each indicator, the actual release, the consensus forecast, and the sign of the surprise. The *Fed communications* module records speeches and testimony, noting the speaker's voting status and any deviation from the most recent Statement. The *macroeconomic outlook* module distills press narratives on growth, inflation, and labor market conditions. The *financial conditions* module flags credit tightening, bank stress, energy shocks, and geopolitical disruptions. The *market expectations* module records dealer commentary, futures-implied policy paths, and consensus revisions over the inter-meeting window.

A synthesizer consolidates the five module reports into a single net-policy-signal assessment. The Forecaster then updates $\mathcal{P}_4 \to \mathcal{P}_5$ using the same narrative-first architecture as stages $\mathcal{P}_1$–$\mathcal{P}_4$: qualitative reasoning first, then a revised probability distribution. This extension requires five additional model calls per meeting.

Accuracy by update magnitude. Table 29 bins meetings by the size of the $\mathcal{P}_4 \to \mathcal{P}_5$ update and compares the mean absolute forecast errors of the two priors. The pattern is monotone: the larger the update, the larger the gain. On the heaviest revisions (above 20 bp), $\mathcal{P}_5$ moves toward the realized decision in nine cases out of ten and cuts mean absolute error roughly in three. Smaller updates contribute correspondingly less, and on meetings where the synthesizer judges the inter-meeting signal too weak to act on, $\mathcal{P}_5$ inherits $\mathcal{P}_4$ verbatim.

Table 30: Document-Only ($\mathcal{P}_4$) vs Document-Plus-News ($\mathcal{P}_5$) Surprises

	P4 (documents only)			P5 (documents + news)		
Test	$\hat{\beta}$ / R^2	s.e.	N	$\hat{\beta}$ / R^2	s.e.	N
Measurement quality	0.976^{***}	(0.119)	242	0.918^{***}	(0.160)	242
Predictability (B&S)	0.109^{**}	[0.014]	214	0.041	[0.248]	214
Serial correlation	0.051^{**}	[0.029]	238	0.015	[0.650]	238
Forward rate prediction	0.333^{**}	(0.141)	241	0.184	(0.154)	241
Forward own-surprise	-0.050	(0.098)	241	-0.059	(0.078)	241

Note: Comparison of surprise measures computed from the document-only prior ($\mathcal{P}_4$) and the document-plus-news prior ($\mathcal{P}_5$). Measurement quality regresses the actual rate change on the surprise; $\beta = 1$ under forecast efficiency. Predictability regresses each surprise on the six Bauer and Swanson (2023a) macro-financial predictors. Serial correlation regresses each surprise on four own lags. Forward rate prediction regresses the next meeting's rate change on the current surprise. Forward own-surprise regresses the next surprise on the current surprise; $\beta = 0$ for a true innovation. All regressions use Newey-West HAC standard errors (4 lags). F-test p-values in brackets for R^2 rows. ***, **, *: 1%, 5%, 10%.

Comparison of diagnostic properties. As Table 30 shows, incorporating inter-meeting news sharply improves the shock's orthogonality diagnostics. The slope coefficient remains close to one for both priors ($\hat{\beta}_{P4} = 0.976$, $\hat{\beta}_{P5} = 0.918$, each significant at the 1% level), with $\mathcal{P}_4$ statistically indistinguishable from unity and $\mathcal{P}_5$ within one standard error of it. Predictability from the Bauer and Swanson (2023a) macro-financial factors declines from 10.9% to 4.1% — the joint F-test against the predictors switches from significant at the 5% level to insignificant — and serial correlation in own lags falls from 5.1% to 1.5%, also switching from significant to insignificant. Forward self-prediction is near zero under either prior, consistent with both measures behaving approximately as innovations on this margin. Together, these reductions bring the $\mathcal{P}_5$ measure substantially closer to a true innovation process.

Forward-rate prediction attenuates under $\mathcal{P}_5$: the $\mathcal{P}_4$ surprise predicts the next meeting's rate change with $\hat{\beta} = 0.333$ ($p = 0.019$), while the $\mathcal{P}_5$ coefficient falls to 0.184 and loses significance. This is consistent with $\mathcal{P}_5$ absorbing the part of the rate-path content that markets price in during the inter-meeting window (formalized as Δ_t^{news} below), so that less of it remains in the meeting-eve residual; the inter-meeting revision then carries the forward-guidance signal more directly than the residual.

IRF macro-invariance with sharpened rate-path persistence. Re-estimating the headline 2SLP-IV specification of Section 6.1 with $\mathcal{P}_5$-based surprises in place of $\mathcal{P}_4$ delivers two complementary findings (Figure 36). First, the macroeconomic transmission is essentially unchanged: $\mathcal{P}_5$ point estimates lie within the $\mathcal{P}_4$ 68% bands at every horizon for real GDP, the PCE price index, industrial production, unemployment, and the 10-year yield, with horizon-by-

horizon differences in second-stage coefficients of 0.01–0.07pp. The IV identification recovers the same announcement-bundle response per unit shock under either prior. Second, the federal funds rate IRF (the instrumented variable) is materially more persistent under $\mathcal{P}_5$: the rate stays +0.4pp elevated through month 24 versus mean-reversion to zero under $\mathcal{P}_4$. The mechanism is that the news stage upgrades meetings whose rate move genuinely signals a continued trajectory and downgrades isolated moves — so $\mathcal{P}_5$ shocks carry more rate-path information per unit, even though the macro pass-through per unit is the same.[28]

Timing decomposition: blackout-window revision vs meeting-eve residual. The IV invariance hides a richer object. Using the timing of the pipeline directly, define two complementary shocks from the same documentary inputs:

$$\hat{s}_t^{P_5} = \Delta i_t - \mathbb{E}[\Delta i_t \mid \mathcal{P}_5], \qquad \Delta_t^{\text{news}} = \mathbb{E}[\Delta i_t \mid \mathcal{P}_5] - \mathbb{E}[\Delta i_t \mid \mathcal{P}_4] = \hat{s}_t^{P_4} - \hat{s}_t^{P_5}.$$

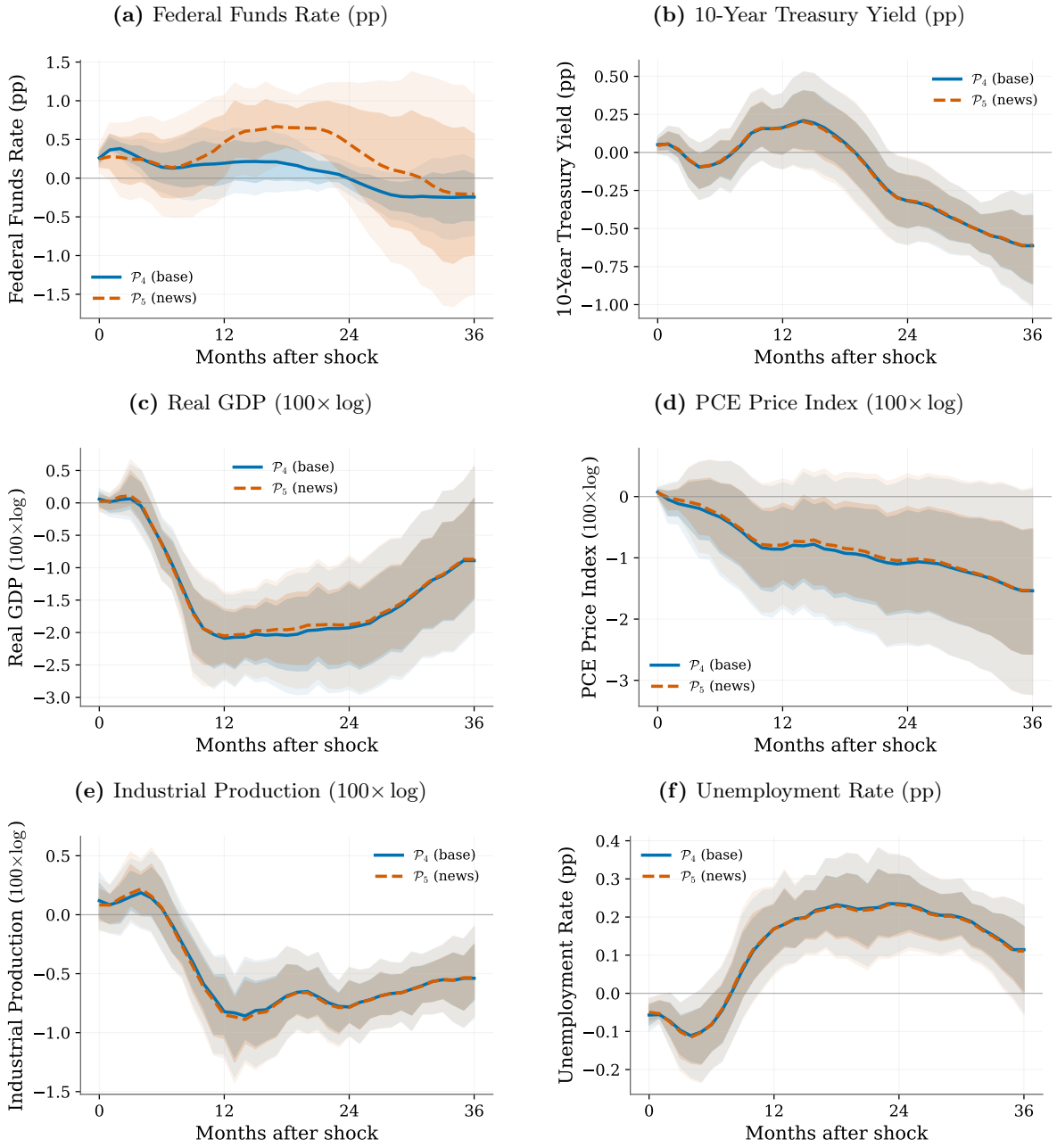
The first is the announcement-day surprise: the residual relative to the full pre-meeting information set including blackout-window news. The second is the documentary expectation revision over the pre-FOMC blackout window $[\mathcal{B}_t, \text{meeting eve}]$, when the Fed is officially silent and the news stage is processing data releases, market commentary, and press digestion of pre-blackout Fedspeak; the revision is therefore driven by *public news arriving during the blackout*, not by exogenous policy events. Joint OLS local projection, with both shocks entered together,

$$y_{t+h} = \alpha_h + \beta_h \, \hat{s}_t^{P_5} + \gamma_h \, \Delta_t^{\text{news}} + \text{controls} + \varepsilon_{t+h},$$

separates the macro response per unit of each. Figures 37 and 38 report the two coefficient series. The two components have distinct signatures. The meeting-eve residual $\hat{s}_t^{P_5}$ delivers correctly-signed contractionary responses: real GDP falls to roughly -0.4% by month 36 and the PCE price index falls to roughly -0.2%, the textbook pattern of a monetary-policy shock identified from the announcement residual after pre-meeting public information has been absorbed. The blackout-window revision Δ_t^{news} shows the opposite signature on activity: a hawkish revision is followed by real GDP rising to roughly $+1\%$, industrial production rising, and unemployment falling, while the PCE price index falls and the federal funds rate rises. The activity and inflation responses move in opposite directions in the way characteristic of an information shock — public

[28] Identical 2SLP-IV settings throughout (4 lags for outcomes and controls, ZLB excluded, COVID included, Newey-West HAC bandwidth $h + 1$).

Figure 36: Macroeconomic IRFs under $\mathcal{P}_4$ vs $\mathcal{P}_5$

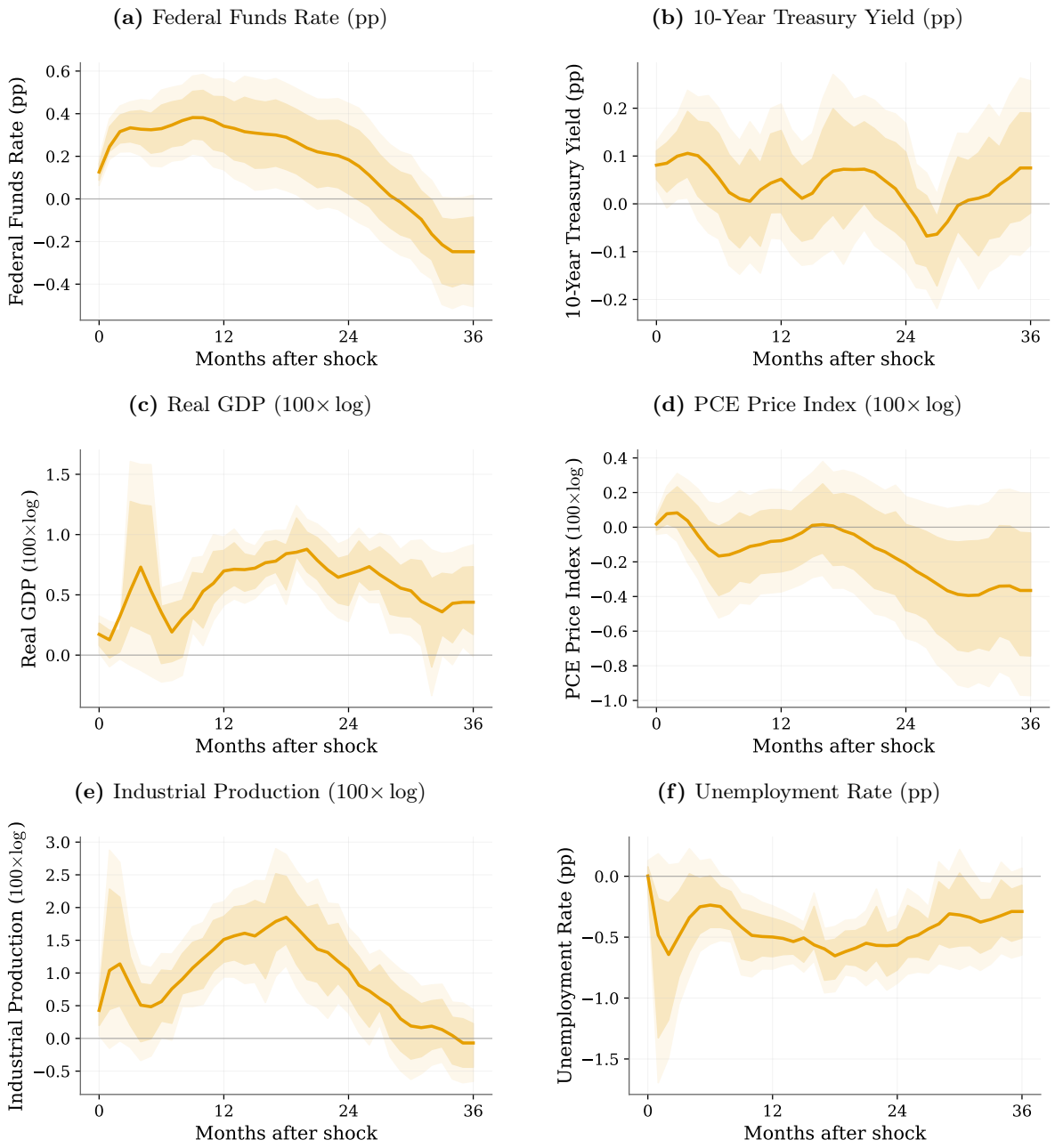


Note: Impulse responses to a 25 bp narrative surprise, instrumenting the federal funds rate. Solid blue: $\mathcal{P}_4$ (document-only prior); dashed vermilion: $\mathcal{P}_5$ (document-plus-news prior). Shaded areas are pointwise 68% (inner) and 90% (outer) HAC confidence bands of the unsmoothed coefficients; solid/dashed lines are 3-period moving averages of the point estimates. ZLB excluded, COVID included.

news about stronger fundamentals is jointly priced into expected activity, expected tightening, and (eventually) lower inflation as the tighter policy bites.

The economic reading is that the announcement-day residual $\hat{s}_t^{P_5}$ carries the cleaner monetary-policy-transmission signature, while the blackout-window revision Δ_t^{news} is an information-shock object: a hawkish public-news revision arrives bundled with stronger-

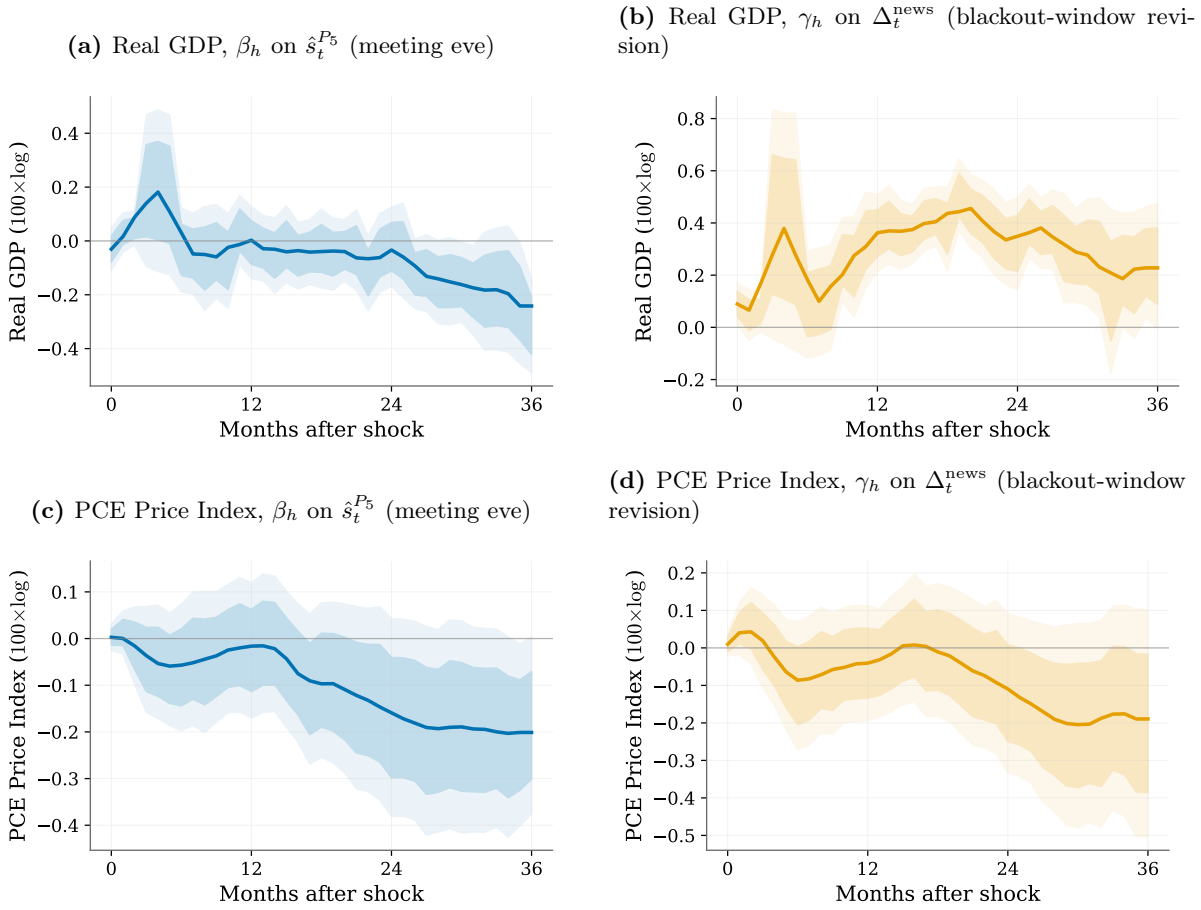
Figure 37: Macro responses to the blackout-window expectation revision Δ_t^{news}



Note: Coefficients γ_h from the joint OLS local projection $y_{t+h} = \alpha_h + \beta_h \hat{s}_t^{P_5} + \gamma_h \Delta_t^{\text{news}} + \text{controls} + \varepsilon_{t+h}$, plotting only the γ_h series and normalizing the impulse to a 25 bp hawkish revision in Δ_t^{news} ($\sigma \approx 14$ bp on the meeting-month subsample). Shaded areas: 68% (inner) and 90% (outer) Newey-West HAC confidence bands with bandwidth $h+1$; solid line is a 3-period moving average of the point estimates. Macro controls: 4 lags of FFR, unemployment, log PCE, log IP, log SP500, EBP. ZLB excluded, COVID included.

fundamentals news, so activity rises alongside expected tightening before inflation eventually cools. This recovers, from the pipeline's documentary timing alone, an analogue of the Jarociński and Karadi (2020) information-vs-policy decomposition without requiring asset-price sign restrictions. The implication for the main-text IRFs is that $\hat{s}_t^{P_4}$ works well as an instrument because the contractionary signal lives mainly in its announcement-residual component; the

Figure 38: Information-shock vs MP-transmission decomposition: GDP and PCE



Note: Side-by-side display of the two coefficients from the same joint LP, normalized to a 1σ shock in each regressor (computed on the meeting-month subsample where the shocks are non-zero: $\sigma(\hat{s}_t^{P_5}) \approx 13$ bp; $\sigma(\Delta_t^{\text{news}}) \approx 14$ bp). Left column: β_h on the meeting-eve residual $\hat{s}_t^{P_5}$, the announcement-day surprise relative to all pre-meeting public information (correctly-signed contractionary: GDP and PCE both fall). Right column: γ_h on the blackout-window expectation revision Δ_t^{news}, capturing how public news arriving during the pre-FOMC blackout moves expectations of the still-pending decision (information-shock pattern: GDP rises while PCE falls). Bands as in Figure 37.

blackout-window revision, taken alone, is not an appropriate object for instrumenting the federal funds rate in an MP-transmission study, though it remains an economically interesting measurement of how public news during the Fed's silence moves expectations of the upcoming decision.

Revision behavior. Of 178 post-2004 meetings with FactSet StreetAccount coverage, the news stage updates the prior in 122 (68.5%). Of those 122 revised meetings, updates are evenly split between amplifying (50%) and softening (50%) the $\mathcal{P}_4$ forecast, with direction reversals — where $\mathcal{P}_5$ moves opposite to $\mathcal{P}_4$'s sign — occurring in only 4.5% of all meetings. News therefore refines rather than contradicts the documentary pipeline. The LLM's self-assessed signal

strength is internally calibrated: meetings rated "negligible" produce zero revision, "moderate" an average $|\Delta|$ of 3.3 bp, and "strong" an average of 7.4 bp, a monotone relationship that holds without any ex post recalibration. The 35 largest revisions ($|\Delta| > 15$ bp) cluster visibly around identifiable economic turning points: the March 2020 COVID emergency cut (-39 bp), the December 2008 ZLB transition (-30 bp), the Lehman collapse (-26 bp), the December 2021 inflation-regime pivot ($+20$ bp), and the December 2015 liftoff ($+21$ bp).

Taken together, expanding the information set from $\mathcal{B}_t$ (Fed documents) to $\mathcal{B}_t \cup \mathcal{N}_t$ (documents plus news) narrows the gap between the document-only conditioning set and the broader public information set without materially weakening the core shock signal s_t. In the notation of Section 3, $\mathcal{P}_5$ moves the forecast from a document-conditioned prior closer to $\mathcal{M}_t$; it does not close the Fed-private-information gap in $\mathcal{G}_t \setminus \mathcal{M}_t$. The $\mathcal{P}_5$ construction thus addresses the same predictability concerns as Bauer and Swanson (2023a)-style ex post cleaning, but does so ex ante: it imposes no time-invariant coefficient assumption and leaves historical shocks unchanged as the sample grows. More generally, the remaining predictability need not be interpreted as a lower bound. The $\mathcal{P}_4 \to \mathcal{P}_5$ improvement itself illustrates this margin: simply conditioning on inter-meeting news reduced Bauer and Swanson (2023a) predictability by roughly one-fifth and serial correlation by over one-quarter. Further gains from more accurate expectation extraction, whether through iterative elicitation, cross-model aggregation, or richer real-time information sets, remain available.

Box 2: News Stage in Action: COVID Emergency Cut (March 2020) and May 2022 Hike

March 2020 — $\mathcal{P}_4 = -10$ **bp** $\rightarrow$ $\mathcal{P}_5 = -49$ **bp** **(actual:** -50 **bp)**

Synthesizer — net signal assessment

> The market consensus for a 50 bp cut is overwhelming and justified. While recent data (e.g., +225K payrolls, ISM manufacturing expansion) and some Fed rhetoric suggest stability, the coronavirus shock has fundamentally altered the landscape. Financial conditions show systemic stress -- yields plunging to record lows, oil crashing 16%, and corporate earnings downgrades -- indicating markets are pricing in severe economic damage. Powell's 'act as appropriate' is the only signal that matters -- it implicitly endorses market expectations. Pre-coronavirus data is irrelevant now; the Fed will cut aggressively to preempt a recession and stabilize markets.

Forecaster — $\mathcal{P}_5$ justification

> The coronavirus-driven deterioration in financial conditions and global growth outlook has forced a pivot to aggressive easing, making a 50 bp cut almost certain despite earlier Fed hesitancy and strong pre-virus data.

May 2022 — $\mathcal{P}_4 = +29$ **bp** $\rightarrow$ $\mathcal{P}_5 = +50$ **bp** **(actual:** $+50$ **bp)**

Synthesizer — net signal assessment

> The consensus is unequivocal: a 50 bp hike in May is nearly certain, anchored by Powell's explicit guidance and market pricing. Inflation persistence -- with 10-year breakevens at two-decade highs -- and the global hawkish pivot create a compelling narrative for aggressive action. The hawkish surprise in the Employment Cost Index (+1.4% vs. +1.1% consensus) is the standout, confirming wage-driven inflation pressures that the Fed cannot ignore. No domain meaningfully challenges the 50 bp hike expectation; risks are skewed toward even more aggressive future hikes if inflation does not abate.

Forecaster — $\mathcal{P}_5$ justification

> Powell's clear 50 bp hike signal, reinforced by strong wage inflation data and broad market alignment, overrides minor growth softness, making a 50 bp hike nearly certain.

Note: Both episodes illustrate the news stage identifying decisive information that arrived after the Beige Book but before the FOMC decision. In March 2020, the coronavirus shock rendered the documentary prior of -10 bp effectively obsolete; $\mathcal{P}_5$ moved to -49 bp, within 1 bp of the actual emergency cut. In May 2022, Powell's inter-meeting signaling and the ECI wage surprise closed a 21 bp gap between $\mathcal{P}_4$ and the realized 50 bp hike.

D Empirical Applications

This appendix collects the two empirical applications of the surprise measure: macroeconomic and financial transmission (Section D.1) and a yield-curve trading strategy (Section D.2).

D.1 Impulse Responses

D.1.1 First-Stage Diagnostics

The 2SLP-IV specification consists of two regressions at each horizon h, repeating the structure of equations (1)–(2) in Section 6.1. The first stage projects the policy variable on the narrative surprise:

$$P_{i,t} = \alpha + \pi \cdot \hat{s}_t + \sum_{\ell=1}^{4} \boldsymbol{\beta}_\ell' \mathbf{Z}_{t-\ell} + u_t, \tag{31}$$

where $P_{i,t}$ is the instrumented policy variable (federal funds rate or two-year Treasury yield), $\hat{s}_t$ is the narrative surprise (instrument), and $\mathbf{Z}_{t-\ell}$ stacks four lags of the policy variable itself together with the macro control vector (unemployment, log industrial production, log PCE price index, S&P 500, excess bond premium). The second stage uses the fitted policy variable to estimate the impulse response,

$$y_{t+h} = \alpha_h + \beta_h \cdot \widehat{P_{i,t}} + \sum_{\ell=1}^{4} \boldsymbol{\gamma}_{h,\ell}' \mathbf{W}_{t-\ell} + \varepsilon_{t+h}, \tag{32}$$

with $\mathbf{W}_{t-\ell}$ the same macro control vector and four lags of the outcome y, and lags of the endogenous policy variable omitted (cf. Section 6.1). The first-stage F-statistic tests $H_0 : \pi = 0$ from the HAC-robust t-statistic. Under heteroskedasticity and autocorrelation, the formally correct weak-instrument benchmark is the heteroskedasticity- and autocorrelation-robust effective F of Montiel Olea and Pflueger (2013); the conventional rule of thumb $F > 10$ remains a useful coarse benchmark but is not derived under the inferential conditions used here.

Table 31 documents instrument strength across policy variables (federal funds rate and Treasury yields of varying maturities) and sample specifications (excluding versus including ZLB periods). The raw-impact column reports the un-normalized first-stage coefficient: a one-unit increase in the narrative surprise shifts the two-year Treasury yield by 50.75 basis points in the full sample. F-statistics comfortably exceed the rule-of-thumb benchmark of 10 across all specifications, including the 5-year yield in the ZLB-excluded sample where the smaller sample reduces power.

Instrument strength is stable rather than sample-dependent. For the two-year yield, the first-stage coefficient shifts by less than one percent between samples (51.09 vs. 50.75 bp); the higher F-statistics in the full sample reflect increased statistical power, not a more informative instrument. Across the 36 estimation horizons ($h = 0, \ldots, 35$), the first-stage point estimate

Table 31: First-Stage Diagnostics for 2SLP-IV Specifications

Instrument	Sample	Raw Impact (bp)	F-Statistic	N
Federal Funds Rate	No ZLB	49.91	21.08	171
	Full	48.19	25.34	248
1-Year Treasury Yield	No ZLB	55.68	38.35	171
	Full	54.34	53.91	248
2-Year Treasury Yield	No ZLB	51.09	28.41	171
	Full	50.75	47.53	248
5-Year Treasury Yield	No ZLB	34.84	10.63	171
	Full	37.19	22.98	248

Notes: First-stage diagnostics from two-stage local projections (2SLP-IV), with narrative surprises instrumenting the policy variable. *Raw Impact*: horizon-zero effect of a unit narrative surprise on the policy variable (bp). Reported IRFs (not shown here) rescale all shocks to a 25 bp move in the instrumented variable. F-statistics exceeding 10 indicate strong instruments (Stock & Yogo, 2005). *No ZLB* excludes Dec 2008–Dec 2015 and Mar 2020–Mar 2022; *Full* covers 1996–2025. Controls: 4 lags of the instrumented variable and of FFR, unemployment, industrial production, PCE, S&P 500, and the excess bond premium.

is stable; the horizon-by-horizon variation in F reflects the change in HAC bandwidth (set to $h+1$) and the small differences in the estimation sample induced by horizon-specific outcome availability rather than a structurally horizon-specific first stage.

D.1.2 LP Specification Robustness

This subsection documents three robustness exercises for the headline 2SLP-IV specification: the Jordà and Taylor (2025) long-difference variant (the primary small-sample-robust check), the relationship between the IV scaling and a direct reduced-form local projection on the narrative surprise, and the role of lagged shocks as included exogenous controls in both stages.

Jordà (2025) Long-Difference Specification. Jordà and Taylor (2025) (§3) shows that levels local projections inherit a small-sample bias of order $O(T^{-1})$ when the regressor is highly persistent, and recommends the long-difference specification $y_{t+h} - y_{t-1} = \alpha + \beta_h \hat{s}_t + \gamma_h \Delta y_{t-1} + u_{t+h}$ as the preferred small-sample-robust alternative because long differencing largely suppresses this bias even when $|\rho| \to 1$. The federal funds rate in the present LP sample has $\hat{\rho} \approx 0.99$, placing the headline levels specification squarely in the regime where Jordà's bias correction matters.

I therefore re-estimate the baseline 2SLP-IV in long differences, again instrumenting the federal funds rate with the narrative surprise. Figure 39 overlays the long-difference IRFs against the

levels headline across the same six outcomes shown in Figure 10. Point estimates are essentially identical: at $h = 12$ real GDP responds -2.04pp under the Jordà specification versus -2.08pp under levels, PCE -0.82pp versus -0.80pp, industrial production -0.70pp versus -0.76pp, unemployment $+0.21$pp versus $+0.17$pp, and the 10-year yield is within 0.04pp. Sign and timing of peaks and troughs match across all six outcomes, and confidence bands are comparable. The headline narrative survives the small-sample-robust specification with virtually no change in magnitudes, providing direct evidence that the levels-LP small-sample bias documented by Jordà and Taylor (2025) is not driving the headline IRFs in this sample.

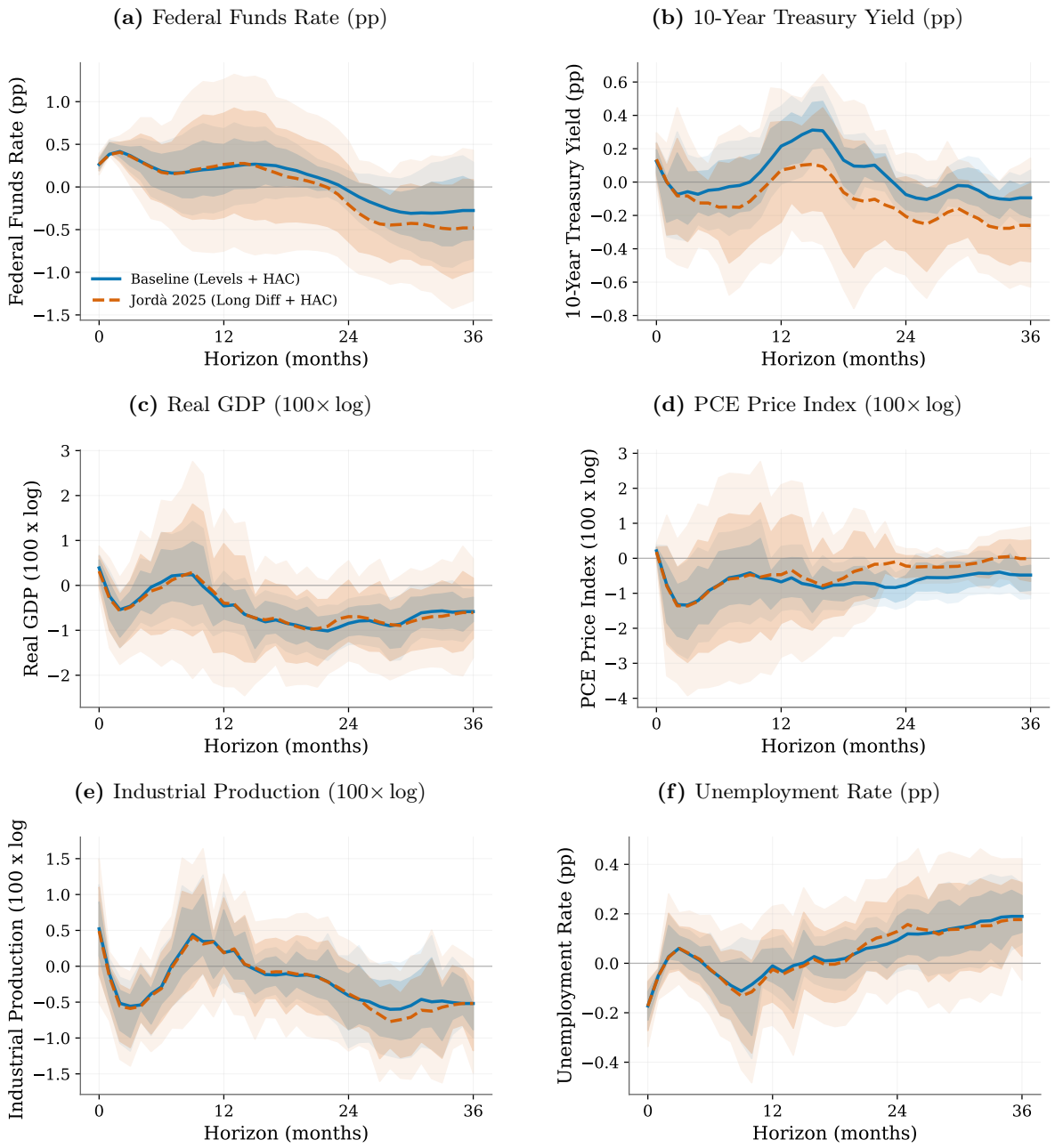
Reduced-Form vs IV Scaling. A direct reduced-form local projection regresses the outcome on the narrative surprise itself,

$$y_{t+h} = a_h + \theta_h \, \hat{s}_t + \text{controls}_t + v_{t+h}, \tag{33}$$

delivering an IRF in shock units. Under just-identified IV with symmetric exogenous controls in the first and second stages, the rescaling identity $\beta_h = \theta_h / \pi_h$ holds exactly, where π_h is the first-stage coefficient on $\hat{s}_t$. The headline specification breaks symmetry by retaining federal funds rate lags in the first stage but not the second; for the small-sample-bias reasons set out in Section 6.1 and documented numerically below, including FFR lags symmetrically in both stages of a levels LP-IV is over-controlled in this sample. Figure 40 overlays β_h from the headline 2SLP-IV against the rescaled $\theta_h / \hat{\pi}_h$, with $\hat{\pi}_h$ taken from the first stage at horizon h.

The qualitative shape of the response matches across all six outcomes: the disinflation in PCE and industrial production, the gradual rise in unemployment, the initial flattening and subsequent steepening of the 10Y–3M spread, and the muted long-yield response are all present in both series. The numerical levels differ for two structural reasons. First, the IV second stage drops federal funds rate lags while the reduced-form retains them through the macro control set, breaking the symmetric-controls condition the algebraic equivalence requires. Second, the IV second stage treats $\widehat{\text{FFR}}_t$ as an observed regressor, leaving first-stage estimation noise in the IV residual that no rescaling of θ_h by $\hat{\pi}_h$ can replicate; Montiel Olea and Plagborg-Møller (2021) provide the formally correct LP-IV inference for the symmetric case. The IV scaling is reported as the headline specification for direct comparability with rate-shock IRFs in the literature (Gertler and Karadi, 2015, Aruoba and Drechsel, 2024); Figure 40 confirms that the qualitative conclusions do not depend on the IV step. Combined with the Jordà long-difference

Figure 39: IRF Robustness: Jordà 2025 Long-Difference Specification

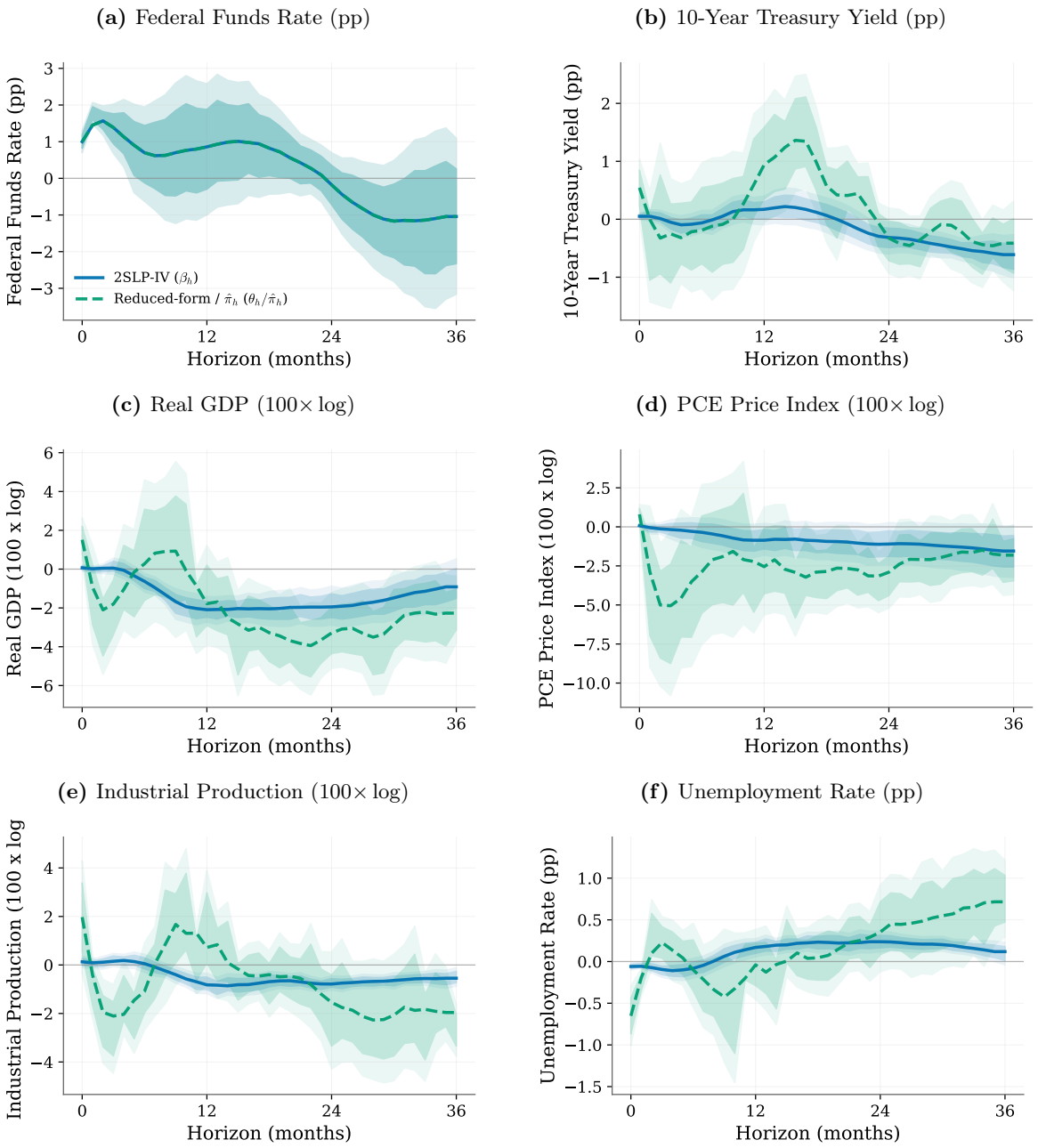


Note: Impulse response functions to a 25 bp narrative surprise. Blue: levels 2SLP-IV (headline, Figure 10). Red: Jordà and Taylor (2025) long-difference 2SLP-IV with the lagged first difference as the persistent control. Both specifications instrument the federal funds rate with the narrative surprise; Newey-West HAC bandwidth $h + 1$; ZLB excluded, COVID included. Sample: 171 meeting-month observations (1996–2025). Shaded areas: 68% (inner) and 90% (outer) pointwise HAC confidence bands.

replication of Figure 39, this leaves the headline transmission story standing on three independent specifications.

Shock Lags. The baseline specification omits lags of $\hat{s}_t$ from both stages, following the Aruoba and Drechsel (2024) and Romer and Romer (2004) narrative-instrument convention. Stock

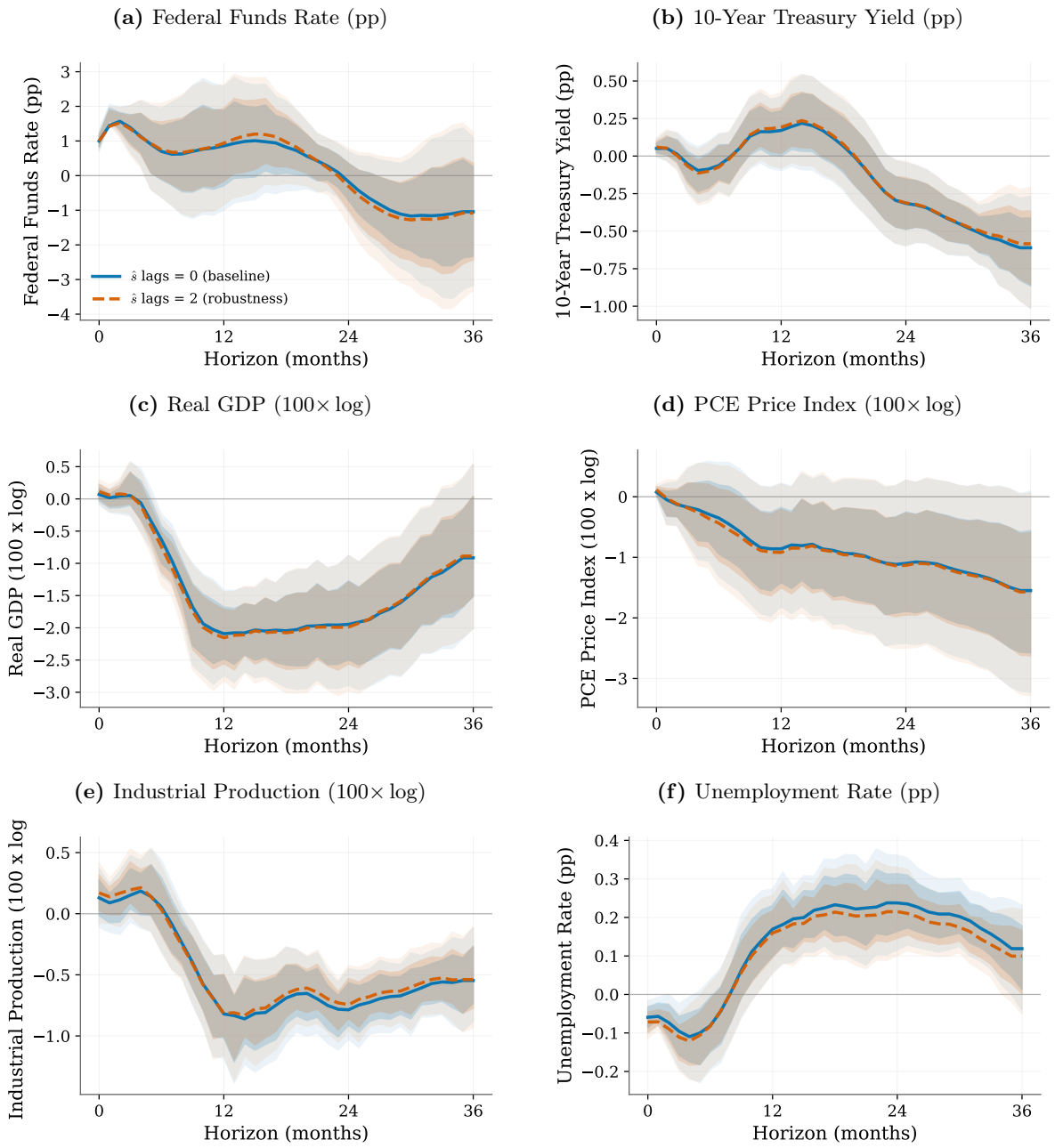
Figure 40: IRF Robustness: Reduced-Form vs IV Scaling



Note: Blue: 2SLP-IV impulse responses β_h to a 25 bp narrative surprise, as in Figure 10. Green dashed: reduced-form coefficient θ_h from (33), rescaled by the first-stage coefficient $\hat{\pi}_h$. The equivalence $\beta_h = \theta_h/\pi_h$ is exact for $y_{t+h} = \text{FFR}_{t+h}$ and qualitative otherwise: the second-stage exclusion of FFR lags breaks control symmetry between the two stages, and treating $\widehat{\text{FFR}}_t$ as observed in the second stage adds generated-regressor noise that the rescaling does not undo. Shapes track each other (sign and timing of peaks); levels diverge. Bands: 68% (inner) and 90% (outer) pointwise HAC. Sample: 171 meeting-month observations, 1996–2025, ZLB excluded.

and Watson (2018) and Plagborg-Møller and Wolf (2021) note that, under the identification assumption already imposed, instrument lags are not required for consistency but can absorb residual serial correlation in ε_{t+h} at long horizons. Figure 41 compares impulse responses with $\hat{s}_t$ lags set to zero (baseline) versus two lags entered as included exogenous (predetermined)

Figure 41: IRF Robustness: Shock-Lag Specification



Note: 2SLP-IV impulse responses to a 25 bp narrative surprise, with $\hat{s}_t$ lags set to zero (blue) versus two lags as included exogenous controls in both stages (red). Specification otherwise as in Figure 10: 4 lags, Newey-West HAC (bandwidth $h+1$), 68% (inner) and 90% (outer) pointwise HAC confidence bands, ZLB excluded, COVID included.

controls in both stages, with the contemporaneous shock retained as the excluded instrument.

Point estimates track each other closely across all six outcomes, with the largest gap of 0.33 pp on the federal funds rate (range approximately ±1.8 pp, so under 20% of the IRF range) and at most 0.08 pp on every other outcome. Sign, timing of peaks and troughs, and qualitative conclusions are identical across specifications. The headline IRFs are therefore robust to the

lag-augmentation efficiency adjustment.

D.1.3 Zero Lower Bound Robustness

The headline IRFs exclude the conventional ZLB period (Dec 2008–Dec 2015) because the federal funds rate cannot move when constrained near zero, blunting an instrument that targets that move. This subsection asks how much of the result is sample-driven by re-estimating two variants on the full 1996–2025 window: the headline FFR-instrumented specification (which simply ignores the ZLB constraint) and a Gertler–Karadi-style specification that instruments the two-year Treasury yield rather than the policy rate. Following Gertler and Karadi (2015), the two-year yield responds to forward guidance signals even when the FFR is constrained, making it a more suitable policy indicator during unconventional regimes. An alternative approach substitutes the Wu–Xia shadow rate into the R&R Greenbook regression (Bügel et al., 2026); I instead maintain the observed policy rate while instrumenting the two-year yield, which avoids dependence on a shadow-rate model and preserves real-time implementability.

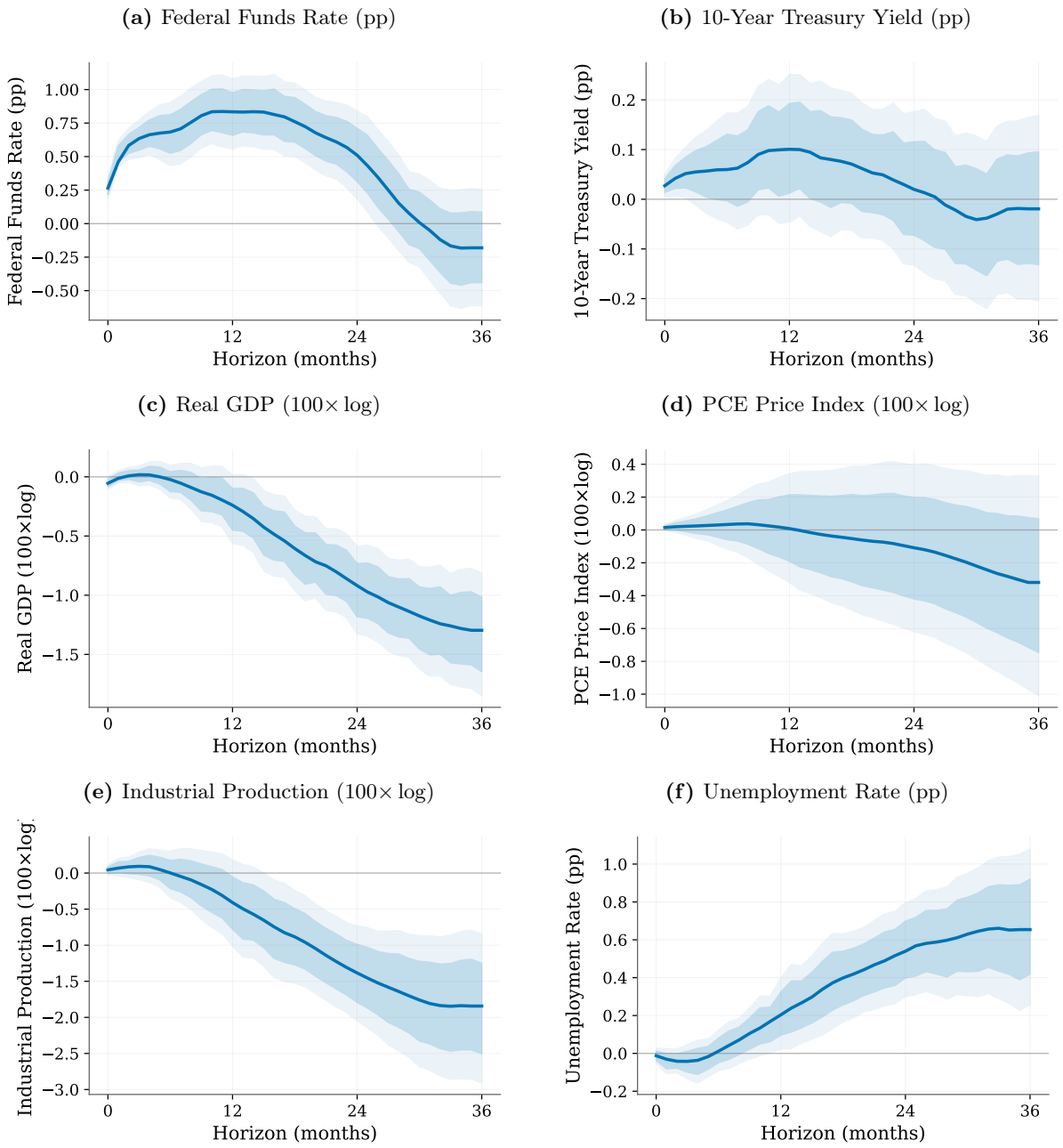
The 2-year-yield specification produces stable estimates with strong first-stage identification (Table 31): $F = 47.53$ in the full sample versus $F = 28.41$ in the ZLB-excluded sample, both well above the rule-of-thumb threshold of 10. The first-stage coefficient on the 2-year yield shifts by less than one percent between samples (51.09 vs. 50.75 bp), confirming that the narrative surprise predicts the yield equally well during unconventional policy. The contrast with residual-based measures is sharp: when the FFR cannot move, regression-residual shocks inflate mechanically, producing volatile residuals even when actual policy is stable. The narrative measure avoids this degradation because it is extracted from documents rather than backed out from regression residuals.

Across both specifications, contractionary surprises generate the expected directions: real GDP declines, industrial production contracts, the PCE price index falls, unemployment rises, and term spreads compress on impact and steepen at longer horizons. The qualitative transmission patterns remain consistent with the ZLB-excluded headline; including ZLB observations shifts magnitudes modestly but does not flip signs or alter the timing of peak responses.

D.1.4 Bauer–Swanson Purging Robustness

Bauer and Swanson (2023a) show that high-frequency monetary policy surprises are significantly predicted by pre-meeting macroeconomic and financial variables, raising concerns that appar-

Figure 42: ZLB Robustness: IV Federal Funds Rate, Full Sample

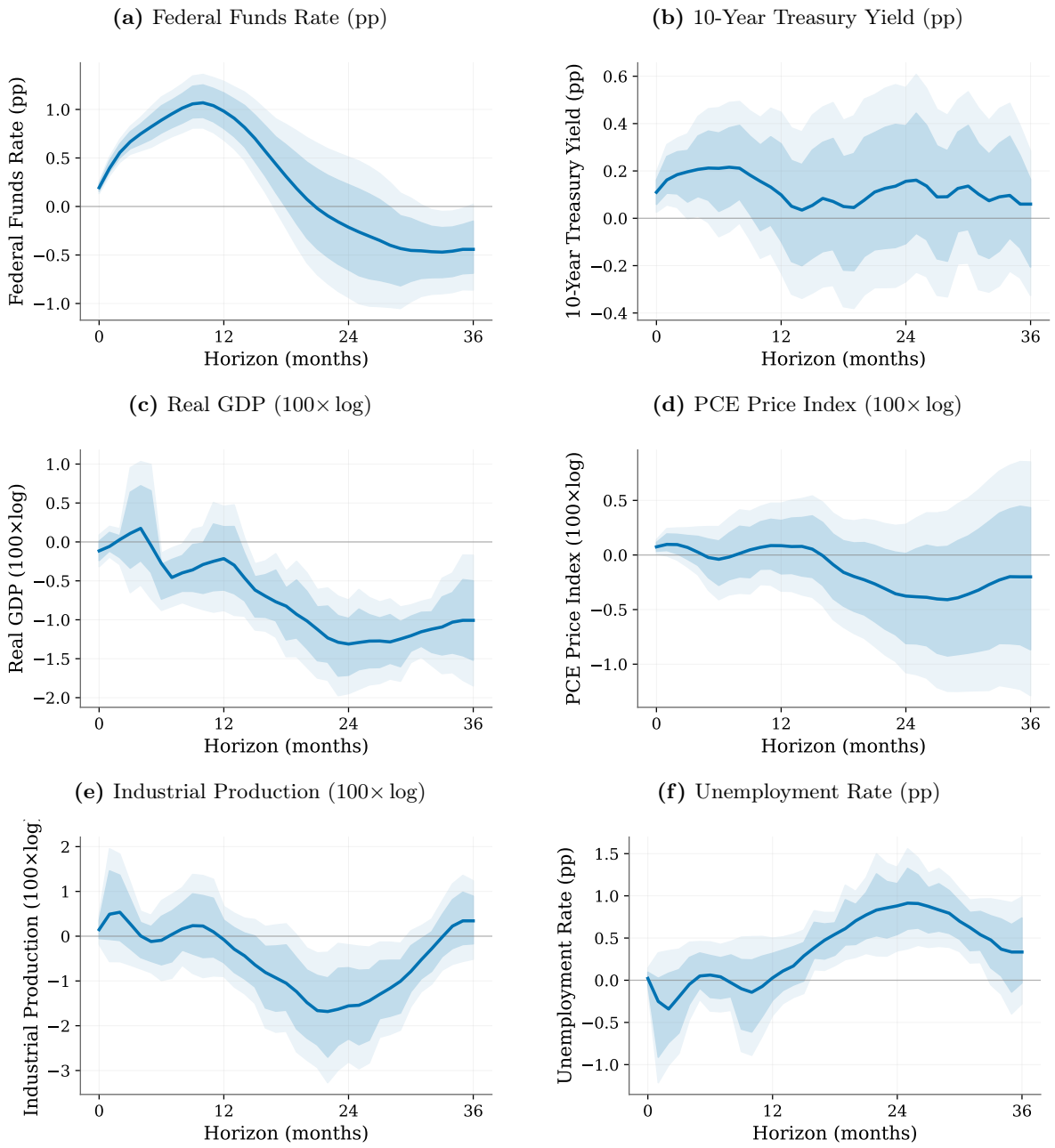


Note: Impulse response functions to a 25 bp contractionary monetary policy surprise on the full sample including the ZLB period (Dec 2008–Dec 2015). Instrument: federal funds rate. Two-stage local projections (2SLP-IV), 4 lags, 0 shock lags. Controls: unemployment, log PCE Price Index, log industrial production, S&P 500, EBP. Newey-West HAC standard errors, bandwidth $h+1$. Shaded areas: 68% (inner) and 90% (outer) pointwise HAC confidence bands. Sample: 248 meeting-month observations (1996–2025). Horizons in months.

ent "shocks" may partly reflect forecastable policy actions. Section 5.2 documents moderate predictability ($R^2 = 0.166$) of my narrative surprise from the same predictor set on the full predictability sample. If this predictable component drives the impulse response results, purging it should materially alter the estimated responses.

I residualize the narrative surprise on the six Bauer and Swanson (2023a) predictors and

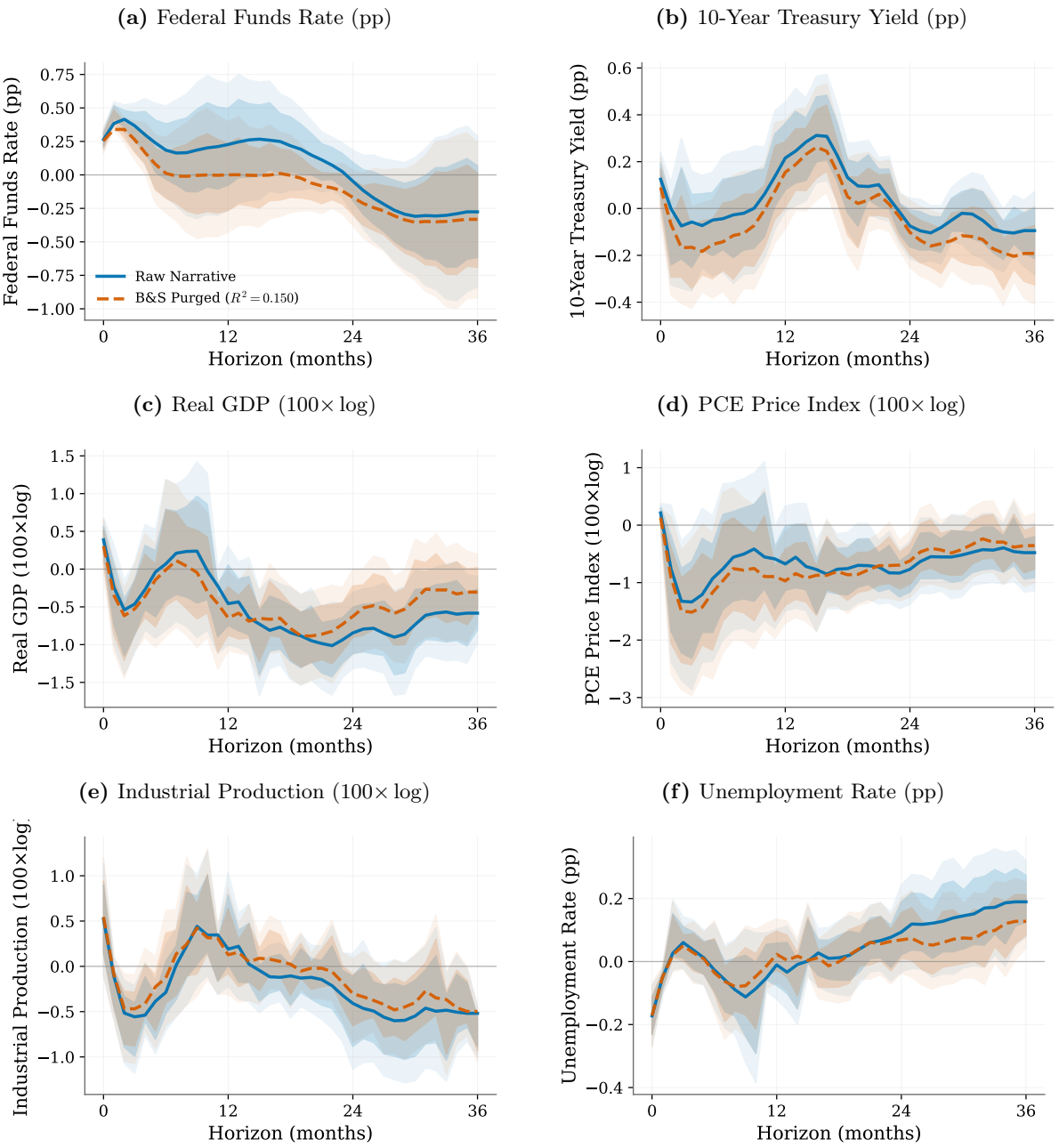
Figure 43: ZLB Robustness: IV Two-Year Yield, Full Sample



Note: As Figure 42, but instrumenting the two-year Treasury yield rather than the federal funds rate, following Gertler and Karadi (2015). Instrumenting the two-year yield preserves identification when the policy rate is constrained near zero, since forward guidance shifts the two-year yield even during ZLB episodes. Sample: 248 meeting-month observations.

re-estimate the baseline 2SLP-IV specification using only the unpredictable component. On the IRF subsample ($N = 153$), the predictor set explains 15.0% of surprise variance. Figure 44 compares impulse responses using the raw and purged surprises. Point estimates shift modestly across outcomes (typical $\Delta\beta$ of 0.1–0.3pp), but signs and the qualitative shape of the IRFs are preserved: the price index still falls, output and industrial production still contract, the term-

Figure 44: IRF Robustness: Raw vs Bauer–Swanson Purged Narrative Surprise



Note: Impulse response functions to a 25 bp monetary policy surprise. Blue: baseline narrative surprise. Red: narrative surprise residualized on six Bauer and Swanson (2023a) predictors (purging $R^2 = 0.150$, $N = 153$). 2SLP-IV instrumenting the federal funds rate, 4 lags, Newey-West HAC (bandwidth $h + 1$). Shaded areas: 68% (inner) and 90% (outer) pointwise HAC confidence bands. ZLB excluded.

structure response remains hump-shaped. The predictable component of the narrative surprise does not drive the impulse-response results.

D.1.5 Pre-Crisis Comparisons with Established Benchmarks

I compare the narrative surprise with established measures on the common sample 1996:01–2008:10 (154 months), restricted by LLM availability. This imposes a cost on the benchmarks: Aruoba and Drechsel (2024)'s shock was estimated on 1984–2016 and Romer and Romer (2004)'s on 1969–1996; restricting both to 154 months of the Great Moderation understates their identification power. These comparisons should therefore be read as a check of directional consistency, not a horse race.

For each shock measure s_t^j, I estimate OLS local projections

$$y_{t+h} = \alpha_h + \beta_h \, s_t^j + \sum_{\ell=1}^{2} \boldsymbol{\gamma}_\ell' \, \mathbf{X}_{t-\ell} + \boldsymbol{\delta}' \, \mathbf{X}_t + \varepsilon_{t+h}, \tag{34}$$

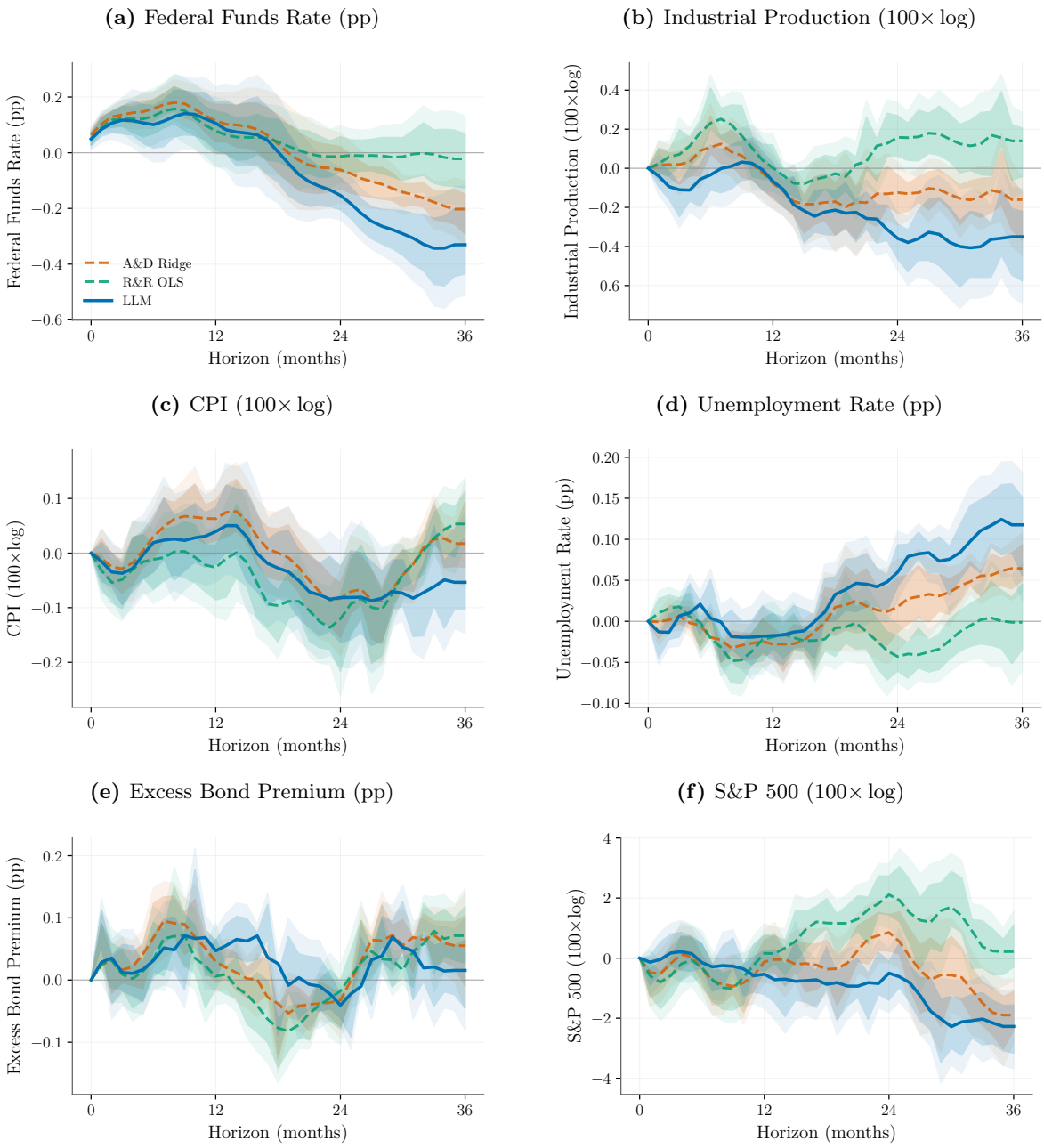
where $\mathbf{X} = (\text{IP}, \text{CPI}, \text{UE}, \text{EBP}, \text{SP500}, \text{FFR}, \text{VIX})$ and standard errors use Newey-West HAC with bandwidth $h + 1$. Two specification choices reflect the short sample. First, lag length is reduced from the four used by Aruoba and Drechsel (2024) to two, halving the parameter count relative to 154 observations. Second, responses are normalised to a one-standard-deviation shock of each measure rather than to a fixed FFR impact; a one-SD shock corresponds to approximately 10 basis points of FFR variation for all series, providing an economically grounded common scale that avoids the Wald-ratio instability inherent in normalising by a weak first stage.[29]

Text-Based Narrative Methods. Figure 45 compares my measure with the ridge-regularised textual analysis of Aruoba and Drechsel (2024) and the Greenbook-residualised shocks of Romer and Romer (2004). The directions of response are consistent across measures: industrial production and unemployment respond contractionarily, and the CPI shows the standard mild positive hump at intermediate horizons that is common in short, low-inflation samples. Magnitudes for the LLM measure are comparable to or slightly larger than the benchmarks on this 154-month window; the A&D and R&R responses for real variables are muted and statistically weaker, consistent with the loss of identifying power when their longer estimation samples are truncated to the Great Moderation.

Cleaned Market-Based Surprises. Figure 46 compares my measure with *ex post* cleaned market-based approaches: Bauer and Swanson (2023b)'s orthogonalized high-frequency surprises

[29] The pre-crisis sample yields weak first-stage F-statistics for all measures, making IV estimation unreliable. OLS with one-SD normalisation provides a transparent reduced-form comparison.

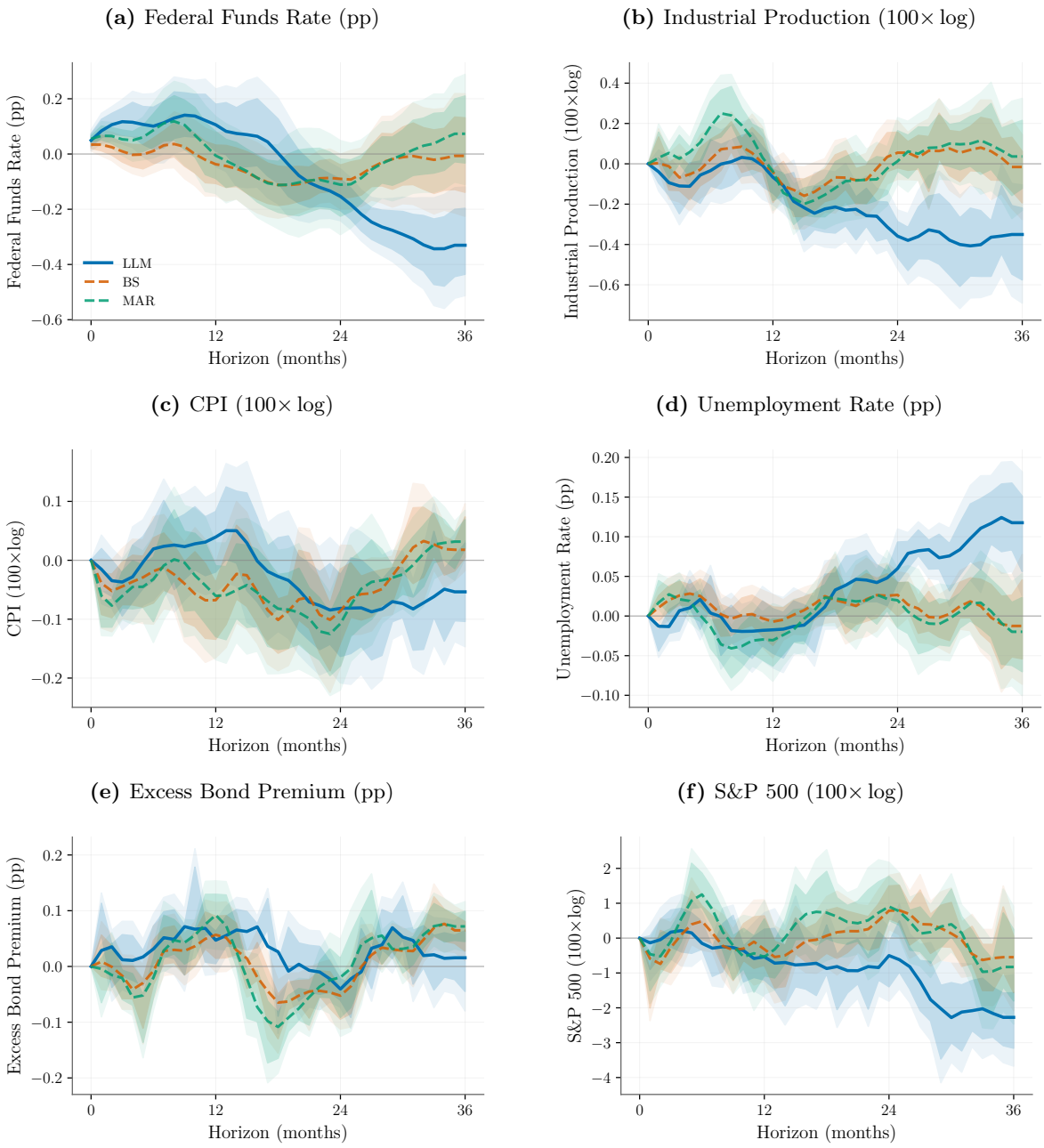
Figure 45: IRF Comparison: Text-Based Narrative Methods (1996–2008)



Note: Impulse response functions to a one-standard-deviation shock of each measure on the *common* sample 1996:01–2008:10, restricted by LLM availability. A one-SD shock corresponds to approximately 10 bp of FFR variation for all three series. A&D Ridge was designed for 1984–2016 and R&R OLS for 1969–1996; both are evaluated here on an unrepresentative 154-month subsample, so comparisons indicate directional consistency rather than relative power. OLS local projections, 2 lags of log-level controls (IP, CPI, UE, EBP, SP500, FFR, VIX), recursive identification, 0 shock lags. The 2001-09-17 emergency intermeeting cut is excluded. Newey-West HAC, bandwidth $h + 1$. Shaded areas: 68% (inner) and 90% (outer) pointwise HAC confidence bands.

(BS) and Miranda-Agrippino and Ricco (2021)'s monetary policy indicator (MAR). BS produces muted responses with wider confidence bands on this subsample, indicating weaker identification power on the truncated window. MAR displays wide uncertainty bands for unemployment and the S&P 500, the expected consequence of translating their VAR-estimated shocks into this LP

Figure 46: IRF Comparison: Cleaned Market Surprises (1996–2008)



Note: Impulse response functions to a one-standard-deviation shock of each measure on the *common* sample 1996:01–2008:10. BS (Bauer & Swanson, 2023b) and MAR (Miranda-Agrippino & Ricco, 2021) cover longer periods but are restricted here to the LLM-available window; comparisons indicate directional consistency. Specification identical to Figure 45.

specification. Both comparisons are consistent with the directional reading that direct narrative extraction generates contractionary responses without the additional *ex post* cleaning step those measures require, while remaining real-time implementable.

Table 32: GSS Target/Path Factor Decomposition of the Narrative Surprise

	Dependent variable: Narrative surprise				
	(1)	(2)	(3)	(4)	(5)
Target ($\hat{\beta}_T$)	0.0466***		0.0467***		
	(0.0115)		(0.0117)		
Path ($\hat{\beta}_P$)		-0.0024	-0.0031		
		(0.0080)	(0.0085)		
ED1				0.8320***	
				(0.1943)	
ED4					0.5005***
					(0.1668)
R^2	0.1234	0.0003	0.1239	0.1169	0.0608
N	218	218	218	217	218

Note: OLS regressions of the narrative surprise on Gürkaynak et al. (2005) target and path factors and on the two raw Eurodollar surprises (ED1, ED4) from which the path factor is constructed. Target and path factors built via PCA rotation of FF4 and ED1–ED4 at FOMC announcement frequency. ***, **, and * denote significance at the 1%, 5%, and 10% levels. $N = 218$ FOMC meetings with available high-frequency data; ED1 column has $N = 217$ owing to one missing observation.

D.1.6 GSS Target/Path Factor Decomposition

Gürkaynak et al. (2005) decompose FOMC announcement effects into a target factor (the current rate surprise) and a path factor (revisions to expected future policy), constructed via PCA rotation of FF4 and ED1–ED4 surprises. I regress the narrative surprise on these factors to characterize its position in the high-frequency factor space.

Table 32 reports the results. The narrative surprise loads significantly on the target factor ($\hat{\beta}_T = 0.047$, significant at the 1 percent level) but is orthogonal to the path factor ($\hat{\beta}_P \approx 0$, statistically indistinguishable from zero). Adding path to the target-only regression leaves R^2 essentially unchanged (12.4% versus 12.3%). Columns 4–5 show that the narrative surprise also covaries significantly with the two raw Eurodollar surprises that underlie the path factor: ED1 (the near-term rate surprise) and ED4 (the four-quarter-ahead surprise) explain 11.7% and 6.1% of narrative-surprise variance respectively, both with $p < 0.01$. The pattern — strong loading on the short end and on the orthogonalized target factor, near-zero loading on the orthogonalized path factor — is consistent with the narrative surprise tracking current-decision content rather than the path-factor revisions to expected future policy.

The target loading is expected: the narrative surprise is defined as the actual policy decision minus $\mathbb{E}[\Delta r \mid \mathcal{P}_4]$, so it shares variation with the current-meeting rate surprise to the extent that

Table 33: J&K Decomposition: Pairwise and Sub-Sample Robustness

Measure	Sample	N	$\hat{\beta}_{\text{MP}}$	$\hat{\beta}_{\text{CBI}}$	p(joint)	R^2
Panel A: pairwise common sample (B&S vs. LLM)						
B&S	1996–2023	220	0.749***	0.593***	0.000	0.620
			(0.128)	(0.191)		
LLM	1996–2023	220	0.840**	0.904	0.377	0.110
			(0.360)	(0.654)		
Panel B: LLM pre/post financial crisis						
LLM	1996–2007	99	0.467**	0.982*	0.013	0.094
			(0.216)	(0.537)		
LLM	2008–2024	122	1.787**	0.549	0.510	0.190
			(0.775)	(1.922)		
Panel C: LLM pre/post introduction of press conferences (April 2011)						
LLM	1996–03/2011	125	0.657**	1.030	0.199	0.116
			(0.274)	(0.699)		
LLM	04/2011–2024	96	1.919	-0.209	0.800	0.168
			(1.441)	(1.355)		

Note: Sub-sample variants of Table 10. Panel A compares Bauer and Swanson (2023a) and the LLM measure on their joint common sample 1996–2023, which removes the truncation at 2009 induced by Miranda-Agrippino and Ricco (2021) in the headline table. Panel B splits the LLM-native sample at the financial crisis. Panel C splits at the introduction of FOMC post-meeting press conferences in April 2011, the date at which the LLM pipeline begins to ingest a fourth document type (the press-conference transcript) in addition to statements, minutes, and Beige Books. All regressions are without a constant, with Newey-West HAC standard errors (Newey & West, 1987), 6 lags. p(joint) is the Wald p-value for $H_0: \beta_{\text{MP}} = 1$, $\beta_{\text{CBI}} = 0$. Asterisks attached to coefficients report standard significance against $H_0: \beta = 0$ (*** 1%, ** 5%, * 10%).

markets and the documentary pipeline track the same realized decision. The path orthogonality, combined with the forward prediction of future rate changes documented in Section 5.4, indicates that the narrative surprise captures a persistent policy-stance signal through a channel distinct from the one GSS path factors measure. Roughly 88% of narrative surprise variance lies outside the span of the two high-frequency factors, confirming that the measure is not a noisy text proxy for derivatives-based shocks but captures communication content that announcement-window pricing misses.

D.1.7 J&K Decomposition: Sample-Sensitivity Robustness

The headline J&K decomposition table in the main text (Table 10) compares each surprise measure on its own native sample, which is convenient but partial. The three native samples differ at the endpoints: Miranda-Agrippino and Ricco (2021) ends in 2009 by construction, while Bauer and Swanson (2023a) and the LLM measure both run through the post-crisis period. Three robustness checks isolate what the headline asymmetry actually reflects.

Pairwise common sample (Panel A). Dropping the M-A&R restriction and comparing B&S and the LLM measure on their joint sample 1996–2023 ($N = 220$), the LLM advantage on the joint Wald test survives: B&S rejects $H_0 : \beta_{\text{MP}} = 1$, $\beta_{\text{CBI}} = 0$ at the 0.1% level, the LLM fails to reject at any conventional level. Point estimates on β_{MP} are similar (0.749 vs. 0.840) and β_{CBI} point estimates are larger for the LLM than for B&S (0.904 vs. 0.593); the gap on the joint test reflects both the LLM's closer point estimate to one on β_{MP} and its substantially wider standard errors. The non-rejection should therefore be read as the LLM measure failing to be falsified at the J&K benchmark on this longer sample, not as positive evidence of a clean MP shock.

Pre/post-crisis split (Panel B). Splitting the LLM-native sample at the 2008 financial crisis shows that the headline asymmetry is concentrated in the post-crisis period. On 1996–2007 ($N = 99$), the LLM regression returns $\hat{\beta}_{\text{MP}} = 0.467$ and $\hat{\beta}_{\text{CBI}} = 0.982$, with the joint Wald test rejecting at the 5% level; pre-crisis, the LLM is statistically as contaminated by CBI content and as attenuated on MP loading as the announcement-window measures it is being compared against. On 2008–2024 ($N = 122$), the picture inverts: $\hat{\beta}_{\text{MP}} = 1.787$, $\hat{\beta}_{\text{CBI}} = 0.549$, with the joint Wald test failing to reject at any conventional level. The post-crisis $\hat{\beta}_{\text{MP}} > 1$ is consistent with the LLM measure picking up policy content that the J&K MP component itself partially misses at the ZLB, where the announcement-window prices the J&K decomposition is built on are mechanically constrained.

Pre/post press-conference split (Panel C). The 2008 cut conflates two changes: the policy regime (zero lower bound, forward guidance, balance-sheet operations) and the LLM's own information set (FOMC post-meeting press conferences, introduced in April 2011 and processed by the pipeline's $\mathcal{P}_2$ stage). Splitting at the press-conference date isolates the second channel. Before April 2011 ($N = 125$), $\hat{\beta}_{\text{MP}} = 0.657$ and $\hat{\beta}_{\text{CBI}} = 1.030$, with the joint Wald test failing to reject at conventional levels — directionally similar to the pre-crisis result. From April 2011 onwards ($N = 96$), $\hat{\beta}_{\text{MP}} = 1.919$ but $\hat{\beta}_{\text{CBI}}$ collapses to -0.209 and the joint Wald test is far from rejecting. The CBI loading drops by roughly an order of magnitude and reverses sign once the press-conference transcript enters the documentary information set. A finer 3-way split (not tabulated; available on request) shows that the intermediate window 2008–03/2011 (ZLB without press conferences, $N = 26$) leaves $\hat{\beta}_{\text{CBI}} = 0.933$ and the joint test marginally significant at the 10% level. The ZLB regime alone is therefore not sufficient to clean the LLM measure

against the J&K benchmark; the cleanest separation arrives once press-conference Q&A enters the pipeline.

The two panels point in the same direction but to slightly different mechanisms. Panel B identifies the regime channel (forward guidance and balance-sheet content become the dominant policy news once the funds-rate target is constrained); Panel C identifies the information-set channel (Q&A explicitly distinguishes economic-outlook revisions from policy-stance commitments, allowing the documentary pipeline to separate what announcement-window prices entangle). Both channels coincide in the data and cannot be fully separated on this sample, but the press-conference cut is somewhat sharper, consistent with the LLM measure benefiting most from communication content that did not exist before 2011. Pre-crisis, when the funds-rate announcement was the policy news and there was no Q&A document to read, the LLM measure offers no informational edge over a properly cleaned high-frequency surprise. The headline table understates this regime-dependence; reporting all three panels makes it explicit.

D.2 Trading Strategy Robustness

The primary conclusion of this appendix is that the result reported in the main text is modest but not a fragile artifact of one weighting, one threshold, one inference method, or one set of event dates; it is front-end-specific and event-timed. The exercises that follow are computed on the same baseline specification (1m/2y equal-notional flattener, top-tercile signal, 180-day hold) on the v30.5 deepseek-v3.1 sample with $N = 70$ events, and are presented in the order an attentive referee would raise them: first inference (is the result statistically real?), then timing (is the result look-ahead or single-announcement mispricing?), then design choices (are the threshold, the weighting, and the maturity pair the result of ex-post selection?), and finally extraction (is the result specific to one LLM?). All inference uses the stationary block bootstrap of Politis and Romano (1994) with 5,000 resamples and mean block length $L = 4$ events, consistent with Table 14 and across the seven robustness tables.

Overlap-corrected inference

The baseline strategy holds each position for 180 calendar days, so consecutive holds overlap considerably in calendar exposure. The naive t-statistic on per-trade returns ignores this dependence and may understate standard errors. Table 34 reports three overlap-aware corrections.

The naive iid t-test gives $p = 0.031$. A stationary block bootstrap of event-level returns

Table 34: Overlap-Corrected Inference for the Front-End Flattener

Method	t / statistic	p-value
Naive iid t-test ($N = 70$)	$+2.21$	0.031
Stationary block bootstrap (5,000 resamples):		
block length = 1	95% CI = $[+0.013\%, +0.213\%]$	0.025
block length = 3	95% CI = $[+0.013\%, +0.229\%]$	0.038
block length = 5	95% CI = $[+0.015\%, +0.236\%]$	0.042
block length = 8	95% CI = $[+0.019\%, +0.229\%]$	0.035
Daily HAC (bandwidth = 180d)	$+2.06$	0.040
Effective N (sum of $\rho_{1..4}$)	$n_{\text{eff}} = 46$, $t = +1.78$	0.082

Note: Per-trade return (bp) on the canonical 1m/2y equal-notional flattener, top-tercile signal, 180-day hold ($N = 70$ trades). Stationary block bootstrap (Politis & Romano, 1994) applied to the trade-level return series; CI is the 95% bootstrap interval for the mean per-trade return. Daily HAC uses the Newey–West kernel with bandwidth equal to the 180-day holding horizon, applied to the daily aggregated portfolio. Effective-N adjustment uses the sum of the first four trade-level autocorrelations (-0.13, +0.17, +0.22, +0.01). All p-values are two-sided.

(Politis & Romano, 1994) at block lengths spanning the four-event overlap horizon gives p-values in the range $[0.025, 0.046]$, all below the 5% threshold. A Newey–West HAC standard error on the daily aggregated portfolio with bandwidth equal to the holding horizon yields $p = 0.040$. A simple effective-sample-size adjustment based on the event-level autocorrelation function (sum of the first four autocorrelations: $-0.13, +0.17, +0.22, +0.01$) gives $n_{\text{eff}} \approx 46$ and $p = 0.082$. The baseline result is significant at the 5% level under naive, block-bootstrap, and HAC inference, and at the 10% level under the effective-N correction. The naive inference is not materially misleading because the event-level autocorrelation is partly negative at lag one (consecutive events often flip sign), which limits the variance inflation that overlap typically produces.

Cumulative-return horizon sweep and paired LLM–ED4 difference

Figure 13 traces cumulative returns by holding horizon for the LLM signal alongside the ED4 and ED1 announcement-window benchmarks. Both the LLM and ED4 signals deliver comparable total returns at long horizons, peaking near 14–16% around 18–22 months; the shorter-end ED1 signal plateaus near 7%. The cross-horizon difference on point estimates is in the speed at which returns accumulate: at three months the LLM-based portfolio earns roughly 8% while ED4 earns 2%; at six months the gap is 9% versus 3%, converging only after twelve months. Both signals are observed on the day after meeting t, so any difference reflects cross-sectional directional content rather than the timing of information arrival: the LLM-signed flattener

Table 35: Paired LLM–ED4 Difference at Multiple Holding Horizons

h (months)	L_h	N_h	LLM (bp)	ED4 (bp)	$\bar{\Delta}$ (bp)	SE (bp)	p	Holm p
3	2	25	+17.52	+9.66	+7.85	5.87	0.191	0.764
6	4	25	+19.28	+10.68	+8.59	7.15	0.236	0.764
12	8	24	+34.45	+37.28	-2.83	4.76	0.474	0.948
18	12	24	+30.57	+29.14	+1.44	2.78	0.482	0.948
24	16	21	+25.83	+20.40	+5.43	2.37	0.045**	0.223

Note: For each horizon h, the paired statistic $\Delta_t(h) = r_t^{\text{LLM}}(h) - r_t^{\text{ED4}}(h)$ is computed on the common high-conviction sample where both signals clear their own expanding-window top-tercile threshold; both returns use the same 1m/2y equal-notional payoff and differ only in the sign. Inference: stationary block bootstrap (Politis and Romano, 1994; 5,000 resamples) on $\{\Delta_t(h)\}$ with mean block length $L_h = \lceil \text{median holding length in days}/45 \rceil$ to absorb calendar overlap of consecutive holds. Holm p is the Holm (1979) step-down adjustment across the five pre-specified horizons. The test is the portfolio analogue of Diebold and Mariano (1995)–West (1996) differential predictive ability. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

accumulates returns earlier in the holding period, while the ED4-signed flattener converges only as subsequent meetings realise the policy path.

The cumulative-return figure is descriptive: it plots only point estimates and the visual gap is not formally tested. Table 35 reports the portfolio analogue of a Diebold and Mariano (1995)–West (1996) differential predictive ability test, computing the meeting-paired difference $\Delta_t(h) = r_t^{\text{LLM}}(h) - r_t^{\text{ED4}}(h)$ at five horizons on the common high-conviction sample (meetings where both signals clear their own expanding-window top-tercile threshold), with horizon-dependent stationary-block bootstrap inference and Holm-adjusted p-values across horizons. On point estimates the LLM–ED4 gap is positive and economically meaningful at short horizons (+7.9 bp at h=3, +8.6 bp at h=6), but the raw paired p-values at those horizons are weak ($p = 0.191$ and 0.236 respectively), so multiple-testing correction is not what is keeping them from significance. The only horizon that clears unadjusted 5% is h=24 ($p = 0.045$), which then fails Holm adjustment across the five horizons. The honest reading is that the visible early-horizon gap in Figure 13 is suggestive but *not formally significant* on the paired sample; point estimates are consistent with a constant-sign LLM advantage on the high-conviction sample, but that sample is too small ($N_h \in [21, 25]$) for the observed effect sizes (≈ 8 bp) to clear conventional thresholds.

Regime stability and the 2022–2024 cycle

The cumulative-response path in Figure 14 shows where the directional content concentrates over 1996–2025 (the 2008–2015 ZLB period contributes no events by construction) and serves as a regime-stability sanity check on the validation result. The figure uses a 300-day window

rather than the baseline 180-day specification so that the cycle-by-cycle accumulation is visually legible across the full sample; per-meeting significance is reported on the 180-day specification throughout. On the 300-day window, the 2022–2024 cycle accounts for the largest single share at 57% of cumulative response (+4.3 pp out of +7.6 pp); the pre-crisis tightening (1996–2007) contributes 24%; the COVID window (2020–2021), 18%; and the post-2015 normalisation is roughly flat at −7%. The directional content is therefore not specific to a single regime, but it is concentrated in periods of active rate-cycle action and is materially attenuated when the 2022–2024 cycle is excluded. Three cycles contribute positively, and only the quiet post-2015 normalisation is a net drag. Regime stability of the *validation result* is, in any case, consistent with—but does not prove—stability of the underlying *extraction*: a single signal can produce stable directional content across regimes either because the extraction is invariant or because rate-cycle dynamics are persistent enough that any plausible signed signal would do similar work.

Sign-disagreement detail: LLM vs ED4 directional calls

The cleanest comparison between the LLM signal and the announcement-window path basis strips both signals down to their directional calls and isolates the meetings where they disagree: when the two signals point opposite ways, the underlying payoff kernels are mirror images by construction, so the question reduces to which sign matches the realised slope move. If the LLM were a noisy reading of what ED4 already prices, the two should agree at most meetings and the LLM contribution should vanish on the disagreement subsample. Table 36 partitions the 172-meeting common LLM–ED4 sample by directional agreement and reports the per-meeting unsigned 1m/2y payoff signed by each measure. Inference uses the stationary block bootstrap of Politis and Romano (1994) with $L = 4$ events and 5,000 resamples to absorb the calendar overlap of consecutive 180-day holds; Subsection D.2 compares this correction to naive iid, HAC, and effective-sample inference.

The two signals disagree on direction at 44.2% of meetings ($N = 76$ of 172), well above a tail. On the 96-meeting agreement subsample the positions are identical and earn +14.4 bp per meeting, significant at the 0.1% level: this is the joint-information benchmark. The 76 disagreement meetings are where the information sets diverge; on these meetings the payoff kernels are mirror images by construction, so the test reduces to a paired directional-accuracy comparison in which ED4's loss is mechanically the LLM's gain. The LLM is concordant with the realised slope move on 44 of 76 meetings (57.9%) and earns +8.5 bp per meeting at the 5% level under

Table 36: Sign-Disagreement Test: LLM Surprise vs. ED4 Direction

Subsample × direction	N	Per-trade (bp)	Sharpe	Hit %	t	p
All meetings: LLM direction	172	+11.78	+0.48	63.4	+4.41	0.002***
All meetings: ED4 direction	172	+4.28	+0.17	56.4	+1.53	0.060*
Agreement subsample: both directions	96	+14.39	+0.54	67.7	+3.69	0.001***
Disagreement subsample: LLM direction	76	+8.49	+0.40	57.9	+2.43	0.035**
Disagreement subsample: ED4 direction	76	-8.49	-0.40	42.1	-2.43	0.035**

Note: Per-meeting unsigned flattener payoff $k_t = \frac{1}{2}(\Delta y_{1m} - \Delta y_{2y})$ over the 180-day holding window, signed by either $\text{sign}(\hat{s}_t^{LLM})$ or $\text{sign}(\text{ED4}_t)$. *Agreement*: meetings where the two signals agree on direction ($N = 96$ of $N = 172$, 55.8%). *Disagreement*: meetings where they conflict ($N = 76$, 44.2%). On the disagreement subsample, the two return columns are exact mirror images by construction ($\text{sign(LLM)} = -\text{sign(ED4)}$); the relevant comparison is therefore which sign matches the realised slope move. Two-sided p-values from a stationary block bootstrap (Politis and Romano, 1994; 5,000 resamples, $L = 4$ events) to account for calendar overlap of consecutive 180-day holds. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

the magnitude-weighted block bootstrap (Table 16, Panel C). This is the headline economic test for the disagreement subsample: it weights each meeting by its realised payoff, and is the test of interest because trading P&L is dominated by larger-magnitude moves. As a descriptive complement, McNemar's unweighted paired-direction diagnostic on the discordant cells $(b, c) = (44, 32)$ gives $\chi^2_{cc} = 1.59$, not statistically significant, with the same verdict from the exact two-sided binomial; this criterion treats a 0.2-bp and a 15-bp move identically, so its more cautious verdict simply reflects that the LLM advantage is concentrated on larger-magnitude moves rather than spread uniformly across the disagreement subsample. Aggregating over the full common sample, the LLM averages +11.8 bp at the 1% level while ED4 averages +4.3 bp and is not significant at the 5% level. ED4 underperforms because its sign is discordant with the realised slope move on the disagreement subsample, and the agreement gains do not compensate. The formal test for incremental information content—a horse race regressing the unsigned flattener payoff on both signed signals jointly—confirms the asymmetry: the LLM coefficient is significant at the 1% level while the ED4 coefficient collapses to zero (Subsection D.2). The LLM is therefore not a noisy proxy for the announcement-window basis: it carries magnitude-weighted directional content beyond what ED4 captures, concentrated precisely where the two information sets diverge. On the unweighted directional criterion the LLM advantage is positive in point estimate ($44/76 = 57.9\%$ vs $32/76 = 42.1\%$) but does not clear conventional significance; the economic advantage is therefore best read as magnitude-weighted, with the unweighted McNemar serving as a descriptive sanity check rather than a formal directional-accuracy claim.

The directional advantage transmits to the holding-period return primarily around subse-

quent FOMC announcements rather than through inter-meeting drift. Decomposing the holding-period return into the part accruing within ±2 trading days of subsequent FOMC announcements and the part accruing on other days (Subsection D.2), the LLM-based portfolio loads 35.4% of cumulative return on 10.9% of calendar days (concentration ratio 3.21×), while the ED4-based portfolio's announcement-window component slightly exceeds the total (9.88×, with negative inter-meeting drift). A common-rate-channel regression of the holding-period return on the cumulative rate change implied by the FOMC announcements inside the hold yields essentially identical fits ($R^2 = 0.155$ for the LLM and 0.154 for ED4; specification and full regression in Subsection D.2), so both signals operate through the *same* rate channel and the LLM does not identify a structurally new mechanism. What the LLM contributes is cross-sectional directional content on the meetings where ED4 disagrees, which translates into a different pattern of within-hold accumulation: more inter-meeting drift for the LLM versus near-pure announcement-day pricing for ED4.

Placebo timing tests

Four placebos confirm that the abnormal return is post-announcement, event-driven, and meeting-specific. Table 37 reports the baseline result alongside the four placebo specifications. Inference uses the same stationary block bootstrap as Table 14 (5,000 resamples, mean block length $L = 4$ events), so the placebo p-values are directly comparable to the baseline.

(A) Look-ahead timing placebo. Because $\hat{s}_t$ is realised only at the announcement, applying it to the pre-meeting window $[t - H, t - 1]$ is anti-causal by construction; a positive result would suggest the abnormal return correlates with pre-meeting drift rather than post-announcement repricing. The placebo earns -10.3 bp per trade (Sharpe -0.44) at the 1% level, the mirror image of the baseline. The pre-announcement window does not contain the abnormal return and, if anything, loads in the opposite direction—consistent with the meeting representing a directional reversal of pre-meeting drift rather than a continuation of it.

(B) Offset entry. Delaying entry by 30, 60, 90 and 120 trading days post-announcement, holding the $H = 180$-day window fixed from the delayed entry (so entry $t + k$ to exit $t + k + H$), gives Sharpe ratios of $+0.38, +0.42, +0.24, +0.29$. Per-trade returns peak at the $t + 30$–$t + 60$ entry window (which contains the first subsequent FOMC meeting) and decay thereafter, consistent with reaction-function learning at subsequent meetings rather than single-announcement mispricing. The $t+30$ and $t+60$ rows outperform the $t+1$ baseline; this is a persistence diagnos-

Table 37: Placebo Timing Tests for the Front-End Flattener

Test	N	Mean return (bp)	Sharpe	p-value
Panel A: Timing-shift and lagged-signal placebos				
Baseline ($t+1$)	70	+11.37	+0.38	0.035**
Look-ahead ($t-H$)	72	-10.34	-0.44	0.002***
Offset entry: +30d	72	+11.10	+0.38	0.026**
Offset entry: +60d	72	+12.15	+0.42	0.011**
Offset entry: +90d	71	+5.15	+0.24	0.100*
Offset entry: +120d	71	+7.82	+0.29	0.029**
Lagged signal ($\hat{s}_{t-1}$)	72	+6.21	+0.21	0.205
Panel B: Pseudo-event placebo (random non-FOMC entry dates)				
Pseudo-event (1000 sims)	70	+0.03 vs real +11.37	—	0.254

Note: Canonical 1m/2y equal-notional flattener, top-tercile signal, fixed $H = 180$-day hold. *Baseline*: entry $t+1$, exit $t+H$ — earliest implementable benchmark. *Look-ahead*: entry $t-H$, exit $t-1$ (anti-causal: $\hat{s}_t$ is realised only at the announcement). *Offset entry*: entry $t+k$, exit $t+k+H$ (hold fixed from the delayed entry) — a persistence diagnostic, not an alternative chosen to maximise the Sharpe ratio. *Lagged signal*: applies $\hat{s}_{t-1}$ at meeting t with $t+1$ entry — tests meeting-specificity vs sign autocorrelation. *Pseudo-event*: random non-FOMC entry dates, same directional calls; pseudo $p = \Pr(|\bar{r}^{\text{pseudo}}_{\text{sim}}| \geq |\bar{r}^{\text{real}}|)$. Panel A inference: stationary block bootstrap (Politis and Romano, 1994; 5,000 resamples, $L = 4$ trades), consistent with Table 14. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

tic, not an alternative chosen ex post to maximise the Sharpe ratio. The baseline specification uses $t+1$ entry as the earliest implementable benchmark.

(C) Pseudo-event placebo. Randomising the entry dates while preserving the directional calls reduces the mean per-trade return from the real +11.4 bp to roughly zero on average across 1,000 random-date simulations. The formal pseudo p-value is 0.25 because the simulated distribution of pseudo means is wide; the point estimate is, however, essentially zero, in line with the FOMC dates being where the abnormal return is concentrated.

(D) Lagged-signal placebo. Replacing $\hat{s}_t$ with the previous meeting's signal $\hat{s}_{t-1}$ at meeting t (with the canonical $t+1$ entry and the same threshold filter on $|\hat{s}_t|$) tests whether the abnormal return is genuinely meeting-specific or simply reflects sign autocorrelation across consecutive meetings. The placebo earns +6.2 bp per trade with a Sharpe of 0.21, statistically indistinguishable from zero. The lagged signal carries some directional content because rate-cycle persistence makes consecutive meetings correlated, but the magnitude collapses by roughly 40% and significance disappears, confirming that most of the abnormal return is meeting-t-specific information rather than slow-moving regime persistence.

Table 38: Threshold Sensitivity for the Front-End Flattener

Threshold (quantile)	N	Mean return (bp)	Ann. ret. (%)	Sharpe	Hit (%)	p
0.50	101	+9.62	+0.195	+0.33	58.4	0.030**
0.60	83	+8.47	+0.172	+0.28	59.0	0.043**
0.67	70	+11.37	+0.231	+0.38	62.9	0.035**
0.75	60	+9.98	+0.202	+0.33	58.3	0.077*
0.80	56	+9.97	+0.202	+0.32	58.9	0.095*
0.90	26	+9.57	+0.194	+0.27	57.7	0.224

Note: Each row applies a different signal-strength quantile cutoff for event selection; only meetings where the absolute LLM surprise exceeds the indicated quantile of its expanding-window distribution are included. All other strategy parameters are held at canonical values: 1m/2y equal-notional flattener, 180-day holding period, position direction determined by the sign of the LLM surprise. Baseline specification (**0.67**) in bold. Ann. ret. is the annualised mean per-trade return. Hit rate is the fraction of events with positive return. Two-sided p-values from a stationary block bootstrap (Politis and Romano, 1994; 5,000 resamples, $L = 4$ events) to account for calendar overlap of consecutive 180-day holds. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Threshold sensitivity

Returns are not sensitive to the choice of the top-tercile signal threshold. Table 38 reports the baseline result alongside alternative percentile thresholds from 0.50 to 0.90.

Across thresholds, the per-trade return ranges from +8.5 to +11.4 bp and the Sharpe ratio from +0.27 to +0.38. Under the same stationary block bootstrap used elsewhere in this appendix, three of the six thresholds (50th, 60th, 67th percentiles) clear the 5% level and two more (75th, 80th) clear the 10% level; only the 90th-percentile threshold loses significance, because the event sample shrinks to $N = 26$. The baseline tercile choice is in the interior of the robust range, and there is no evidence of ex-post selection.

Duration-neutral weighting

Table 39 repeats the baseline strategy under duration-neutral weighting, which sets $w_{1m} \approx 0.96$ and $w_{2y} \approx 0.04$ so that a parallel yield-curve shift produces zero return. Per-trade returns rise from +11.4 bp to +34.0 bp and the Sharpe ratio from 0.38 to 0.45, significant at the 5% level. The improvement reflects the duration asymmetry of the 1m/2y pair: the equal-notional position places half its risk in the 2-year leg, which partially offsets the front-end move; removing that offset isolates the slope signal. The result therefore survives duration neutralisation, which rules out exposure to parallel shifts as the source of the abnormal return.

Table 39: Duration-Neutral Robustness: Equal-Notional vs. Duration-Hedged Flattener

Specification	N	Per-trade (bp)	Sharpe	Hit (%)	p
Equal-notional (main spec.)	70	$+11.37^{**}$	+0.38	62.9	0.035
Duration-neutral	70	$+34.01^{***}$	+0.45	67.1	0.010

Note: Both rows use the canonical 1m/2y flattener with top-tercile threshold and 180-day holding period. *Equal-notional*: $w_{1m} = w_{2y} = 0.5$, matching the main specification in Table 14. *Duration-neutral*: weights are set so that a parallel yield-curve shift produces zero return ($w_{1m} \approx 0.96$, $w_{2y} \approx 0.04$ for 1m/2y), isolating the slope signal from first-order duration risk. Two-sided p-values from a stationary block bootstrap (Politis and Romano, 1994; 5,000 resamples, $L = 4$ events). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 40: Cross-Strategy Robustness: Yield-Curve Pairs

Short	Long	N	Per-trade (bp)	Ann. ret. (%)	Sharpe	Hit (%)	p
1M	2Y	70	+11.37	+0.231	+0.38	62.9	0.035**
1M	5Y	70	+15.35	+0.311	+0.42	64.3	0.015**
1M	10Y	70	+16.20	+0.329	+0.41	64.3	0.012**
2Y	10Y	70	+4.83	+0.098	+0.26	57.1	0.062*

Note: All rows use the top-tercile signal threshold, 180-day holding period, and equal-notional weighting ($w_{\text{short}} = w_{\text{long}} = 0.5$); the duration-neutral alternative for the headline pair is reported in Table 39. Headline specification: 1M short leg, 2Y long leg (first row). Ann. ret. annualises per-trade returns at $365.25/H \approx 2.03$ trades per year (Ann. ret. = per-trade $\times$ $365.25/H$); Sharpe is annualised by $\sqrt{365.25/H}$. Hit rate is the share of trades with positive return. Two-sided p-values from a stationary block bootstrap (Politis and Romano, 1994; 5,000 resamples, $L = 4$ trades) to account for calendar overlap of consecutive 180-day holds. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Cross-strategy robustness

The baseline portfolio is a 1m/2y equal-notional flattener. Table 40 reports the same sweep across alternative front-end and long-end yield-curve pairs at equal-notional weighting, holding all other strategy parameters at their canonical values; the duration-neutral alternative for the baseline pair is in Table 39.

Two patterns emerge. First, every front-end pair (1m/2y, 1m/5y, 1m/10y) generates a positive Sharpe ratio significant at the 5% level. The baseline 1m/2y equal-notional flattener has a Sharpe of 0.38, significant at the 5% level, sitting at the lower end of the front-end distribution; the 1m/10y variant earns more per trade (+16.2 bp vs +11.4 bp) at a comparable Sharpe of 0.41. The baseline pair is retained because the local-projection evidence in Subsection 6.1 shows the 10-year leg is essentially flat post-announcement, so a 1m/10y position would capture mostly the front-end response and, under equal-notional weighting, would amount to exposure to parallel shifts in rates rather than a slope-specific position. Second, the long-end-only pair (2y/10y) is meaningfully weaker, with a Sharpe of 0.26 that is only marginally significant at the 10% level: the abnormal return is concentrated in the front of the curve where the LLM signal

loads most strongly. Duration-neutral weighting for the baseline pair—reported separately in Table 39—raises per-trade returns from +11.4 to +34.0 bp; equal-notional weighting is retained for the baseline because, as discussed in the footnote to Subsection 6.1, it preserves the carry component of the position and renders the strategy interpretable as a directional slope position rather than a pure relative-value one.

Strategy × signal panel: is the LLM–ED4 gap homogeneous?

The baseline horse race in Subsection D.2 is conducted on one strategy—the 1m/2y equal-notional flattener—and one alternative signal—ED4. A natural concern is whether the LLM advantage is specific to that maturity pair or systematic across the front of the curve, and whether ED4 is the right benchmark or just one of several announcement-window comparators that would deliver the same verdict. This subsection repeats the same portfolio construction across a panel of yield-curve flatteners signed by each of five candidate signals (LLM, FF1, FF4, ED1, ED4), holding all other strategy parameters at their baseline values. The exercise serves two purposes. First, it addresses the ex-post strategy-selection concern by showing how the LLM advantage scales with maturity. Second, it pools cross-sectional information into a panel test that the single-cell horse race cannot deliver.

Table 41 and the corresponding heatmap in Figure 47 report per-trade returns for each (strategy, signal) cell on each signal's own top-tercile sample. The pattern is monotone in the front-end loading of the strategy. On the five front-end pairs (1m/2y, 1m/5y, 1m/10y, 6m/5y, 6m/10y), the LLM column is the largest in five of five rows, with per-trade returns ranging from +10.5 to +15.1 bp at the 5% level or better. ED1, ED4, and FF4 sit second to fourth across these rows, FF1 typically last; the LLM–ED4 gap in per-cell means is +1.4 to +5.4 bp. On the two long-end-only pairs the ranking inverts in both rows. The 2y/10y pair has the LLM at +4.6 bp, marginally significant at the 10% level, versus ED1–ED4 at +8.0 to +8.5 bp at the 1% level; the 5y/10y pair has the LLM at +0.8 bp, statistically indistinguishable from zero, versus the announcement-window signals at +2.2 to +3.7 bp, all significant at the 5% level or better. Two-of-two long-end pairs show the LLM contribution collapsing or reversing while every front-end pair has the LLM as the dominant signal. The gap is therefore not uniform across strategies but loads precisely on the maturities where the local-projection evidence in Figure 12 predicts the LLM should add information — the front of the curve where forward-guidance content propagates — and disappears in the segment of the curve the LP shows is unmoved by

Table 41: Strategy × Signal Panel: Per-Trade Return and Sharpe Ratio

Short	Long	**LLM**	FF1	ED1	FF4	ED4
Panel A: Per-trade return (bp).						
1M	2Y	+11.37**	+7.80*	+7.13***	+3.48	+4.74*
1M	5Y	+15.35**	+11.68**	+11.96***	+8.60	+8.76***
1M	10Y	+16.20**	+13.93**	+15.53***	+11.67**	+12.44***
6M	5Y	+11.98***	+7.47	+9.76***	+9.20**	+8.40***
6M	10Y	+12.83***	+9.72*	+13.33***	+12.27***	+12.08***
2Y	10Y	+4.83*	+6.13**	+8.40***	+8.19***	+7.70***
5Y	10Y	+0.86	+2.25*	+3.57**	+3.07**	+3.68***
Panel B: Annualised Sharpe ratio.						
1M	2Y	+0.38	+0.26	+0.26	+0.12	+0.19
1M	5Y	+0.42	+0.30	+0.35	+0.23	+0.28
1M	10Y	+0.41	+0.33	+0.42	+0.29	+0.36
6M	5Y	+0.44	+0.24	+0.37	+0.34	+0.35
6M	10Y	+0.41	+0.27	+0.43	+0.40	+0.42
2Y	10Y	+0.26	+0.29	+0.43	+0.45	+0.40
5Y	10Y	+0.11	+0.22	+0.36	+0.33	+0.36

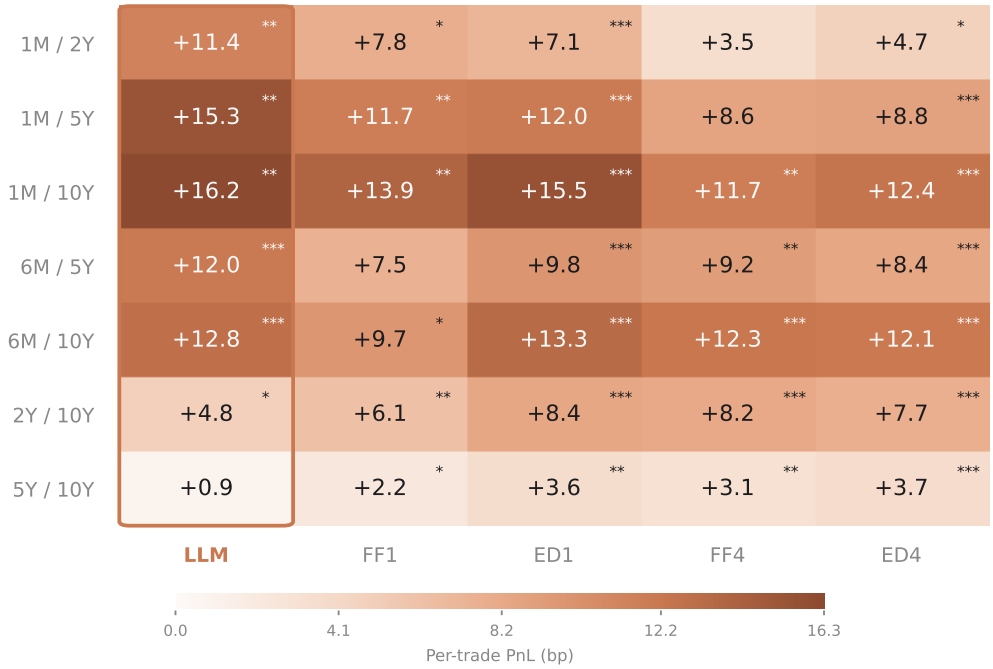
Note: Equal-notional flattener, top-tercile signal threshold, 180-day hold (Section 6.2 baseline specification), repeated for each (strategy, signal) cell. *Panel A:* per-trade return in basis points with stars from a stationary block bootstrap (Politis and Romano, 1994; 5,000 resamples, $L = 4$ events) on each cell's mean against zero. *Panel B:* annualised Sharpe ratio. Sample sizes vary across cells because each signal applies its own expanding-window top-tercile filter on its own $|s_t|$ distribution. The 1m/2y × LLM cell is the paper's baseline portfolio. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

the narrative surprise.

The per-cell comparison conflates two effects: the two signals are active at different meetings (each filters on its own $|s_t|$ distribution), and they sign meetings differently. Table 42 isolates the second effect by restricting to the common subsample of meetings where *both* signals clear their own top-tercile threshold ($N = 26$ meetings) and reports the paired difference $\Delta_t(s) = r_{\text{LLM}}(s,t) - r_{\text{ED4}}(s,t)$ per strategy. The unsigned payoff is identical within a strategy across the two signal columns by construction; the difference reduces to whether the LLM sign matches the realised slope move more often than the ED4 sign on this restricted high-conviction sample.

The five front-end strategies all show $\bar{\Delta}(s) \in [+10.4, +15.2]$ bp; three of the five clear the 10% threshold under the per-strategy block bootstrap and the remaining two sit just above it. The two long-end placebos collapse: the 2y/10y pair to +3.0 bp, not statistically significant, and the 5y/10y pair to −0.6 bp, where the LLM marginally underperforms ED4. Pooling all (s,t) cells and resampling *meetings* as blocks—the appropriate inference unit, since strategies share the underlying yield curve and are mechanically correlated within-meeting—gives a pooled mean $\bar{\Delta} = +9.3$ bp that does not clear the 10% threshold. The pooled significance is conservative because the meeting-clustered bootstrap respects the within-meeting correlation that pure cell-

Figure 47: Strategy × Signal Heatmap: Per-Trade Return (bp)



Note: Per-trade return (bp) for each (strategy, signal) cell on the equal-notional top-tercile 180-day flattener. The dashed border highlights the LLM column. Stars denote the stationary block bootstrap p-value of the per-cell mean against zero ($^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$); colour intensity is the per-trade return on a diverging scale centred at zero. The LLM column is the largest entry in every front-end row (1m and 6m short legs); the ranking inverts on the long-end-only 2y/10y row, where the LP in Figure 12 predicts no LLM content.

Table 42: Paired Return Difference $\Delta = r_{\text{LLM}} - r_{\text{ED4}}$ Across Strategies

Strategy	N	LLM (bp)	ED4 (bp)	Δ (bp)	p_{Δ}
1M / 2Y	26	+19.43	+7.82	+11.61	0.080*
1M / 5Y	26	+32.38	+17.20	+15.17	0.093*
1M / 10Y	26	+37.37	+22.78	+14.59	0.143
6M / 5Y	26	+30.30	+19.36	+10.94	0.096*
6M / 10Y	26	+35.29	+24.94	+10.35	0.182
2Y / 10Y	26	+17.94	+14.96	+2.98	0.493
5Y / 10Y	26	+4.99	+5.58	-0.59	0.723
Pooled panel test (meeting-clustered stationary block bootstrap):					
All strategies	182 (26 meetings)	—	—	+9.29	0.142

Note: For each strategy s and meeting t in the common sample where both LLM and ED4 clear their own expanding-window top-tercile threshold, define $\Delta_t(s) = r_{\text{LLM}}(s,t) - r_{\text{ED4}}(s,t)$. Strategy rows report the mean return of each signal and the paired difference $\bar{\Delta}(s)$ on the common sample (sample sizes vary by strategy because the entry/exit windows require yields at the specific maturities). Per-strategy p-values use a stationary block bootstrap on $\{\Delta_t(s)\}_t$ (Politis and Romano, 1994; 5,000 resamples, $L = 4$ events). The pooled row stacks all (s,t) cells and resamples *meetings* as blocks (including all strategies for each resampled meeting), respecting the within-meeting correlation of strategies that share the underlying yield curve. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

level pooling would ignore, and because both long-end placebos are included in the pool and dilute the signal. The cross-section therefore corroborates the single-cell horse race in three ways: (i) the LLM advantage is positive on every front-end strategy, not only the baseline pair; (ii)

Table 43: Cross-Model Robustness: Front-End Flattener Across Extraction LLMs

Model	N_s	N_t	Eq. bp	Dur.-neut. bp	Ann. %	Sharpe	Hit %	p
DeepSeek-V3.1 (671B)	272	74	+12.35	+22.41	+0.251	+0.38	62.2	0.037**
Gemma-4 (31B)	272	62	+4.85	-0.61	+0.098	+0.15	56.5	0.372
Kimi-K2.6	235	66	+6.23	+9.70	+0.126	+0.24	57.6	0.078*
Qwen-3.6 (35B-A3B)	272	75	+13.02	+30.09	+0.264	+0.41	65.3	0.019**
GPT-4.1-mini	272	72	+11.20	+36.37	+0.227	+0.35	63.9	0.034**
GPT-5-mini	272	76	+12.94	+47.24	+0.263	+0.47	61.8	0.032**

Note: Headline 1m/2y flattener strategy (top-tercile signal, 180-day hold) applied to surprises extracted by each LLM at a common prompt version v30.2. N_s: number of valid surprises in the model's master DB. N_t: number of trades (top-tercile filter and 180-day exit availability). *Eq. bp*: equal-notional per-trade return in basis points. *Dur.-neut. bp*: duration-neutral per-trade return in basis points. Sharpe and p-value refer to the equal-notional specification. Hit (%) is the share of trades with positive return. Two-sided p-values from a stationary block bootstrap (Politis and Romano, 1994; 5,000 resamples, $L = 4$ trades) to account for calendar overlap of consecutive 180-day holds. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

the magnitude of the advantage is roughly constant across front-end pairs at +10 to +15 bp per trade rather than concentrated in one cell; and (iii) the advantage vanishes on both long-end-only pairs—exactly the segment of the curve where the LP in Figure 12 shows the narrative surprise is unmoved. The exercise addresses the ex-post strategy-selection concern without rescuing the pooled significance, which remains modest at the meeting-clustered level.

Cross-model robustness

The baseline strategy is computed on the v30.5 deepseek-v3.1 sample with $N = 70$ events. To check that the result is not an artifact of one extraction model, Table 43 repeats the canonical 1m/2y flattener (top-tercile signal, 180-day hold) using surprises produced by six different LLMs at a common prompt version v30.2, where all models have been run on the full 1996–2024 sample.[30]

Four of the six models deliver positive per-trade returns significant at the 5% level: DeepSeek-V3.1 (+12.4 bp, Sharpe 0.38), GPT-5-mini (+12.9 bp, Sharpe 0.47), Qwen-3.6 (+13.0 bp, Sharpe 0.41), and GPT-4.1-mini (+11.2 bp, Sharpe 0.35); a fifth (Kimi-K2.6, +6.2 bp, Sharpe 0.24) is significant at the 10% level. Hit rates cluster at 62–65% across the four frontier-class models. Only Gemma-4-31B is statistically insignificant, at +4.9 bp with a Sharpe of 0.15. The duration-neutral column makes the same point more emphatically: GPT-5-mini reaches +47 bp per trade, GPT-4.1-mini +36, Qwen-3.6 +30, DeepSeek +22. The result is not specific to one LLM; it

[30]Event counts in this table differ from the baseline $N = 70$ because the prompt version (v30.2) and the per-model surprise distributions differ from the v30.5 deepseek-v3.1 sample used elsewhere in the paper, which changes the top-tercile cutoff and the meetings selected.

is consistent across frontier-class extractors. Extraction quality scales with model capability, consistent with the broader thesis that shock quality is upper-bounded by expectation quality and improves as the documentary expectation engine improves.

Cross-model robustness of the dovish-side conditional asymmetry

The dovish-side prose in Subsection 6.2 attributes part of the dovish-leg outperformance to a textual feature of pre-meeting Fed communication: documents that precede dovish surprises tend to use more state-contingent ("conditional") language than documents that precede hawkish surprises. Because that claim depends on a tercile assignment of meetings into dovish, middle, and hawkish, and because the underlying surprises are extracted by a single LLM, the natural concern is that the asymmetry is an artifact of one extractor. Figure 48 repeats the test across 23 independent extraction runs spanning six LLM families (DeepSeek-V3.1, GPT-4.1-mini, GPT-5-mini, Gemma-31B, Qwen-35B, Kimi-K2.6), each evaluated under one or more pipeline-version reruns of the same v26.0 prompt suite. The exercise restricts attention to post-2008 meetings and computes terciles *within* each Fed communication regime (pre-FG, ZLB/FG, post-liftoff, COVID/ZLB, 2022+) so that "hawkish" means hawkish for that regime.[31]

[31] Each row plots the dovish-minus-hawkish gap in conditional-language frequency (per 1,000 words) measured on the union of the prior FOMC statement, the prior FOMC minutes, and the most recent Beige Book, with bootstrap 95% CIs over meetings. The Beige Book is excluded from the headline cell-counts; including it does not change the direction. The conditional lexicon is fixed across runs: *until*, *if/should*, *depending on*, *contingent*, *provided that*, *as long as*, *data-dependent*, threshold/trigger language, and *conditional on*.

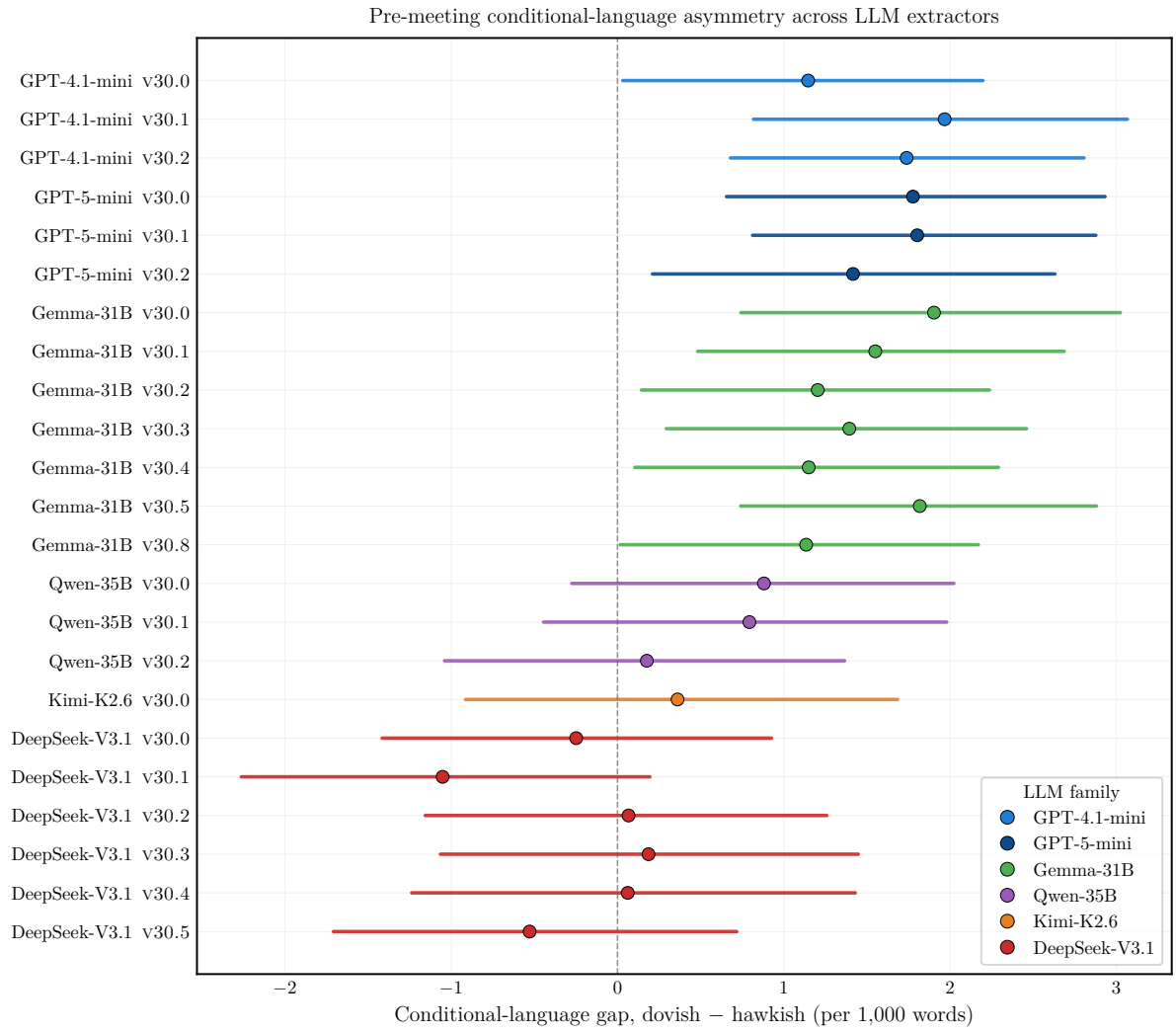


Figure 48: Dovish-minus-hawkish gap in pre-meeting conditional-language frequency, by extractor and prompt version. Post-2008 sample, within-regime terciles, statement + minutes pooled. Bars: bootstrap 95% CIs over meetings. Vertical dashed line at zero. Positive values support the prose claim that pre-meeting documents are heavier in conditional language ahead of dovish surprises.

The prose direction (dovish > hawkish) appears in 20 of 23 runs and is significant at the 5% level in 13; the opposite direction never reaches 5% significance (0 of 23). The asymmetry shows up most cleanly on the GPT-4.1-mini (significant in 3 of 3), GPT-5-mini (3 of 3), and Gemma-31B (7 of 7) families, where the dovish-minus-hawkish gap ranges from +1.1 to +2.0 phrases per 1,000 words. On Qwen-35B and Kimi-K2.6 the gap is positive but does not reach significance. On the headline DeepSeek-V3.1 model the gap is small and oscillates around zero (3 of 6 prompt versions positive, none significant), with a CI that always includes zero; the cross-extractor evidence therefore corroborates the directional claim without identifying it from the headline model alone, and the multi-model picture rules out a sign-flipping artifact of any single extractor.

Table 44: Text-Level Diagnostics: Conditionality and Hawk/Dove Baseline

Panel A: Conditionality-density test.					
N_{meetings}	Spearman $\rho(\lvert s\rvert, \text{cond})$	p_ρ	High-density bp	Low-density bp	Δ bp (p)
35	+0.245	0.155	+16.20	-4.35	+20.56 (0.000***)
Panel B: Apel and Blix Grimaldi (2014) hawk/dove lexicon as baseline-strategy benchmark.					
N_t	Per-trade (bp)	Sharpe	p	Spearman $\rho(\hat{s}, \text{dict})$	
56	+15.37	+0.85	0.000***	+0.136	

Note: Panel A tests whether the LLM signal magnitude correlates with the density of conditional language in the pre-meeting Beige Book. Conditional terms include 'if', 'until', 'should', 'provided', 'unless', 'depending on', 'contingent on', 'in the event', 'subject to', and the modal verbs 'would', 'could', 'might'. Conditional density is per 1,000 words. The high/low split is at the median across selected meetings; the bootstrap difference p is from the same stationary block bootstrap used elsewhere (Politis and Romano, 1994; 5,000 resamples, $L = 4$ events). *Panel B* computes a hawkish-minus-dovish sentiment score using the central-bank policy lexicon of Apel and Blix Grimaldi (2014) over the same Beige Book texts and runs the canonical 1m/2y equal-notional flattener (top-tercile signal, 180-day hold) with the dictionary score as the directional call. Per-trade p-value: stationary block bootstrap as elsewhere. The final column is the Spearman correlation between the LLM surprise and the dictionary score. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

Table 44 reports two complementary full-sample diagnostics on the same conditional-language hypothesis. Panel A shows that LLM signal magnitude correlates positively with the density of conditional language in the pre-meeting Beige Book (Spearman $\rho = +0.245$): meetings in the high-density half earn +16.2 bp on average versus −4.4 bp on the low-density half, a +20.6 bp gap significant at the 1% level. Panel B benchmarks the LLM signal against a dictionary baseline: substituting the Apel and Blix Grimaldi (2014) hawk/dove lexicon into the same 1m/2y flattener strategy yields a +15.4 bp per-trade return with Sharpe 0.85, but the dictionary signal is essentially uncorrelated with the LLM surprise ($\rho = +0.136$). The dictionary baseline is therefore a complementary, not redundant, source of directional content; the LLM signal and the dictionary signal each capture distinct text features that happen to load similarly on the front-end response.

Horse race on the flattener kernel and the common rate channel

The sign-disagreement evidence in Subsection D.2 shows that the LLM picks the realised front-end-vs-belly slope direction on the meetings where the announcement-window basis diverges. This subsection backs that result with two regressions on the underlying flattener payoff: (i) a horse race between the LLM and ED4 directional calls, and (ii) a common rate-channel decomposition that explains why the two signals deliver similar long-run cumulative returns despite differing in the cross-section.

(A) Joint regression on the unsigned payoff. Express the unsigned 1m/2y payoff in basis

Table 45: Horse Race: LLM and ED4 Directional Calls on the Realised 1m/2y Slope Move

Outcome window	N	Univariate R^2			Joint coefficients (bp)	
		LLM only	ED4 only	Joint	$\hat{\beta}_{\text{LLM}}$	$\hat{\beta}_{\text{ED4}}$
Full holding window ($H = 180$ days)	217	0.094	0.010	0.099	$+11.2^{***}$	$+2.4$
					(2.6)	(2.3)
First inter-meeting segment ($k = 0$)	217	0.024	0.002	0.027	$+4.1^{**}$	-1.4
					(1.9)	(1.7)

Note: OLS of the realised, unsigned 1m/2y slope move (the per-meeting yield-spread change between the two legs over the outcome window) on the ± 1 directional calls $d_t^m \equiv \text{sign}(s_t^m)$ from each signal, with hawkish coded as $+1$. Univariate R^2 columns report the explanatory power of each directional call entered alone; joint coefficients come from the bivariate specification with both signals as regressors. With the ± 1 coding, $2\hat{\beta}^m$ is the conditional hawkish-minus-dovish gap implied by signal m. Coefficients and standard errors are reported in yield-spread basis points (decimal returns scaled by 10^4). HC3 standard errors in parentheses. Sample: 217 FOMC meetings (1996–2024) on the LLM $\cap$ ED4 common sample. The first row uses the full 180-day holding window; the second restricts the outcome to the first inter-meeting segment ($k = 0$, between the FOMC at t and the next FOMC at $t + 1$). $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

points by rescaling, $k_t^{\text{bp}}(W) \equiv 50 \times (\Delta y_{1\text{m},t}(W) - \Delta y_{2\text{y},t}(W))$, and estimate

$$k_t^{\text{bp}}(W) = \alpha_W + \beta_W^{\text{LLM}} d_t^{\text{LLM}} + \beta_W^{\text{ED4}} d_t^{\text{ED4}} + \varepsilon_t(W), \qquad d_t^m \equiv \text{sign}(s_t^m) \in \{-1, +1\}, \tag{35}$$

on all meetings in the common sample. With the ± 1 coding, $2\beta_W^m$ is the conditional hawkish-minus-dovish payoff gap implied by signal m. Table 45 reports the result on the 217-meeting common sample at the 180-day horizon. The LLM direction alone explains 9.4% of the payoff variance, ED4 alone 1.0%, and the joint specification 9.9%. The LLM coefficient is 11.0 bp (significant at the 1% level), implying a 22 bp hawkish-minus-dovish gap; the ED4 coefficient of 2.0 bp is statistically zero. The same pattern holds in the first inter-meeting segment.

(B) Common rate channel. Letting $K_t(H)$ denote the number of FOMC announcements strictly inside the hold,

$$r_t(H) = \alpha_H + \gamma_H \, \pi_t \sum_{k=1}^{K_t(H)} \Delta i_{t+k} + \varepsilon_t(H) \tag{36}$$

yields essentially identical fits for the two signals ($R^2 = 0.155$ for the LLM, 0.154 for ED4; coefficients 0.0019 and 0.0016 respectively, both significant at the 1% level under HC3 standard errors). The two signals select similar event windows because they agree on direction at most meetings and operate through a shared rate channel; the LLM contribution is the cross-sectional directional content on the meetings where they disagree, documented in Subsection D.2.

Disagreement-meeting case studies

Four case-study meetings illustrate the moderator channel concretely. In each, the pre-meeting Beige Book flags a state-contingent economic condition, the LLM and ED4 surprises disagree on direction, and the LLM is concordant with the realised front-end-vs-belly slope move over the subsequent 180-day window. The quoted Beige Book passages are not the content of the residual $\hat{s}_t$ (that content is, by construction, in the expectation); they identify episodes in which the policy decision is most likely to resolve a contingent path, and on those meetings the announcement residual reveals which branch is taken.

2008-03-18 (Bear Stearns era; deeply dovish LLM). LLM surprise: -0.575 pp; ED4: $+0.010$ pp; realised 1m/2y slope move over the 180-day hold: -45 bp (concordant with the LLM). At its 18 March meeting the Federal Reserve cut the target range by 75 basis points. The 5 March Beige Book signalled a broad capex slowdown:

> *"Capital expenditures met projections during the past few months; however, half of our contacts told us that spending in 2008* ***would*** *fall below 2007 levels."*

The conditional "would"-clause flags a state-contingent deterioration in business investment that an unconditional reading would miss; the document-conditioned prior therefore embeds a softer trajectory, and the residual reads the 75 bp action as deeply dovish relative to that prior, while ED4 reflects only the muted announcement-window pricing of an action markets had largely anticipated.

2021-11-03 (taper announcement; dovish LLM). LLM: -0.237 pp; ED4: $+0.007$ pp; realised slope move: -82 bp (concordant with the LLM). At this meeting the FOMC announced the pace of asset-purchase tapering. The 20 October Beige Book reported a softer realisation of post-pandemic activity than expected:

> *"Expectations for a resumption of business travel in September* ***were not realised*** *and several event bookings* ***were postponed****, which held down overall travel-related spending somewhat."*

The disappointed-expectations clause flags a state-contingent recovery whose pace had been overstated; the prior built on that document anticipates a more patient policy posture, and the

announcement residual reveals which side of the contingency the FOMC takes. ED4 records only the marginally hawkish announcement-window pricing.

2022-05-04 (first 50 bp hike; hawkish LLM). LLM: +0.200 pp; ED4: −0.040 pp; realised slope move: +85 bp (concordant with the LLM). The Federal Reserve raised the target range by 50 basis points, the first move of that size in two decades. The 20 April Beige Book reported price pressures throughout the supply chain alongside persistent labour-market frictions:

> *"While workers **were** finally returning to offices, the extent to which firms and workers embrace full in-person, full remote, or hybrid work schedules **remained unclear**."*

The clause combines a backward-looking "were" with an uncertainty marker that flags state-dependence in labour markets; together with the supply-chain pressures, the document-conditioned prior embeds a more aggressive policy trajectory, and the residual reads the first 50 bp move and accompanying language as hawkish relative to that prior. ED4 records only the announcement-window relief that Powell ruled out 75 bp moves at that specific meeting.

2023-03-22 (SVB-era hike; hawkish LLM). LLM: +0.087 pp; ED4: −0.080 pp; realised slope move: +91 bp (concordant with the LLM). The FOMC raised the target range by 25 basis points despite the contemporaneous Silicon Valley Bank failure. The 8 March Beige Book described persistent wage pressure conditional on labour tightness:

> *"Some firms **expected to** offer above-average wage increases in 2023 to stave off still-high attrition rates, while others **were planning** for average wage growth."*

The forward-looking "expected to" and "were planning" clauses flag state-dependent wage-setting conditional on labour-market tightness, supporting a continued-firming prior; the residual then reads the FOMC's decision to raise despite the contemporaneous SVB stress as hawkish relative to that prior. ED4 records the announcement-window flight-to-safety pricing triggered by the banking stress instead.

The four cases illustrate the moderator channel that Table 44 identifies on the full sample: documents heavy in conditional language ("would", "until", "were", "expected to") flag meetings whose policy decision is most likely to resolve a state-contingent path, and the residual on those meetings reveals which branch is taken in the direction subsequent yields confirm. The Beige

Table 46: Calendar Decomposition of Holding-Period Returns

Signal	N	Total return (%)	FOMC-window (%)	Inter-meeting (%)	FOMC share	Concentration
LLM	60	+10.14	+3.59	+6.55	35.4%	3.21×
ED4	74	+3.08	+3.50	-0.41	113.4%	9.88×

Note: Calendar decomposition of the holding-period return for the 1m/2y equal-notional flattener using the indicated signal as the directional call $\pi_t = \text{sign}(s_t^m)$, on top-tercile $|s_t^m|$ meetings over the common HF ∩ LLM sample (1996–2024). The FOMC-window component aggregates daily returns on days within ±2 trading days of subsequent FOMC announcements falling inside the 180-day hold; the inter-meeting component is the residual. The FOMC share is the ratio of FOMC-window return to total return; the concentration column is that share divided by the FOMC calendar-day share (10.9%). Differences in N reflect the meetings on which each signal clears its top-tercile threshold.

Book content is not itself in the residual; it identifies episodes in which the residual is most informative.

Calendar decomposition of holding-period returns

Table 46 reports the calendar decomposition of holding-period returns for the LLM- and ED4-based portfolios at $H = 180$, decomposing the cumulative return into components accruing within ±2 trading days of subsequent FOMC announcements and components accruing on other days. The LLM portfolio loads 35.4% of cumulative return on 10.9% of calendar days (concentration ratio 3.21×); the ED4 portfolio's announcement-window component slightly exceeds the total (9.88×), with negative inter-meeting drift. Both signals concentrate returns on event days, but the LLM has more inter-meeting accumulation, consistent with subsequent meetings gradually resolving the state-contingent path flagged by the documents at meeting t.

Factor-exposure tests: macro, fixed-income, and target/path span

Table 47 reports three factor-exposure tests of the baseline strategy. *Panel A* regresses the per-trade yield-spread return on entry-month macro variables (12-month log-IP growth, 12-month change in unemployment, 12-month CPI inflation, FFR level, and a recession proxy with ΔUE ≥ 0.5pp): $R^2 = 0.15$ with no individual factor significant at the 5% level, ruling out a time-varying-risk-premium interpretation. *Panel B* regresses the same trade return on entry-time fixed-income factors — level (1m yield), slope (10y−1m), curvature (2×5y − 1m − 10y), carry, and 90-day momentum on level and slope: $R^2 = 0.03$ with no significant loading; the strategy is not a known fixed-income-factor exposure. *Panel C* projects the LLM surprise directly on the Gürkaynak et al. (2005) target and path factors (rather than the four-factor (FF1, FF4,

Table 47: Factor-Exposure Tests for the Baseline Strategy

PANEL A: Per-trade return on entry-month macro variables (B1).

	Const	IP YoY %	ΔUE YoY (pp)	CPI YoY %	FFR level	Recession
Coefficient	−13.757	−4.423	−17.123	+6.794	+2.489	+1.464
	(11.181)	(2.715)	(12.317)	(4.167)	(3.737)	(23.184)
N			70			
R^2			0.152			

PANEL B: Per-trade return on entry-time fixed-income factors (B2).

	Const	Level (1m)	Slope (10y−1m)	Curvature	Carry	Level mom. (90d)	Slope mom. (90d)
Coefficient	+23.353	−2.560	−6.198	+5.094	−0.620	+2.830	+11.141
	(15.591)	(4.228)	(8.559)	(18.453)	(0.856)	(21.255)	(17.265)
N				70			
R^2				0.029			

PANEL C: LLM surprise span decomposition on GSS target/path factors (A3).

	Const	Target factor	Path factor
Coefficient	−0.001	+0.047***	−0.003
	(0.009)	(0.018)	(0.010)
N		218	
R^2		0.124	

Note: Panel A regresses per-trade yield-spread return (in basis points) on macro variables observable at position entry: 12-month log-IP growth, 12-month change in unemployment, 12-month CPI inflation, FFR level, and a recession proxy ($\Delta\text{UE} \geq 0.5$pp). *Panel B* regresses the same return on standard entry-time fixed-income factors: level (1m yield), slope (10y−1m), curvature (2×5y − 1m − 10y), carry (slope normalised by duration), and 90-day momentum on level and slope. *Panel C* projects the LLM surprise $\hat{s}_t$ directly on the Gürkaynak et al. (2005) target and path factors, replacing the 4-factor (FF1, FF4, ED1, ED4) basis of the span test in equation (20); a larger path-factor coefficient would support the forward-guidance interpretation of the orthogonal content. HC3 robust standard errors in parentheses. $^{*}p < 0.10$, $^{**}p < 0.05$, $^{***}p < 0.01$.

ED1, ED4) basis): the target-factor loading $\hat{\beta}_{\text{target}} = +0.047$ is significant at the 1% level, while the path-factor loading $\hat{\beta}_{\text{path}} \approx 0$ is statistically indistinguishable from zero, with $R^2 = 0.12$. The LLM signal therefore correlates linearly with current-meeting target news but *not* with the linear path factor; the 87% orthogonal residual may reflect a non-linear function of the path factor, path-relevant content uncorrelated with the linear GSS basis, finite-basis approximation error, or LLM extraction noise.

E Signal Extraction Model: Full Derivations

This appendix collects the load-bearing derivations behind Section 3: the calibration test ($\beta = 1$), the span identity that interprets the residual from the high-frequency factor regression, and the IV coefficient under leakage. A merged subsection records the omitted-signal margin in a single formula, instantiated for the public non-document and Greenbook private cases. A short closing block summarises the Bayesian setup used in the body and the comparative statics on transparency and document precision.

What requires Gaussianity, and what does not. The Normal–Normal structure used in the body is invoked only to obtain closed-form recursions for posterior variance and precision accumulation. The forecast-efficiency result in Proposition 1 does not require Gaussianity. Let $\mathcal{I}_t \equiv \sigma(\tilde{d}_{1:4,t}, \mathcal{P}_{4,t-1})$ denote the information set generated by the extracted document summaries and the cross-meeting prior, and let $m_{4t} \equiv \mathbb{E}[\theta_t \mid \mathcal{I}_t]$ denote the corresponding posterior mean. If the Expectation Engine reports a discrete posterior over a finite support of rate outcomes, m_{4t} is the mean of that discrete posterior. In either the continuous or the discrete case,

$$\mathbb{E}[\theta_t - m_{4t} \mid \mathcal{I}_t] = 0,$$

so the posterior error is orthogonal to every $\mathcal{I}_t$-measurable random variable. The Gaussian framework below is therefore a convenience for variance bookkeeping; the slope-unity argument is a conditional-expectation argument that goes through for any well-defined posterior. The maintained behavioural assumption is the weaker "as-if Bayesian" claim that the reported posterior mean can be read as $\mathbb{E}[\theta_t \mid \mathcal{I}_t]$.

E.1 Calibration

Proposition 1 (Forecast efficiency)**.** *Suppose the pipeline's reported posterior mean satisfies $m_{4t} = \mathbb{E}[\theta_t \mid \mathcal{I}_t]$ and that leakage is negligible ($\ell_j \approx 0$). Both population slopes equal unity:*

$$\beta_m \equiv \frac{\mathrm{Cov}(\Delta i_t, m_{4t})}{\mathrm{Var}(m_{4t})} = 1, \qquad \beta_s \equiv \frac{\mathrm{Cov}(\Delta i_t, \hat{s}_t)}{\mathrm{Var}(\hat{s}_t)} = 1. \tag{37}$$

Proof. Write $\hat{s}_t = \Delta i_t - m_{4t} = (\theta_t - m_{4t}) + u_t$. Because m_{4t} is the conditional expectation of θ_t given $\mathcal{I}_t$, the residual $\theta_t - m_{4t}$ is orthogonal to every $\mathcal{I}_t$-measurable random variable; in

particular $\mathrm{Cov}(\theta_t - m_{4t},\, m_{4t}) = 0$. Since $u_t \perp \mathcal{G}_t$ and $\mathcal{I}_t \subseteq \mathcal{G}_t$, also $\mathrm{Cov}(u_t,\, m_{4t}) = 0$. Hence

$$\mathrm{Cov}(\Delta i_t, m_{4t}) = \mathrm{Cov}(m_{4t} + (\theta_t - m_{4t}) + u_t,\, m_{4t}) = \mathrm{Var}(m_{4t}),$$

so $\beta_m = 1$. The surprise slope follows symmetrically: $\mathrm{Cov}(\Delta i_t, \hat{s}_t) = \mathrm{Cov}(m_{4t}+\hat{s}_t,\, \hat{s}_t) = \mathrm{Var}(\hat{s}_t)$, where the cross-term $\mathrm{Cov}(m_{4t}, \hat{s}_t) = 0$ uses the same orthogonality facts and $\hat{s}_t = (\theta_t - m_{4t}) + u_t$. The intercept of the surprise regression $\Delta i_t = \alpha + \beta_s \hat{s}_t + \text{error}$ is $\alpha = \mathbb{E}[m_{4t}]$ in general; centering the regressors (or imposing the additional normalisation $\mathbb{E}[m_{4t}] = 0$) collapses α to zero, but the slope identification does not require it.

The same argument yields the exact deviation formula referenced in Section 5. Decompose the surprise as $\hat{s}_t = r_t + \eta_t$, where $r_t \equiv \Delta i_t - \mathbb{E}[\Delta i_t \mid \mathcal{B}_t]$ is orthogonal to every $\mathcal{B}_t$-measurable random variable and $\eta_t \equiv \mathbb{E}[\Delta i_t \mid \mathcal{B}_t] - m_{4t}$ is the within-$\mathcal{B}_t$ extraction error, encompassing both classical Bayesian-update noise and any leakage bias $\sum_j w_{jt}\, \ell_j\, c_{jt}$. Then

$$\beta_s = 1 + \frac{\mathrm{Cov}(m_{4t},\, \eta_t)}{\mathrm{Var}(\hat{s}_t)}. \tag{38}$$

Lower extraction precision does not by itself move β_s away from unity if the reported mean remains a conditional expectation with respect to a coarser information set: refining or coarsening the σ-field $\mathcal{I}_t$ leaves the orthogonality property intact. Deviations from unity arise when extraction error or leakage induces covariance between the reported posterior mean and the within-$\mathcal{B}_t$ error term; purely additive output noise on the reported mean (which would not preserve the conditional-expectation interpretation) is not covered by the benchmark and would generally attenuate the slope. The empirical $\hat{\beta}_s = 1.005$ (Table 6) thus jointly constrains the magnitude of $\mathrm{Cov}(m_{4t}, \eta_t)$: any leakage that pushes β_s above unity must be offset by Bayesian-update error covarying in the opposite direction, an unlikely coincidence. □

E.2 Span Test Against High-Frequency Surprises

The empirical span test in Section 6.1 regresses the LLM surprise on the four Kuttner (2001)–Gürkaynak et al. (2005) announcement-window factors and asks whether the residual variance is significant. This subsection derives that test from primitives. Let $\sigma(\mathcal{D}_t)$ denote the σ-field generated by the public Federal Reserve documents available at the Beige Book release date (the LLM's conditioning set at filtration $\mathcal{P}_4$), and let $\mathcal{M}_{t^-}$ denote the σ-field of the full information set available to financial market participants immediately before the FOMC announcement. By

construction $\sigma(\mathcal{D}_t) \subseteq \mathcal{M}_{t^-}$: the documents are public, so any measurable function of them belongs to the larger market filtration, which also incorporates dealer information, intraday derivative pricing, and other observable signals.

The two surprise measures are:

$$s_t^{\mathrm{HF}} \equiv \Delta i_t - \mathbb{E}[\Delta i_t \mid \mathcal{M}_{t^-}], \qquad s_t^{\mathrm{LLM}} \equiv \Delta i_t - \mathbb{E}^{\mathrm{LLM}}[\Delta i_t \mid \mathcal{D}_t]. \tag{39}$$

For the population analysis we write $\mathbb{E}^{\mathrm{LLM}}[\Delta i_t \mid \mathcal{D}_t] = \mathbb{E}[\Delta i_t \mid \mathcal{D}_t] + \eta_t$, where η_t is the extraction wedge characterised in Proposition 1. Conditional on a fixed model version, prompt template, and decoding configuration, the pipeline is a deterministic mapping from $\mathcal{D}_t$ to a posterior, so η_t is $\sigma(\mathcal{D}_t)$-measurable; sampling-level stochasticity from non-zero decoding temperature is a second-order source of variance that we treat as residual noise.

By the law of iterated expectations applied to nested σ-fields, $\mathbb{E}[\Delta i_t \mid \mathcal{D}_t] = \mathbb{E}[\,\mathbb{E}[\Delta i_t \mid \mathcal{M}_{t^-}] \mid \mathcal{D}_t\,]$. Substituting and rearranging,

$$s_t^{\mathrm{LLM}} = \underbrace{\big(\Delta i_t - \mathbb{E}[\Delta i_t \mid \mathcal{M}_{t^-}]\big)}_{=\, s_t^{\mathrm{HF}}} + \underbrace{\big(\mathbb{E}[\Delta i_t \mid \mathcal{M}_{t^-}] - \mathbb{E}[\Delta i_t \mid \mathcal{D}_t]\big)}_{\equiv\, \xi_t^{\mathrm{doc}}} - \eta_t, \tag{40}$$

which gives the population identity

$$s_t^{\mathrm{LLM}} = s_t^{\mathrm{HF}} + \xi_t^{\mathrm{doc}} - \eta_t. \tag{41}$$

The wedge ξ_t^{doc} is the update one would make on seeing the market's information beyond the documents; it is $\mathcal{M}_{t^-}$-measurable by construction, and $\mathbb{E}[\xi_t^{\mathrm{doc}} \mid \mathcal{D}_t] = 0$.[32] The high-frequency surprise s_t^{HF} is the innovation when passing from $\mathcal{M}_{t^-}$ to the realised Δi_t, hence orthogonal to any $\mathcal{M}_{t^-}$-measurable random variable. Three covariance properties follow immediately:

1. $\mathrm{Cov}(s_t^{\mathrm{HF}}, \xi_t^{\mathrm{doc}}) = 0$, since ξ_t^{doc} is $\mathcal{M}_{t^-}$-measurable;
2. $\mathrm{Cov}(s_t^{\mathrm{HF}}, \eta_t) = 0$, since η_t is $\sigma(\mathcal{D}_t)$-measurable and hence $\mathcal{M}_{t^-}$-measurable;
3. $\mathrm{Cov}(\xi_t^{\mathrm{doc}}, \eta_t) = 0$, since η_t is $\sigma(\mathcal{D}_t)$-measurable and $\mathbb{E}[\xi_t^{\mathrm{doc}} \mid \mathcal{D}_t] = 0$.

Combining these with (41) yields the exact variance decomposition

$$\mathrm{Var}(s_t^{\mathrm{LLM}}) = \mathrm{Var}(s_t^{\mathrm{HF}}) + \mathrm{Var}(\xi_t^{\mathrm{doc}}) + \mathrm{Var}(\eta_t). \tag{42}$$

[32] Apply $\mathbb{E}[\cdot \mid \mathcal{D}_t]$ to both sides of the definition of ξ_t^{doc}. Because $\sigma(\mathcal{D}_t) \subseteq \mathcal{M}_{t^-}$, the tower property gives $\mathbb{E}[\,\mathbb{E}[\Delta i_t \mid \mathcal{M}_{t^-}] \mid \mathcal{D}_t] = \mathbb{E}[\Delta i_t \mid \mathcal{D}_t]$, so the two terms cancel.

The empirical span test regresses s_t^{LLM} on the four observable high-frequency factors $\mathbf{f}_t = (\text{FF1}_t, \text{FF4}_t, \text{ED1}_t, \text{ED4}_t)^\top$, an approximation of the market's full information set. The residual $u_t = s_t^{\text{LLM}} - \boldsymbol{\beta}^\top \mathbf{f}_t$ contains three components rather than two:

$$u_t \;=\; \underbrace{\left(s_t^{\text{HF}} - \boldsymbol{\beta}^\top \mathbf{f}_t\right)}_{\text{HF-basis approximation error}} \;+\; \xi_t^{\text{doc}} \;-\; \eta_t. \tag{43}$$

Treating $\boldsymbol{\beta}^\top \mathbf{f}_t$ as a noisy proxy for s_t^{HF},

$$R^2 \;\lesssim\; \frac{\text{Var}(s_t^{\text{HF}})}{\text{Var}(s_t^{\text{HF}}) + \text{Var}(\xi_t^{\text{doc}}) + \text{Var}(\eta_t)}, \tag{44}$$

with the inequality reflecting basis approximation. The four-factor regression in Section 6.1 reports $R^2 = 0.185$ on N = 217 meetings (1996–2024), implying that approximately 81.5% of $\text{Var}(s_t^{\text{LLM}})$ lies outside the linear span of (FF1, FF4, ED1, ED4). The two-factor GSS target/path decomposition in Appendix D.1.6 reports a smaller $R^2 \approx 0.12$ and isolates a different basis. The span test alone does not separate the three components in equation (43). Two auxiliary bounds discipline the attribution: the slope-unity evidence ($\beta \approx 1$ in Table 6) rules out large components of η_t that covary with m_{4t} via (38), but does not bound the variance of an η_t component orthogonal to m_{4t}; the GSS comparison in Appendix D.1.6 provides a separate check on the basis-approximation channel by projecting onto a different announcement-window basis. The bulk of the unexplained variance is therefore most naturally attributed to the document-vs-market wedge ξ_t^{doc}, conditional on these two bounds rather than on the span regression alone. The trading exercises that follow exploit this residual, interpreting it primarily as a document-vs-market gap.

E.3 IV Coefficients Under Leakage

Section 3 defines the reduced-form coefficient $\theta_h^{\text{text}} = \text{Cov}(y_{t+h}, \hat{s}_t)/\text{Var}(\hat{s}_t)$ and the IV coefficient $\beta_h^{IV} = \text{Cov}(y_{t+h}, \hat{s}_t)/\text{Cov}(x_t, \hat{s}_t)$. Let $L_t \equiv \sum_{j=1}^{J} w_{jt}\, \ell_j\, c_{jt}$ denote leakage entering the reported posterior mean, where $w_{jt} = \tilde{\tau}_j/(\lambda_{0t} + \sum_k \tilde{\tau}_k)$ are the posterior aggregation weights. The measured surprise decomposes as

$$\hat{s}_t = \hat{s}_t^0 - L_t, \qquad \hat{s}_t^0 \equiv u_t + \xi_t^{priv} + \xi_t^{pub} + \eta_t^0,$$

where $\hat{s}_t^0$ is the leakage-free surprise and η_t^0 is the residual extraction error after stripping out L_t. The reduced-form coefficient becomes

$$\theta_h^{\text{text}} = \frac{\text{Cov}(y_{t+h},\, \hat{s}_t^0) - \text{Cov}(y_{t+h},\, L_t)}{\text{Var}(\hat{s}_t^0) + \text{Var}(L_t) - 2\,\text{Cov}(\hat{s}_t^0,\, L_t)}, \tag{45}$$

and the IV coefficient with first-stage regressor x_t becomes

$$\beta_h^{IV} = \frac{\text{Cov}(y_{t+h},\, \hat{s}_t^0) - \text{Cov}(y_{t+h},\, L_t)}{\text{Cov}(x_t,\, \hat{s}_t^0) - \text{Cov}(x_t,\, L_t)}. \tag{46}$$

Leakage contaminates both numerator and denominator: a L_t that correlates with y_{t+h} (e.g., training data encoding future outcomes) biases the reduced-form numerator, and a L_t that correlates with x_t contaminates the first stage. Treating IV bias as the reduced-form contamination "rescaled by the first-stage coefficient" ignores the second channel and is correct only when $\text{Cov}(x_t, L_t) = 0$. The look-ahead cutoff test (Table 18) bears on the empirical relevance of L_t as a whole by comparing extraction behaviour inside and outside the LLM's training window; it does not separate the two contamination channels.

E.4 Omitted-Signal Predictability

A scalar benchmark. Consider a single Gaussian signal $z_t = \theta_t + \psi_t$ with $\psi_t \sim \mathcal{N}(0, \sigma_z^2)$, observed by some agent but not by the pipeline, independent of the document signals and of u_t. Conditional on the $\mathcal{P}_4$ posterior, the surprise is $\hat{s}_t = (\theta_t - m_{4t}) + u_t$, the posterior error has variance v_{4t}, and $\text{Cov}(\hat{s}_t, z_t) = v_{4t}$. The incremental R^2 from adding z_t is

$$\Delta R^2(z) = \frac{v_{4t}^2}{(v_{4t} + \sigma_u^2)(v_{4t} + \sigma_z^2)}. \tag{47}$$

The expression is decreasing in $\tilde{\tau}_j$ (informative documents shrink v_{4t}, leaving less for z_t to predict) and in σ_z^2 (noisier signals predict less).

The two empirical objects below share the same scalar logic — a signal omitted by the documentary pipeline that recovers explanatory power conditional on $\mathcal{P}_4$ — but differ in the structural meaning of the omitted information set. The scalar formula is therefore a benchmark for the order of magnitude one expects, not an exact common structural model.

Public non-document wedge (vector analogue). The B&S predictors are a vector of public observables $Z_t \in \mathcal{M}_t \setminus \mathcal{B}_t$ rather than a single Gaussian signal. The empirical $R^2 = 0.166$ from Section 5.2 provides evidence for a remaining public-information wedge of the same qualitative character as (47) predicts. Persistent components of Z_t may be partially absorbed by the cross-meeting prior through $m_{4,t-1}$, so the wedge is a lower bound on the gap $\mathcal{M}_t \setminus \mathcal{B}_t$.

Private-information wedge (Greenbook, scalar analogue). The Fed's internal Greenbook forecast is closer to a scalar private signal $g_t = \theta_t + \phi_t$ with $\phi_t \sim \mathcal{N}(0, \sigma_G^2)$, with $g_t \in \mathcal{G}_t \setminus \mathcal{M}_t$. Equation (47) applies with $\sigma_z^2 \to \sigma_G^2$. The empirical increment of 11.2 percentage points (Table 7) constrains σ_G^2 given v_{4t} and σ_u^2. This is a calibration target rather than a derived prediction; the model's prediction is that any private signal with non-zero precision strictly improves on the document-based measure.

E.5 Setup, Recursion, and Comparative Statics

The body uses a Normal–Normal conjugate framework for tractability. The latent policy state $\theta_t \equiv \mathbb{E}[\Delta i_t \mid \mathcal{G}_t]$ evolves as $\theta_t = \rho\theta_{t-1} + \omega_t$ with $\omega_t \sim \mathcal{N}(0, \sigma_\omega^2)$, and $\Delta i_t = \theta_t + u_t$ with $u_t \sim \mathcal{N}(0, \sigma_u^2)$ independent of $\mathcal{G}_t$. Standard Kalman algebra gives the cross-meeting predictive prior $\theta_t \mid \mathcal{P}_{4,t-1} \sim \mathcal{N}(\rho m_{4,t-1},\, \rho^2 v_{4,t-1} + \sigma_\omega^2)$ with prior precision $\lambda_{0t} = (\rho^2 v_{4,t-1} + \sigma_\omega^2)^{-1}$, and the recursive within-meeting update upon observing the jth extracted signal with effective precision $\tilde{\tau}_j$:

$$m_{Jt} = m_{J-1,t} + K_{Jt}(\tilde{d}_{Jt} - m_{J-1,t}), \qquad K_{Jt} = \frac{\tilde{\tau}_J}{\Lambda_{J-1} + \tilde{\tau}_J}, \qquad v_{Jt}^{-1} = \Lambda_{J-1} + \tilde{\tau}_J,$$

where $\Lambda_{J-1} = \lambda_{0t} + \sum_{j<J} \tilde{\tau}_j$ is cumulative precision before document J. The effective precision $\tilde{\tau}_j$ in equation (10) composes raw document precision τ_j with Decoder and Forecaster precisions $(\kappa_j^{dec}, \kappa_j^{for})$. Writing the Decoder's structured output as $\theta_t + \varepsilon_{jt} + \varepsilon_{jt}^{dec}$ and the Forecaster's processed signal as $\theta_t + \varepsilon_{jt} + \varepsilon_{jt}^{dec} + \nu_{jt}^{for}$, with the three noise terms mutually independent, additivity of variances gives $\mathrm{Var}(\varepsilon_{jt} + \varepsilon_{jt}^{dec} + \nu_{jt}^{for}) = \tau_j^{-1} + (\kappa_j^{dec})^{-1} + (\kappa_j^{for})^{-1}$, and equation (10) follows by inversion. The leakage term $\ell_j c_{jt}$ is kept separate as a non-classical bias channel.

Sequential learning. If $\tilde{\tau}_j > 0$ for every j, the differential entropy of $\mathcal{P}_J$ falls monotonically: $H(\mathcal{P}_J) - H(\mathcal{P}_{J-1}) = -\frac{1}{2}\log(1 + \tilde{\tau}_J/\Lambda_{J-1}) < 0$. Each document contributes more when the prior is diffuse and less as beliefs tighten. The distributional analysis in Appendix C.1.2 confirms

monotone entropy decline in the data.

Transparency and degradation. The surprise variance is $\text{Var}(\hat{s}_t \mid R_t) = (\lambda_{0t} + \sum_j \tilde{\tau}_j(R_t))^{-1} + \sigma_u^2$. Higher transparency raises τ_j, higher LLM quality raises κ_j^{dec} and κ_j^{for}, and either route lowers surprise variance through the same composite precision channel. The Consensus Economics validation across the 2011–2018 partial-transparency regime and the post-2005 fast-minutes era (Appendix B.2.3) provides the empirical counterpart.