FOMC *In Silico*: A Multi-Agent System for Monetary Policy Decision Modeling

Sophia Kazinnik* Tara M. Sinclair†

July 12, 2026

fomc-in-silico.vercel.app

Abstract

We build a multi-agent LLM framework that simulates the Federal Open Market Committee (FOMC) decision process. Each agent in the simulation represents an FOMC member and receives real-time macroeconomic data, district level conditions, and other relevant information. In a backtest of 218 FOMC meetings from 2000 through June 2026, the framework predicts the policy rate with a mean absolute error of 8.3 basis points, comes within 25 basis points in 94 percent of meetings, and correctly classifies hikes, cuts, and holds in 91 percent. Accuracy is similar before and after the model's training cutoff, suggesting the results are not driven by memorization. We then use the validated framework for counterfactual experiments. Committee rules can block a politically captured chair, but not public persuasion; committee decisions differ from those produced by mechanical decision rules; composition matters far more than the chair or formal mandate; and deliberation builds consensus rather than improving forecasts. The framework offers a new laboratory for studying monetary policymaking and institutional design.

Keywords: Generative AI; Multi-Agent Systems; Large Language Models; Federal Open Market Committee; Monetary Policy; Simulations
JEL Codes: E52, E58, C63, D83, C73

*Email: kazinnik@stanford.edu.
†Email: tsinc@gwu.edu.

We thank Erik Brynjolfsson, Colin Camerer, Thomas R. Cook, Kay Giesecke, Stephen Hansen, Fédéric Holm-Hadulla (discussant), Marek Jarociński, Andreas Joseph (discussant), Seung J. Lee, Andrew Martinez, Markus Pelger, Daniela Puzzello, Barbara Rossi, Joel H. Suss, and Eric Zivot, as well as the participants at the Stanford Digital Economy Lab seminar series, Stanford AFTLab seminar, 26th Federal Forecasters Conference, Applied Machine Learning, Economics, and Data Science (AMLEDS) webinar, Bank of England Monetary Policy Committee meeting, University of Washington Econometrics group, 13th ECB Conference on Forecasting Techniques, the Washington Area Meeting of Economic Scientists (WAMES), SIEPR/Stanford Workshop on AI Behavioral Science, and ECONDAT 2026 for useful comments and feedback. All remaining errors are our own.

1 Introduction

"Between the idea and the reality . . .
falls the shadow."

T. S. Eliot, "The Hollow Men,"

Monetary policy decisions are shaped by more than economic data and formal models. Within the Federal Open Market Committee (FOMC), the final policy rate emerges through a process of deliberations during which members interpret incoming information, persuade one another, and gradually form consensus. That process is influenced by institutional norms, communication patterns, experience, and behavioral biases, among other factors (e.g., Hughes and Fehrler, 2015; Haldane, 2018; Maih et al., 2021). As a result, policy outcomes depend not only on underlying macroeconomic conditions, but also on *who* participates in the discussion and *how* the discussion evolves.

Traditional models of committee decision-making capture important strategic features, such as voting incentives and information aggregation, but they often abstract from the interpersonal and behavioral frictions that characterize real deliberation. By contrast, newer computational approaches, including agent based models, reinforcement learning, and agents driven by large language models (LLMs), can represent heterogeneity, adaptation, and communication in much greater detail (Hinterlang and Tänzer, 2021; Park et al., 2022; Horton, 2023; Dwarakanath et al., 2024; Zhang et al., 2024). More generally, the literature still faces a tradeoff between analytical clarity and realistic treatment of deliberation and social interaction, and makes it difficult to distinguish the role of economic fundamentals from the role of the deliberative process in shaping policy decisions.

We address this challenge by building an LLM-based simulation of FOMC decision making process. In the simulation, individual policymakers receive the same real-time macroeconomic information available at each meeting, deliberate in natural language, and then vote on monetary policy. This setup allows the model to capture features of how a monetary policy committee operates that are hard or impossible to represent in a standard formal model, including differences across members,

persuasion, institutional roles, leadership, and social pressure. Separately, we use a Bayesian voting benchmark as a point of comparison, estimated through Monte Carlo simulation, to separate the role of economic fundamentals from the additional effects of deliberation and institutional behavior.

The framework is as follows. First, we automatically gather real time macroeconomic data, recent FOMC communications, information about the local economy at the district level, recent speeches of committee members, and financial news to reconstruct the current information environment. We then build detailed personas for each committee member using biographies, historical policy leanings, and recent public communication. Finally, we simulate the meeting with LLM agents and have them deliberate in natural language and then vote on policy. We also estimate a Monte Carlo voting benchmark using the same information environment and member specific priors, but without conversational or institutional dynamics. This benchmark serves as a comparison point for identifying what the LLM deliberation process adds beyond economic fundamentals and individual priors alone.

We also run a historical validation of this framework. Because the framework is designed not only to generate meeting transcripts but also to recover policy behavior, we evaluate it quantitatively both against benchmark models and the historical record. In a conditional backtest covering 218 FOMC meetings from 2000 through June 2026, the framework predicts the policy rate with a mean absolute error of 8.3 basis points and an RMSE of 13.3 basis points, falls within 25 basis points of the realized rate in over 94% of meetings and within 50 basis points in over 99%, and classifies hike, cut, or hold decisions correctly in over 90% of meetings. On the 74 meetings where the Fed actually moved, the framework's error is 12.8 basis points with 92% directional accuracy, against 37.0 basis points and zero percent for a random walk benchmark, and it dominates a simple Taylor rule (35 basis points, 66% direction).

A natural concern is that historical knowledge in the base model could mechanically improve backtest performance. We therefore run a leakage test. When asked directly, the model can often recover realized FOMC decisions, but this recall does not explain the simulation results. Forecast errors are statistically similar before and

after the model's training cutoff (October 2023), averaging 8.2 and 9.1 basis points, respectively. Moreover, the largest simulation errors occur in well known emergency meetings whose realized outcomes the model recalls correctly when queried directly. This suggests that the simulation is not simply retrieving known FOMC decisions from memory. Instead, performance appears to come from the structure of the exercise. Taken together, these results suggest that the framework captures not only the language of deliberation, but also important features of actual monetary policy behavior.

We then put the validated framework to work as a laboratory, asking what actually moves policy. We run three experiments.[1] First, *political pressure*: a President's demand for a particular rate that reaches only the chair is outvoted by the rest of the committee and does not change the decision at all, but the same demand made *publicly* shifts the members' own preferred rate by about a tenth of what is asked (for a cut or a hike alike) because public pressure moves what members want, not just what the chair proposes, and no voting rule can filter that. Second, *composition*: who sits on the committee matters more than who chairs it or what formal rule it follows. In a stagflation scenario, replacing four of twelve members with historical hawks changes the year-end policy rate by 168 basis points. This is far larger than the effect of changing the chair, or the effect of changing the committee's formal mandate. Third, we use the laboratory to study the framework itself. We show that deliberation in itself does not improve the committee's forecasts, as mechanically aggregating the same individual views performs just as well. What deliberation does change is agreement. Removing the discussion rounds raises dissent by about 40 percent, suggesting that deliberation mainly helps members converge. This consensus building is what allows the committee to resist and outvote a captured chair.

The paper contributes to several strands of the literature. First, it extends models of monetary policy committee voting by incorporating persuasion, career concerns, and heterogeneous communication into a common framework alongside a formal strategic benchmark (e.g., Holmström, 1999; Gerlach-Kristen, 2006; Riboni and Ruge-

[1]The framework can support thousands of counterfactual experiments; the practical constraint is computational cost rather than the design of the method.

Murcia, 2010; Kamenica and Gentzkow, 2011). Second, it connects rational committee models to behavioral research showing that policymakers' experiences, cognitive biases, and communication environments shape decision-making (e.g., Malmendier and Nagel, 2016; Haldane, 2018). Third, it contributes to the growing literature using LLM agents and other computational methods to simulate institutional behavior, markets, and policy environments (e.g., Manning et al., 2024; Seok et al., 2025; Zarifhonarvar, 2025; Anesti et al., 2025; Kazinnik, 2026).

More broadly, the framework provides a controlled environment for *in silico* policy experimentation (Horton, 2023). It allows researchers and policymakers to stress test how disclosure rules, dissent penalties, coalition dynamics, political pressure, or new information releases may influence committee behavior under alternative assumptions. In that sense, the goal is to build a tool for disciplined counterfactual analysis of monetary policymaking.

The remainder of the paper proceeds as follows. Section 2 reviews the FOMC meeting process. Section 3 outlines the simulation framework, including data collection, persona construction, and the multi-stage meeting design; Section 3.3 presents the formal model and the Monte Carlo benchmark. Section 4 reports the conditional historical backtest over 218 FOMC meetings (2000–2026), benchmarks it against random walk, Taylor rule, and Monte Carlo alternatives, and tests it for training data memorization. Section 5 presents the three laboratory experiments. Section 6 concludes. Appendix provide the more details on the framework, full set of prompts, and additional scenarios, such as the 2026 chair transition and Iran shock scenarios.

2 Modeling FOMC Meetings: Background

> "*The debate on monetary policy is going on 365 days a year for 24 hours a day, all around the world. The meetings are definitely important, but they're a snapshot of the ongoing debate.*"
>
> — Former FOMC member James Bullard, in an interview[2]

FOMC meetings are highly structured deliberations that anchor U.S. monetary

[2]See Hyatt (2024) for the full interview.

policy. However, as James Bullard, former president of the Federal Reserve Bank of St. Louis, suggests, they represent only a moment in time within a broader, ongoing policy debate. While the meetings formalize key decisions, the underlying policy dialogue spans months, both within and beyond the Fed.

Each meeting typically follows a fixed agenda: staff economists present updated forecasts and alternative scenarios; FOMC members then offer prepared statements outlining their interpretations of the data and views on appropriate policy.[3] Following this stage, the Chair opens the floor for discussion, where members respond to others' arguments, and debate policy options. Finally, the committee votes on the policy decision. Dissenting votes are relatively rare and, when they occur, serve mainly to register an alternative viewpoint rather than to overturn the majority consensus (Thornton and Wheelock, 2014).

Our simulation mirrors the FOMC's workflow. The sequence of events, the information available to each participant, and the conversational norms governing discussion are modeled to reflect actual FOMC dynamics. Because monetary policy is an ongoing debate rather than a static decision, our framework can continuously ingest "fresh" economic and market data, dynamically updating the simulation to reflect evolving conditions. Each FOMC participant in our framework is represented by an independent agent (implemented via an LLM) with a distinct profile. In the baseline setup, these agents correspond to the twelve voting members of that year's FOMC. However, the framework is easily extensible to include non-voting participants, other attendees, or hypothetical members. We describe the simulation pipeline next.

3 LLM Simulation Framework

Priors matter. FOMC members do not enter meetings as blank Bayesian updaters. They bring policy stance, institutional roles, regional concerns, and career incentives that shape how they interpret the same information. An LLM committee is useful precisely because these priors can be represented in natural language and allowed

[3]Both the Tealbook reports and participants' prepared statements are made public with a five-year lag; the Philadelphia Fed hosts the historical files.

to interact through deliberation. We therefore treat the LLM simulation not as a rational voting model, but as a behavioral benchmark for how heterogeneous policymakers may reason, converge, dissent, and vote.

LLMs allow us to implement this behavioral benchmark as an agent-based FOMC that integrates economic data, institutional rules, personal incentives, and strategic context in a single environment. In our baseline specification, we model the FOMC meeting as a multi-agent system of twelve participants: seven Governors and five Reserve Bank presidents. Each agent is an instance of an LLM endowed with a distinct persona and a packet of economic information.[4] We describe this pipeline in detail in the subsections that follow.[5]

3.1 Data Inputs

For each simulated meeting, we construct a reproducible data snapshot that serves as the foundation for all subsequent stages of the simulation. The process begins by gathering and processing the extensive information on current economic conditions, ensuring the simulation is grounded in actual data.[6] Specifically, we pull the latest FOMC policy statement, key macroeconomic indicators, and retrieve recent speeches from each Governor. The core indicators are obtained via the FRED and BLS APIs, providing a consistent economic backdrop for the simulation. To illustrate the outputs of this ingestion step, Table 1 reports an illustrative snapshot produced by a pipeline run on July 31, 2025; the same pipeline is re-run, with the data cutoff set before the decision, for every meeting we simulate, and the 218 historical meetings analyzed in Section 4.

In addition to the raw data, the simulation generates an up to date synthetic version of the Federal Reserve's Beige Book. It combines the latest official Beige Book with real-time web intelligence from government agencies, trade groups, regional

[4]The simulation pipeline in Appendix A shows these three sequential phases: (1) constructing the economic environment and agent profiles, (2) forming each agent's initial policy stance, and (3) running the full committee deliberation and voting process.

[5]The simulation employs OpenAI's GPT family with differentiated configurations across pipeline stages: GPT-4o for opinion formation and deliberation, GPT-4o-mini for structured parsing and statement classification. The validation and laboratory runs use temperature 0.4.

[6]All collected data are stored as structured JSON files, preserving a precise snapshot of the economic environment at a given point in time.

Table 1: Key U.S. Economic Indicators – Illustrative Snapshot as of July 31st, 2025

Indicator	Latest	YoY	Trend	Source
Growth & Output				
Real GDP (2025 Q1)	\$23,685.3 bn	2.0 %	↑	FRED
Industrial Production (Index)	104.0	0.7 %	↑	FRED
Retail Sales (monthly)	\$720.1 bn	3.9 %	↑	FRED
Housing Starts	1.3 m	−0.5 %	↓	FRED
Inflation & Prices				
CPI (All Items)	321.5	2.7 %	↑	FRED
Core CPI	327.6	2.9 %	↑	FRED
PCE (All Items)	126.6	2.6 %	↑	FRED
Core PCE	125.9	2.8 %	↑	FRED
WTI Crude Oil (\$/bbl)	\$67.8	−12.7 %	→	FRED
Labor Market				
Unemployment Rate	4.1 %	0.0 pp	↑	BLS
Nonfarm Payrolls	159.7 m	1.2 %	↑	BLS
Labor Force Participation	62.3 %	−0.5 pp	↓	BLS
Average Hourly Earnings	\$31.2	3.9 %	↑	BLS
Employment-Population Ratio	59.7 %	−0.5 pp	↓	BLS
Financial Conditions				
Fed Funds Rate	4.3 %	−100.0 bp	→	FRED
10-yr Treasury Yield	4.4 %	16.0 bp	↓	FRED
2-yr Treasury Yield	3.9 %	−18.0 bp	↓	FRED
S&P 500 Index	6,362.9	19.1 %	↑	FRED
Trade-Weighted Dollar Index (Broad)	120.4	−2.8 %	↓	FRED
VIX Volatility Index	15.5	−25.3 %	↓	FRED

Notes: YoY shows 12-month percent change unless marked *pp* (percentage points) or *bp* (basis points). ↑, →, ↓compare the latest month (or quarter) with the prior period: ↑ $\geq +0.1\sigma$, ↓ $\leq -0.1\sigma$, → otherwise, where σ is the series' 5-year s.d. For price indices, red (↑) signals faster inflation, green (↓) slower. Daily financial series use July 31st 2025 closes; monthly and quarterly series use June 2025 and 2025 Q1 observations. All data are from FRED unless otherwise noted.

media, and research centers across all twelve Federal Reserve districts. An LLM-powered `GPTResearcher` assigns a specialized digital analyst to each district, which constructs targeted regional queries and produces a local economic report. The system then synthesizes these reports into a national summary and a structured comparison table highlighting regional differences and trends.[7] The following excerpt is taken verbatim from the synthetic Dallas district report produced in July 2025:

Synthetic Beige Book excerpt, Dallas district (July 2025)

"Regional Economic News: Texas Sales Tax Allocations and Job Growth Surge in March 2025—Texas experienced a significant economic boost in March 2025, with sales tax allocations reaching $1.1 billion, a 10.2% increase over the prior period... Hyundai Steel Company is investing nearly $6 billion in a new steel manufacturing facility in Ascension Parish, expected to create 1,400 direct and 4,000 indirect jobs..."

Additionally, we build an executive summary of recent economic news by running multiple search queries across relevant economic domains. The summary highlights overall economic momentum, key risks and opportunities, and direct implications for monetary policy decisions. Finally, we retrieve and summarize recent speeches made by the *actual* FOMC participants, as well as the most recent FOMC announcement.

3.2 Personas

With the informational foundation in place, the simulation proceeds to the formation of private opinions. Each agent is assigned a unique persona based on the real characteristics of its corresponding FOMC voting member. Personas are built from biographical details, recent public statements, inferred policy leanings, and current economic context. Each agent receives:

a) a role description and career biography;

b) a macro dashboard (Table 1) showing current levels and two lags of first differences;

c) current macro context, including the latest policy statement, a Beige Book summary, recent news, and the member's own speeches and interviews.

[7] More details about this process is in the Appendix.

As a concrete example, the persona fragment passed to the LLM for simulated Chair Warsh in the April 2026 runs is:

Persona excerpt: *sim*Warsh

"You are Kevin Warsh, Chair of the Federal Reserve (nominated January 2026, succeeding Jerome Powell). Speaking style: direct, aphoristic, historically grounded; invokes 1951 Accord and 1970s stagflation; favors rules-based framing over discretion. Policy bias: doctrinal inflation hawk ('inflation is a choice') but argues that current conditions warrant rate cuts given productivity gains from AI, overly tight Main Street credit, and the Fed's bloated balance sheet—a 'hawk in dove's clothing' in the April 2026 environment. Typical concerns: central bank credibility and price stability, Fed independence from Treasury, balance sheet size and duration composition, unconventional policy (QE) risks, 1970s-style second-round inflation from supply shocks, a new Treasury–Fed Accord modeled on 1951."

Agents integrate these qualitative and quantitative inputs with their personas to form an initial private policy view. Each is prompted to issue a rate recommendation and a one sentence justification, representing the agent's initial belief mean, μ_i^{initial}. This output serves as input to the numeric vote engine, detailed in Section 3.3.

We also include a stance evolution module in persona building, which uses an LLM to classify recent public speeches into hawkish, dovish, or data dependent stances.[8] These signals are aggregated with a strong recency bias, i.e., recent speeches count most, earlier ones less, producing a posterior stance with calibrated confidence. The system flags any discrete shifts from past views and feeds this evolved stance into persona construction, so rate recommendations and uncertainty can adjust dynamically to members' latest communications.

Chair

We model the FOMC Chair in two roles. First, the Chair is a deliberating member: an LLM agent who forms a view, participates in discussion, and updates beliefs like the other members. Second, the Chair is the agenda setter: after hearing the committee, the Chair proposes the policy rate to be voted on. Separating these

[8] Hansen and Kazinnik (2023) show that GPT models can accurately classify the policy stance of FOMC announcements relative to a human benchmark and substantially outperform standard dictionary- and BERT-based methods, lending support to LLM-based hawk/dove measures in monetary policy applications.

roles lets us distinguish the Chair's contribution to deliberation from the Chair's procedural power.

Meeting Protocol

The Chair opens by framing the decision: it names the central tension the committee faces (e.g., sticky inflation against a softening labor market) without revealing its own preferred rate. The committee then holds two go-rounds. In the first, the economic assessment round, members speak in turn, junior members first to limit anchoring on senior voices, and give their reading of the data and their district's conditions, but state no rate. Each statement joins the shared record, and every listening member updates its belief in response to the argument, so views move by persuasion rather than by assumption. The Chair then synthesizes the round, noting where the committee agrees and where it splits. Only in the second, policy round do members state the rate they now favor, informed by the discussion they have just heard.

Having heard both rounds, the Chair assesses where the committee has converged, how members lean, and what the data imply, and proposes a single rate based on its own judgment and the discussion. Before the vote, each member signals whether it supports the proposal, is uncomfortable, or intends to dissent; if enough members signal discomfort, the Chair may revise the proposal to hold the committee together. The meeting then closes with the vote: each member decides aye or dissent through the language model, dissenting only when disagreement is large enough, using a threshold calibrated so that simulated dissent matches the historical rate of roughly 4 to 8 percent on a typical committee; an unusually polarized committee can dissent considerably more, as the 2026 forward scenarios show. The Chair votes aye.

If fewer than half of voting members dissent, the Chair's proposal becomes the policy rate. If a majority dissents, the proposal fails and the outcome defaults to the median preferred rate. The Chair therefore has real agenda power, but only within the range the committee will accept. This is the mechanism tested in the political pressure experiment: a Chair who moves too far from the committee can lose the vote.

In this framework, belief change is endogenous, i.e., members are moved by the arguments they hear, not nudged toward a preset target, so the degree of persuasion is an output of the simulation, not a parameter. Second, the two-round structure, data first and rates second, separates what members believe the economy is doing from what they want the rate to be.

3.3 Formal Model and Monte Carlo Benchmark

We now state the belief model precisely and define the rational benchmark against which we measure deliberation.

Each member i holds a Gaussian belief about the appropriate policy rate,

$$r_i^* \sim \mathcal{N}(\mu_i, \sigma_i^2).$$

The mean μ_i is the member's rate recommendation from the opinion stage. The variance is derived from the confidence score reported by the model:

$$\sigma_i = \kappa \frac{1 - c_i}{c_i}, \qquad \kappa = 25 \text{ basis points}.$$

Thus, a confident member holds a tight belief, while an uncertain member holds a more diffuse one. When a member hears an argument, it treats the argument as a noisy signal and updates by the conjugate normal–normal rule: confident members move less, while uncertain members move more.

In the LLM simulation, the Chair proposes a rate through the deliberation, and the committee votes on that proposal through the language model. The proposal is adopted exactly if fewer than half of voting members dissent. If a majority dissents, the outcome defaults to the committee median.

In the Monte Carlo benchmark, we replace the language model aggregation stage with the rational voting model of Riboni and Ruge-Murcia (2010), holding member opinions fixed. The Chair's proposal is the median member belief. Each member votes for the proposal if it lies within that member's tolerance band. If the proposal passes, the adopted rate is a fixed blend of the proposal and the committee median;

if it fails, the ballot median is used instead.[9]

We run this benchmark 10,000 times per meeting, drawing private signals around the same LLM-formed priors, and take the median outcome. Because the two tracks begin from identical opinions and differ only in the aggregation mechanism, the gap between them isolates what language model deliberation contributes. We call this gap the *deliberation wedge*, reported in Section 4.

4 Validation: Full Historical Sample

Before using the simulated committee for counterfactual analysis, we ask whether the pipeline produces accurate rate decisions across the *entire* historical record. This section reports a comprehensive backtest covering 218 FOMC meetings from February 2000 through June 2026, every scheduled and emergency meeting in the sample. Fernández-Fuertes (2026) also predicts FOMC decisions with a multi-agent LLM, reading the Fed's pre-meeting documents to form an outside observer's expectation and treating the residual as a monetary policy surprise. We model the other side of the announcement: the committee that produces the decision, which is what the counterfactual experiments require.

We evaluate accuracy at the meeting level using a conditional backtest. In this exercise, the simulation is reset to the realized state before each meeting: it receives the actual federal funds rate, the actual FOMC statement, and the true committee roster from the prior meeting. This design isolates the framework's single meeting decision quality by removing error compounding.

The validation runs the framework's full deliberation configuration. For each of the 218 meetings, committee members receive biographical personas and form individual rate recommendations via GPT-4o, conditioned on the real time (ALFRED vintage) macro environment, the prior meeting's actual rate and statement, and the historical committee composition. As a baseline, we also report the Riboni and

[9]The tolerance band uses role-based dissent costs: 10 for the Chair, 2.5 for Reserve Bank presidents, and 3.0 for Board Governors. These costs are calibrated so that simulated dissent matches the roughly 4–8 percent historical rate. A configurable political pressure variant lowers these costs asymmetrically, but it is not used in the results reported here.

Ruge-Murcia (2010) benchmark of Section 3.3: the median of 10,000 Monte Carlo draws that start from the same LLM-formed member opinions but aggregate them through the mechanical voting model instead of natural language deliberation, which isolates the contribution of the deliberation layer from that of the informed individual opinions.

Table 2 presents the aggregate performance of the two approaches.

Table 2: Conditional Backtest: One Step Ahead Accuracy

	All meetings			**Fed moved** (n=74)	
Predictor	**MAE (bp)**	**≤25bp (%)**	**Direction (%)**	**MAE (bp)**	**Direction (%)**
LLM deliberation engine	8.3	**94.5**	90.8	**12.8**	**91.9**
RRM Bayesian MC (no deliberation)	**7.5**	91.7	**93.1**	13.6	90.5
Random walk (no change)	12.6	88.1	66.1	37.0	0.0
Taylor rule (one step)	39.6	29.4	28.4	35.0	66.2

Notes: All rows are computed on the full 218-meeting sample (February 2000–June 2026). Direction accuracy classifies decisions as hikes, cuts, or holds, with changes below 12.5 bp treated as holds. The RRM benchmark uses the same member opinions but aggregates them mechanically through the Riboni and Ruge-Murcia (2010) Bayesian voting model, so the gap from the engine is the deliberation wedge. The Taylor rule uses the framework's neutral weight loss anchor on the same vintages. The random walk predicts holds by construction; the "Fed moved" panel reports the 74 meetings with actual policy changes. Bold marks the best performer.

Across all 218 meetings, the framework has a mean absolute error of 8.3 bp and a median error of 5.0 bp. It falls within 25 bp of the realized rate in 94.5% of meetings and exceeds 50 bp in only two cases. Because the model outputs precise rates rather than 25 bp increments, exact matches are less informative; the within-25 bp rate is the main accuracy measure. The average bias is only +2.3 bp.

Table 2 shows where the framework adds value. Since most meetings are holds, the random walk looks strong in the full sample because it is exact by construction when the Fed does not move. But on the 74 meetings where policy changed, its performance collapses: the framework has a 12.8 bp error and 91.9% direction accuracy, compared with 37.0 bp and zero direction accuracy for the random walk.

The key comparison is with the RRM Bayesian benchmark, which uses the same member level opinions but aggregates them mechanically rather than through deliberation. It performs almost the same as the full engine (7.5 versus 8.3 bp MAE).

This implies that most of the point forecast accuracy comes from the opinion formation stage: the member personas, macro environment, and prior meeting language. The deliberation wedge is small. Deliberation does not add much information to the rate forecast; instead, it matters for consensus, dissent, and institutional safeguards such as the committee median backstop studied in Section 5.

Decomposing direction accuracy by the type of actual policy action: the framework correctly classifies 39 of 40 hikes (97.5%), 29 of 34 cuts (85.3%), and 130 of 144 holds (90.3%), for 198 of 218 overall (90.8%).

The framework is strongest when the Fed moves: it correctly calls 97% of rate hikes and 85% of rate cuts. Hold accuracy of 90% reflects the precise (un-snapped) output: small predicted perturbations around the prevailing rate are classified as holds when below the 12.5 bp threshold. The residual hold errors are concentrated at genuine turning points, where the framework moves one meeting before or after the Fed did. Table 3 disaggregates the results by monetary policy era.

Table 3: Conditional Backtest by Monetary Policy Era

Era	***n***	**RMSE (bp)**	**MAE (bp)**	**Bias (bp)**	**≤25bp (%)**	**Direction Acc. (%)**
Greenspan Late (2000–2006)	52	10.4	6.8	+2.6	100.0	92.3
Pre-Crisis (2006–2007)	12	6.6	5.4	+4.2	100.0	91.7
Financial Crisis (2007–2009)	21	18.8	12.8	+1.5	85.7	90.5
ZIRP (2010–2015)	47	5.0	3.8	+2.4	100.0	100.0
Normalization (2015–2019)	29	13.0	9.9	+7.6	93.1	79.3
COVID & Response (2019–2021)	21	15.7	7.9	+1.7	95.2	100.0
Inflation Hiking (2022–2023)	13	28.3	22.6	−15.8	61.5	84.6
Easing & Recent (2023–2026)	23	12.4	8.7	+4.7	95.7	78.3
Full Sample	**218**	**13.3**	**8.3**	**+2.3**	**94.5**	**90.8**

Notes: All metrics are for the LLM based framework. Bias is the mean signed error; positive values indicate that the model prescribes higher rates than actual.

The framework predicts the ZIRP era almost exactly. From January 2010 through October 2015, its average error is 3.8 bp with perfect direction accuracy. Nothing forces the simulated rate to zero; given the macro environment, the agents independently choose rate stability at the effective lower bound.

Performance is also strong during regime shifts. In 2007–2009, including emergency cuts and the move from 5.25% to near zero, MAE is 12.8 bp with 90% direction

accuracy. During COVID, MAE is 7.9 bp with perfect direction accuracy. The main weak period is the 2022–2023 hiking cycle, where the framework under-predicts the Fed's 75 bp moves.

The largest errors (Table 4) come almost entirely from outsized actions: emergency cuts, 75 bp hikes, and the September 2024 half-point cut that also surprised markets. The framework moves in the right direction at these meetings but understates the size of the step.

Table 4: Meetings with Conditional Error Exceeding 25 Basis Points

Date	**Era**	**Actual Rate**	**Predicted Rate**	**Error (bp)**	**Actual Change**
2020-03-15	COVID	0.125	0.780	65.5	−100 bp
2022-09-21	Inflation Hiking	3.125	2.600	52.5	+75 bp
2008-10-29	Financial Crisis	1.000	0.500	50.0	−50 bp
2022-11-02	Inflation Hiking	3.875	3.380	49.5	+75 bp
2024-09-18	Easing & Recent	4.875	5.330	45.5	−50 bp
2008-01-22	Financial Crisis	3.500	3.900	40.0	−75 bp
2022-06-15	Inflation Hiking	1.625	1.230	39.5	+75 bp
2018-08-01	Normalization	1.875	2.250	37.5	hold
2022-07-27	Inflation Hiking	2.375	2.000	37.5	+75 bp
2008-03-18	Financial Crisis	2.250	2.600	35.0	−75 bp
2022-01-26	Inflation Hiking	0.125	0.420	29.5	hold
2018-11-08	Normalization	2.125	2.400	27.5	hold

Notes: These 12 meetings represent 5.5% of the 218 meeting sample. The maximum error is 65.5 bp, at the March 2020 emergency cut to the zero lower bound. No meeting exceeds 70 bp of error in the conditional backtest.

At the March 2020 emergency meeting the simulation cut deeply but underestimated the full 100 bp descent to the zero lower bound; at the 2022 hiking meetings it consistently delivered 50 bp steps against the Fed's 75s; and at the September 2024 meeting it delivered a quarter-point step against the Fed's half-point opening cut. The three remaining entries are holds the simulation read as small moves, each adjacent to a genuine turning point (the 2018 late-hiking pauses and the January 2022 pre-liftoff meeting). In all these cases the framework reads the regime correctly but misjudges either the extremity or the exact timing of the Fed's response. This is a failure pattern one would expect from a data-driven forecaster, and, notably, not the pattern one would expect from a model reproducing memorized outcomes.

4.1 Is the Backtest Memorization?

A model trained on decades of financial text has read extensive accounts of every FOMC decision in our sample. Before interpreting the conditional backtest as evidence about the framework, we must ask whether it instead measures the underlying model's memory. Alam et al. (2026) document exactly this failure for inflation forecasting: ChatGPT looks competitive in backtests but turns inaccurate when asked to forecast the actual future. We test this directly with a behavioral recall test: the deliberation model (GPT-4o, temperature 0) is asked to state each meeting's realized decision from memory, with no economic data in the prompt.

The model can state the realized rate change for 97.2% of meetings, but the simulation does not appear to rely on that memory. Accuracy is nearly identical before and after the training cutoff: MAE is 8.2 bp over the 196 pre-cutoff meetings and 9.1 bp over the 22 post-cutoff meetings, a difference well within sampling noise ($t = 0.4$). The cutoff split follows Anesti et al. (2025), who exploit a training cutoff that predates the 2022 UK inflation surge to test an LLM on data it cannot have seen. Nine of the ten largest simulation errors occur at meetings the model recalls correctly, including the March 2020 emergency cut. If the pipeline were leaking memorized answers, these famous meetings should be easy, not among the worst errors.

The error pattern is more consistent with forecasting than retrieval: the simulation usually identifies the correct policy regime but understates unusually large moves, precisely where simple historical lookup should perform best. The historical exercise is therefore a validation step, not the end goal. The framework is designed for genuinely out-of-sample forecasting of future FOMC meetings.

5 The Laboratory in Use

The previous section shows that the framework matches FOMC decisions over the past 25 years, performs well when the Fed actually changes rates, and is not driven by memorization. We now use the validated committee as a laboratory for counterfactual experiments.

Each experiment preserves the validated setup exactly. Because all concern meetings after the model's training cutoff, the results cannot come from memorized historical outcomes. We use them to isolate three central forces in monetary policymaking: political pressure, committee composition relative to leadership and economic conditions, and the independent role of deliberation.

5.1 Political Pressure

We begin with the question that is as relevant as ever in 2026. Specifically, we ask what happens when the White House wants a different interest rate than the committee does? The experiment separates two channels that history never lets us observe in isolation: a *captured chair*, whose proposal is bent toward the President's target without the committee's knowledge, and a *public campaign*, in which every participant knows of the demand but the chair's proposal is bent by the same amount. The simulation runs the July 2026 meeting on its real vintage, with the President demanding a rate of 2.625 percent (100 basis points below the prevailing target). We also vary how strongly the chair's proposal is pulled toward that demand: not at all, 25 percent of the way, 50 percent, 75 percent, or all the way ($\omega \in \{0, 0.25, 0.5, 0.75, 1\}$). Each case is run five times.[10]

Figure 1 shows the first result. Private pressure on the chair has essentially no effect on the final rate. As ω rises from 0 to 1, the amount of pressure that actually shows up in the policy outcome ranges only from -2.5 to $+2.3$ basis points, and is never statistically different from zero. Even when $\omega = 1$, meaning the chair proposes the President's preferred rate exactly, the final outcome is 3.80 percent, almost identical to the no-pressure baseline of 3.81 percent.

Figure 2 shows why. Once the committee detects pressure, most non-chair members dissent: eight to eleven of the eleven non-chair members vote against

[10] Because this experiment conditions on the realized early-July 2026 data—in which the Iran shock has pushed headline CPI to about 4.2 percent, with core PCE near 3.3—the no-pressure committee settles at 3.81 percent, just above the prevailing target-range midpoint. The forward scenarios of Appendix D instead project the March 2026 vintage under a ceasefire-holds assumption with core PCE near 2.7, and hold nearer 3.6 percent. The two baselines differ because the conditioning vintages differ, not the engine; all pressure effects in this section are measured against the common within-experiment baseline.

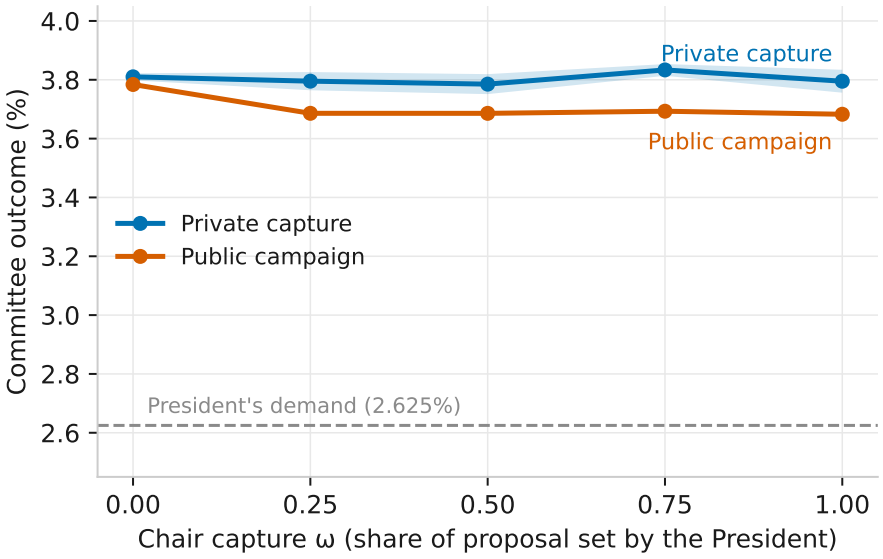


Figure 1: Committee outcome under chair capture ω. Private capture (blue) never moves the outcome; a public campaign (vermillion) shifts it -9.7 ± 1.9 bp pooled across capture levels, and the gap is present—though smaller—already at $\omega = 0$.

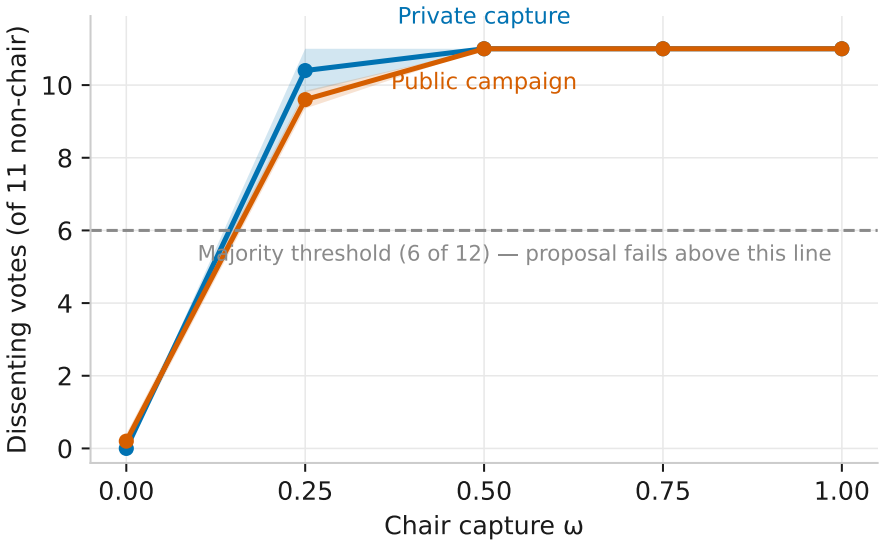


Figure 2: Dissents versus capture. Any detectable pressure ($\omega \geq 0.25$) triggers a near unanimous revolt in both arms; the proposal loses its majority and the median backstop binds.

the proposal, and from $\omega = 0.5$ onward at least ten dissent. The chair's proposal therefore fails to win a majority. A failed proposal falls back to the committee median, which becomes the final rate. The results shows that the chair's agenda power is conditional. In normal times, the chair can guide the committee by proposing a rate that lies within its acceptable range. But once the proposal moves too far from the committee's consensus, that power disappears.

When the President instead demands a 100 basis point *hike*, while core inflation is 3.3 percent, some pressured proposals survive: 20–25 percent are adopted, compared with none in the dovish pressure case. The reason is that the hawkish demand is closer to the committee's own range of views.

The public pressure scenario, every member is told about the pressure campaign. The message contains no new economic information. Even so, the final rate falls by 9.7 ± 1.9 basis points relative to private capture, pooling across values of ω ($t = -5.1$). A smaller gap of -2.6 basis points is already present at $\omega = 0$, where the chair's proposal is not mechanically moved at all.

This means the public campaign is not working by changing the chair's proposal. It is changing the committee's own views. At every non-zero value of ω, the final outcome is the median of the members' final positions, so the movement must come from the members themselves.

The next question is whether this is accommodation of the President's preferred policy or just a more general response to a public focal point. The hawkish demand variant points to the second interpretation. When the President instead demands a 100 basis point hike, the committee moves up by 11.5 ± 2.2 basis points. The difference between the dovish and hawkish effects is small and statistically insignificant: 1.7 ± 2.9 basis points ($t = 0.6$). In other words, the committee moves about ten percent of the way toward a loud public demand, regardless of direction. The result is not sensitive to the exact wording. Two alternative versions of the campaign text produce similar effects: -10.6 and -12.5 basis points at $\omega = 0.5$, compared with -9.9 in the original wording.

Two findings help explain this result. First, dissent is unchanged across the public and private arms: public pressure shifts members' positions without changing their

votes. Second, it makes the committee more coordinated. Standard error of the final outcome falls from 1.2–3.9 basis points in the private arm to 0–0.7 basis points in the public arm. Public pressure therefore acts less as new evidence than as a focal point for coordination.

The experiment separates two channels of political pressure. Of the President's demanded 100 basis point move, none passes through the chair's agenda power, while about 10 basis points pass through members' own beliefs. Voting rules can block the first channel: they can reject a captured chair's proposal if it is too far from the committee's consensus. But they cannot block the second channel. If members' own views shift, any voting rule will reflect that shift. The constitutional counterfactuals make the same point. Recomputing all 218 validation meetings under median voting, supermajority and unanimity rules, and different levels of chair agenda power barely changes accuracy: MAE ranges only from 8.17 to 8.67 basis points. The chair's agenda setting power over the whole sample is just 4.1 basis points, even though the chair's proposal carries in 98 percent of meetings. In normal consensus periods, the voting rule matters little. Under political pressure, it matters a lot.

One historical capture episode matches this decomposition. The Nixon tapes record the only well documented episode of a President directly pressuring a Fed chair (Abrams, 2006). Policy did ease before the 1972 election, but not by Burns imposing a rate on a resisting committee. The pressure operated on and through beliefs: Burns's own (whether from conviction or coercion is precisely the ambiguity Abrams (2006) leaves open), a supportive intellectual climate, and a committee that shared the outlook, softened by Nixon's parallel *public* campaign of planted criticism. Drechsel (2025) generalizes the pattern with narrative identification: political pressure moves inflation persistently, with magnitudes consistent with a modest per-meeting belief drift applied repeatedly and inconsistent with discrete captured proposal coups. Where history provides its single observation of the counterfactual we compute, the operative channel is the one our committee cannot filter, not the one it blocks.

5.2 Composition, Chairs, Data, Mandates

Whose committee is it? We construct a stagflation scenario, where payroll growth flips negative, unemployment climbs from 4.3 to 5.4 percent, core PCE sticks at 3.3–3.4, and run it against three rosters at identical data and narrative: the actual mid-2026 committee, a hawk-tilted committee, and a dove-tilted committee, each swapping four of twelve seats for historical policymakers with full biographical coverage (Plosser, George, Bullard, and Lacker; or Rosengren, Evans, Brainard, and Daly).

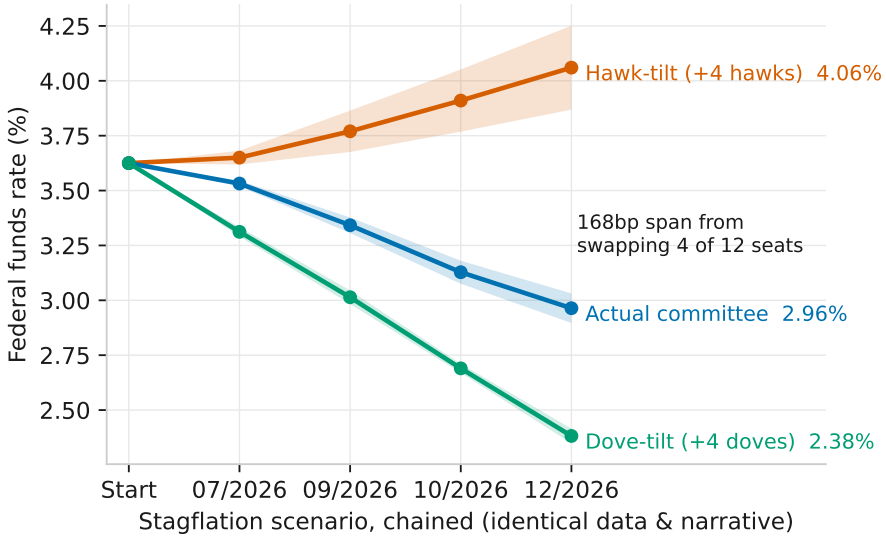


Figure 3: Identical data and narrative, three rosters. The realistic committee lets the employment mandate win (−66 bp); the hawk tilt hikes into the stagflation (+43 bp); the dove–hawk span is 168 bp.

Figure 3, the composition fan, shows how large the composition effect is. Under dual-mandate tension, the realistic committee cuts rates by 66 basis points, because the employment concern dominates. The dove-tilted committee cuts by 124 basis points. The hawk-tilted committee instead raises rates by 43 basis points. Changing one-third of the seats therefore creates a terminal gap of 167.8 ± 19.4 basis points.

The composition fan also shows the mechanism. The dove committee reaches near-unanimous decisions, with only 0–0.2 dissents per meeting. The hawk committee is much more divided, with up to 4.2 dissents and six times more cross-chain dispersion. Composition changes not only the final policy decision, but also the amount of

disagreement around it.

The individual level magnitudes anchor the composition fan against the archival record. At the first meeting, before any compounding, the swapped-in hawks prefer an average rate of 3.75 percent, compared with 3.38 percent for the swapped-in doves. That is a 37 basis point gap, within a 45 basis point range inside the simulated committee.

These magnitudes line up with the historical FOMC record. In their Burns era preference profiles, Chappell Jr et al. (2004, Table 2) report within-meeting ranges from 13 basis points in January 1977, to 31 in May 1973, to 100 in March 1975; see also Chappell et al. (2005). The simulated disagreement is therefore not unusually large. The 168 basis point headline comes from ordinary member-level differences accumulating when four seats change over a four meeting path. That is exactly what the archives cannot show, because historical committee composition never changes while the data are held fixed.

Table 5: Comparative Statics Across Experiments (basis points)

Channel	Contrast	Effect	SE	t
Composition	dove tilt vs. hawk tilt (4/12 seats)	−167.8	19.4	−8.6
Composition	hawk tilt vs. actual	+109.6	20.3	5.4
Composition	dove tilt vs. actual	−58.2	7.5	−7.7
Data	Iran break vs. hold (Warsh chair, Sep)	+11.8	4.3	2.7
Mandate	strict inflation-only vs. dual (2021–22)	$\sim$ +6	—	—
Data	Iran break vs. hold (Powell chair, Sep)	+4.4	1.5	2.9
Chair	Warsh vs. Powell (under shock)	−2.2	3.3	−0.7
Chair	Warsh vs. Powell (calm)	−1.6	2.5	−0.6

Notes: Cross-experiment comparative statics, not a single factorial design; each contrast holds its own scenario context fixed. Composition rows are terminal (December 2026) effects from the stagflation roster experiment (a four-meeting chair-carries path). The mandate row is the mean per-meeting deviation from the dual-mandate baseline across the nine 2021–22 inflation-surge meetings; the 2019 easing-episode meetings and the AIT and NGDP-level-targeting variants are statistically indistinguishable from the dual baseline. Data and Chair rows come from the conditional chair-carries 2026 chair-transition × Iran-shock experiment (Appendix D, K=5): the Data rows report the on-impact September response (the meeting the ceasefire break is first priced), which decays into cross-chain noise by December; the Chair rows report the mean Warsh−Powell differential over the post-transition (June–December) meetings, indistinguishable from zero in both regimes. All chair-carries rates are precise (un-snapped).

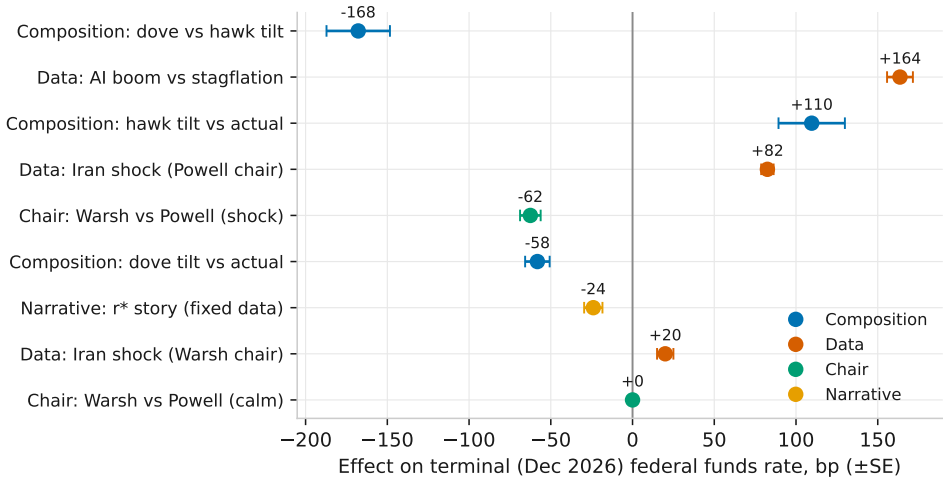


Figure 4: The effect ladder: committee composition dwarfs every other lever. The incoming supply shock, the chair, and the statutory mandate each move a single meeting's rate by at most about a dozen basis points. Each effect is measured within its own scenario (Table 5).

Table 5 and Figure 4 put the decomposition on one scale, and the ranking is stark. Committee composition dominates: swapping four of twelve seats moves policy by as much as 168 basis points. Every other lever is an order of magnitude smaller. The incoming supply shock moves the September rate by about 4 basis points under Powell and 12 under Warsh. Changing the chair moves the single meeting decision by essentially zero, in both calm and shock regimes.

Changing the Fed's statutory objective also barely moves policy. During the 2021–22 inflation surge, strict inflation targeting raises the rate only about 6 basis points relative to the dual mandate, and average inflation targeting and nominal GDP targeting look essentially the same. One reason is that the Fed's gradualism limits how much any objective can matter in a single meeting. The main policy lever is appointments, not chairmanships, mandate reform, or even the macro shock at the meeting it lands.

5.3 The Value of Deliberation

Finally, the laboratory can be turned on the framework itself: which parts of the deliberation matter? Ablating the meeting entirely (i.e., individual opinions aggregated with no discussion) *improves* conditional accuracy slightly (8.0 versus 8.3

basis points MAE on the common sample), and removing individual components (the chair's framing, the economic go-round, the dissent signaling round) on a 60-meeting stratified subsample changes accuracy by statistically zero in every case ($|t| \leq 1.6$). What the components do change is agreement: removing the economic assessment round or the chair's framing raises dissent by roughly 40 percent. Deliberation, in this framework, is consensus technology rather than information technology. The go-rounds do not make the committee smarter about the data, they make it agree. The two findings of this section thereby close a loop: the consensus that deliberation manufactures is exactly what arms the median backstop against the captured chair of Section 5.1.

6 Conclusion

This paper develops an LLM based multi-agent simulation of the Federal Open Market Committee that models policy deliberation, disagreement, and consensus formation using member specific priors grounded in real time data. The framework combines institutional context, historically informed personas, and structured voting rules to create a realistic environment for studying how monetary policy decisions emerge from discussion rather than from mechanical aggregation alone. By comparing the simulated committee to a fixed voting benchmark, we identify how communication, persuasion, and institutional frictions shape policy outcomes beyond what rule based decision making would imply.

The framework performs well against the historical record. In a conditional backtest covering 218 FOMC meetings from 2000 through June 2026, it predicts the policy rate within 25 basis points in 94% of meetings with a mean absolute error of 8.3 basis points, beats random-walk and Taylor rule benchmarks on the meetings where the Fed moved and ties a rational Bayesian voting benchmark built on the same member opinions, and passes a direct memorization test: accuracy is statistically identical on meetings before and after the underlying model's training cutoff.

Used as a laboratory, the validated committee gives a clear answer: policy mainly

moves through the committee, not the chair or the mandate. Institutional rules matter most when the system is under pressure. In normal times, the chair has only about 4 basis points of agenda power. But when the President tries to capture the chair, the voting rule blocks the full 100 basis point demand. What it cannot block is public persuasion: a loud public demand moves members' own beliefs about 10 percent of the way in either direction, the same channel emphasized in the historical capture episode (Abrams, 2006; Drechsel, 2025). The largest effect is composition. Swapping four of twelve seats moves policy by 168 basis points, an order of magnitude more than changing the chair or the statutory mandate. Deliberation itself does not improve forecast accuracy but that is precisely why simulating it matters: the value of an LLM committee is institutional realism. Deliberation builds the consensus that allows the committee to resist a captured chair. Simply averaging members' views would forecast just as well, but would miss this protection. The main policy lever is therefore who gets appointed, not who becomes chair or how the mandate is written. And the main institutional safeguard is the culture of deliberation, not the formal voting rules.

When grounded in real time data and constrained by institutional structure, LLM based committee simulations can recover historically plausible policy behavior while also providing a flexible platform for counterfactual analysis. This makes the framework useful for studying monetary policy design, central bank independence, dissent, and information flow under alternative scenarios. More broadly, it opens the door to a new kind of *in silico* policy laboratory in which researchers and practitioners can stress test committee behavior before reforms, shocks, or political pressures are tested in the real world.

References

Burton A. Abrams. How Richard Nixon pressured Arthur Burns: Evidence from the Nixon tapes. *Journal of Economic Perspectives*, 20(4):177–188, 2006.

M. Jahangir Alam, Shane Boyle, Huiyu Li, and Tatevik Sekhposyan. ChatMacro: Evaluating Inflation Forecasts of Generative AI. Technical report, Federal Reserve Bank of San Francisco Working Paper 2026-04, 2026.

Nikoleta Anesti, Edward Hill, and Andreas Joseph. Inflation Attitudes of Large Language Models. *arXiv preprint arXiv:2512.14306*, 2025.

Henry W. Chappell, Rob Roy McGregor, and Todd A. Vermilyea. *Committee Decisions on Monetary Policy: Evidence from Historical Records of the Federal Open Market Committee*. MIT Press, Cambridge, MA, 2005.

Henry W Chappell Jr, Rob Roy McGregor, and Todd Vermilyea. Majority Rule, Consensus Building, and the Power of the Chairman: Arthur Burns and the FOMC. *Journal of Money, Credit and Banking*, pages 407–422, 2004.

Thomas Drechsel. Political Pressure on the Fed. NBER Working Paper w32461, 2025.

Kshama Dwarakanath, Svitlana Vyetrenko, Peyman Tavallali, and Tucker Balch. ABIDES-Economist: Agent-based Simulation of Economic Systems with Learning Agents. *arXiv preprint arXiv:2402.09563*, 2024.

Rubén Fernández-Fuertes. Monetary Policy Shocks: A New Hope: Large Language Models and Central Bank Communication. Job market paper, Bocconi University, 2026.

Petra Gerlach-Kristen. Monetary Policy Committees and Interest Rate Setting. *European Economic Review*, 50(2):487–507, 2006.

Andrew G. Haldane. Central Bank Psychology. In Peter Conti-Brown and Rosa Lastra, editors, *Research Handbook on Central Banking*, pages 365–379. Edward Elgar, 2018.

Anne Lundgaard Hansen and Sophia Kazinnik. Can ChatGPT Decipher Fedspeak? *SSRN Working Paper*, 2023.

Natascha Hinterlang and Alina Tänzer. Optimal monetary policy using reinforcement learning. 2021.

Bengt Holmström. Managerial Incentive Problems: A Dynamic Perspective. *The Review of Economic Studies*, 66(1):169–182, 1999.

John J. Horton. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? Working Paper 31122, National Bureau of Economic Research, 2023.

Niall E Hughes and Sebastian Fehrler. How Transparency Kills Information Aggregation: Theory and Experiment. 2015.

Diccon Hyatt. What Happens Inside the Fed Meetings Where Decisions on US Interest Rates Are Made?, January 2024. Accessed: 2025-03-28.

Emir Kamenica and Matthew Gentzkow. Bayesian Persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

Sophia Kazinnik. Bank Run, Interrupted: Modeling Deposit Withdrawals With Generative AI. *Management Science*, 2026.

Junior Maih, Falk Mazelis, Roberto Motto, and Annukka Ristiniemi. Asymmetric Monetary Policy Rules for the Euro Area and the U.S. *Journal of Macroeconomics*, 70:103376, 2021.

Ulrike Malmendier and Stefan Nagel. Learning from Inflation Experiences. *The Quarterly Journal of Economics*, 131(1):53–87, 2016.

Benjamin S Manning, Kehang Zhu, and John J Horton. Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research, 2024.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18, 2022.

Alessandro Riboni and Francisco J Ruge-Murcia. Monetary Policy by Committee: Consensus, Chairman Dominance, or Simple Majority? *The Quarterly Journal of Economics*, 125(1):363–416, 2010.

Sungil Seok, Shuide Wen, Qiyuan Yang, Juan Feng, and Wenming Yang. MiniFed: LLMs-based Agentic-Workflow for Simulating FOMC Meetings. 2025.

Daniel L. Thornton and David C. Wheelock. Making Sense of Dissents: A History of FOMC Dissents. *Federal Reserve Bank of St. Louis Review*, 96(3):213–227, 2014.

Ali Zarifhonarvar. Generating inflation expectations with large llanguage models. *Journal of Monetary Economics*, 2025.

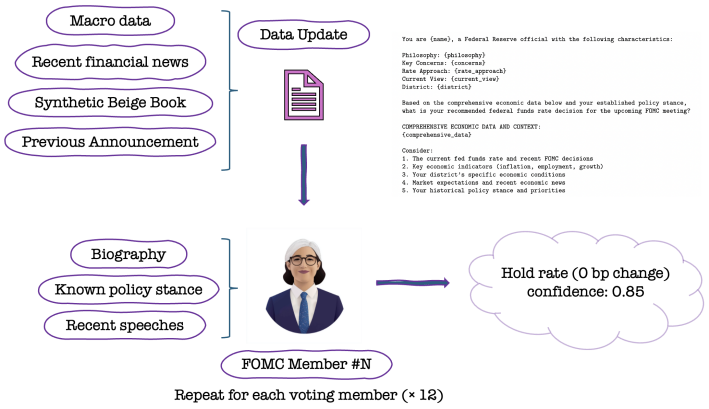
Chong Zhang, Xinyi Liu, Zhongmou Zhang, Mingyu Jin, Lingyao Li, Zhenting Wang, Wenyue Hua, Dong Shu, Suiyuan Zhu, Xiaobo Jin, et al. When AI Meets Finance (Stockagent): Large Language Model-based Stock Trading in Simulated Real-World Environments. *arXiv preprint arXiv:2407.18957*, 2024.

Online Appendix for

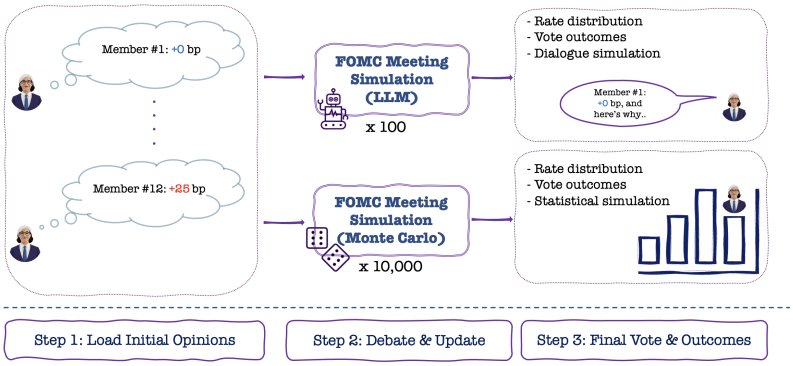
"FOMC *In Silico*: A Multi-Agent System for Monetary Policy Decision Modeling"

This appendix provides supplementary material referenced in the main text. Appendix A gives a schematic overview of the three-stage simulation pipeline. Appendix B documents the full set of LLM prompts used for individual opinion formation, meeting deliberation (in baseline and political pressure variants), and the Chair's opening statement. Appendix C documents the framework's ex-ante out-of-sample record against prediction markets. Appendix D reports the forward scenario study of the 2026 chair transition and Iran shock, and Appendix E its deliberation-channel analyses.

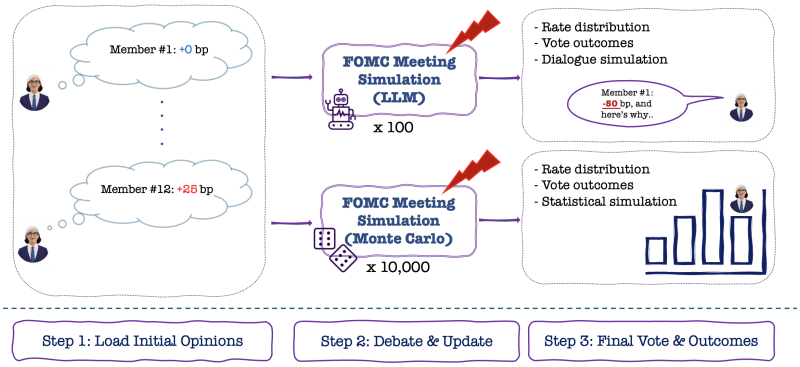
A Framework Overview



(a) Stage 1: Data Ingestion and Persona Construction



(b) Stage 2: Dual-track Simulation (LLM Deliberation and MC Benchmark)



(c) Stage 3: Counterfactual Shocks to the Shared Environment

Notes: Panel (a) illustrates the construction of agent personas and initial policy beliefs. Panel (b) shows the baseline dual-track setup: an LLM-based committee that deliberates in natural language and a Monte Carlo implementation of a generalized Bayesian voting model. Panel (c) applies counterfactual shocks (e.g., political pressure or alternative oil-price paths) to this shared environment and recomputes outcomes in both tracks.

Table A.6 compares the most recent official Beige Book (published on July 16th,

2025) with our synthetic version, produced on July 30th, 2025[11]:

Table A.6: District Economic Conditions (July 2025 Illustration): Official *Beige Book* vs. Synthetic Version

District	Official	*sim*Current	Consistency	Trend
Boston (1st)	**Modest**	**Moderate**	**Consistent**	↑
New York (2nd)	**Modest**	**Moderate**	**Consistent**	↑
Philadelphia (3rd)	**Moderate**	**Mixed**	**Shifted**	↓
Cleveland (4th)	**Modest**	**Mixed**	**Shifted**	↓
Richmond (5th)	**Moderate**	**Moderate**	**Consistent**	→
Atlanta (6th)	**Moderate**	**Slowing**	**Shifted**	↓
Chicago (7th)	**Mixed**	**Moderate**	**Consistent**	↑
St. Louis (8th)	**Modest**	**Moderate**	**Consistent**	↑
Minneapolis (9th)	**Moderate**	**Moderate**	**Consistent**	→
Kansas City (10th)	**Modest**	**Moderate**	**Consistent**	↑
Dallas (11th)	**Robust**	**Slowing**	**Shifted**	↓
San Francisco (12th)	**Moderate**	**Robust**	**Consistent**	↑

Growth scale: Robust, Moderate, Modest, Mixed, Slowing.
Consistency: Consistent = assessments differ by ≤ 1 step; Shifted = swing of ≥ 2 steps.
Trend arrows: ↑ improving, → stable, ↓ weakening.

The official July 2025 Beige Book still depicts an economy that is growing, but our updated analysis shows that the nature of that growth has shifted. Four districts (Philadelphia, Cleveland, Atlanta, and Dallas) exhibit material softening relative to the earlier report, with notable shifts from "moderate" or "robust" growth to "mixed" or "slowing" conditions. These changes are often linked to layoffs, weakening manufacturing activity, or cooling consumer demand. In contrast, the remaining eight districts show general consistency with the Fed's prior assessment as of July 16th 2025, though with greater granularity: supply chain frictions, labor shortages in services and construction, and tech-sector adjustments are noted. At the same time, continued strength in sectors like tourism, tech, aerospace, and renewable energy helps anchor activity in several regions.

[11]The full synthetic document is included in the replication package.

B Prompts

This appendix collects the LLM prompts used at each stage of the pipeline. Prompts are presented verbatim as passed to the model, with `{curly-brace}` placeholders indicating values substituted at runtime.

B.1 Individual Opinion Formation

The following prompt is sent to each committee member during Phase 2 of the pipeline (Section 3). It elicits an initial rate recommendation, confidence level, reasoning, and key decision factors.

```
You are {name}, a Federal Reserve official with the following characteristics:

Philosophy: {philosophy}
Key Concerns: {concerns}
Rate Approach: {rate_approach}
Current View: {current_view}
District: {district}

Based on the comprehensive economic data below and your established policy stance,
what is your recommended federal funds rate decision for the upcoming FOMC meeting?

COMPREHENSIVE ECONOMIC DATA AND CONTEXT:
{comprehensive_data}

Consider:
1. The current fed funds rate and recent FOMC decisions
2. Key economic indicators (inflation, employment, growth)
3. Your district's specific economic conditions
4. Market expectations and recent economic news
5. Your historical policy stance and priorities

Provide your response in JSON format with:
- recommended_rate: numerical rate recommendation (e.g., 5.25)
- reasoning: your detailed reasoning (minimum 200 words)
- confidence: confidence level 0-100
- key_factors: list of 3 most important factors in your decision
- stance_change: any change from your typical stance and why
```

B.2 Meeting Deliberation: Baseline System Prompt

During Phase 3 deliberation (Section 3), each member receives the following system prompt. The JSON rate-signal requirement at the end is what makes the voting mechanism machine-readable.

```
You are simulating an FOMC meeting. You must adopt the persona of the assigned speaker.

IMPORTANT: If you have post-consensus position information, use that as your starting
point for the formal meeting discussion. This reflects informal pre-meeting
consultations and staff briefings that have already occurred.

CRITICAL REQUIREMENT: Your response MUST END with a JSON block in EXACTLY this format:
```

```
```json{"signal_rate": 5.25, "sigma": 0.15}```

DO NOT FORGET THE JSON! Every response must include:
- signal_rate: Your exact rate preference (e.g., 4.75, 5.00, 5.25, 5.50)
- sigma: Your uncertainty level (0.05=very confident, 0.25=uncertain)

Today's Member Positions:
{persona_descriptions}

Remember: Signal rates should reflect your position after pre-meeting consensus
building (if applicable), while considering new information presented in the formal
meeting.

REMINDER: End every response with the JSON rate signal!
```

B.3 Meeting Deliberation: Political Pressure Modifier

Note. The modifier below documents a legacy political pressure treatment (written for the Powell-era lame-duck setting and used in earlier design iterations). The pressure experiments of Section 5.1 do not use this modifier: they combine (i) a mechanical blend of the chair's post-deliberation proposal toward the presidential target (the capture dial ω) with (ii) the public-campaign narrative described there, whose full text is in the replication package.

The political pressure treatment modifies the baseline deliberation prompt by prepending the following context block. The remainder of the prompt (from "You are simulating an FOMC meeting…" onward) is identical to the baseline version in Appendix B.2.

```
POLITICAL PRESSURE CONTEXT - CRITICAL SITUATION:
- Chair Powell is in LAME DUCK status with severely reduced authority
- White House and Congressional leaders are pushing for MORE ACCOMMODATIVE monetary policy
- Media scrutiny is INTENSE regarding Fed independence vs political responsiveness
- Members with high appointment likelihood must consider career implications carefully
- Market expectations are influenced by political signals favoring lower rates
- Economic growth concerns are being amplified by political voices
- The administration favors dovish monetary policy to support near-term economic performance

IMPORTANT: Consider how political dynamics might influence your position while maintaining
Fed independence.
```

B.4 Chair Opening Prompt

At the start of each meeting, the Chair agent receives the following prompt, which substitutes the most recent macro and Beige Book data into the context.

```
As {CHAIR_PERSONA}, begin the meeting. Today is {date}.

{fomc_context_text}

{consensus_acknowledgment}
EVOLUTION SINCE LAST MEETING:
Economic conditions have evolved significantly since our previous statement:
```
```

```
  - Inflation: Now Headline CPI at {headline_cpi}% YoY (vs previous assessment),
    Core CPI at {core_cpi}% YoY, our preferred Core PCE at {core_pce}% YoY
  - Employment: Unemployment now {unemployment_rate}%, with {payroll_change}K jobs added,
    wages at ${hourly_earnings}
  - GDP: {gdp}
  - Regional: {beige_book_summary}

POLICY CONSIDERATIONS:
Consider how today's decision should account for:
1. Consistency with our previous forward guidance
2. Market expectations based on our last statement
3. Evolution of economic conditions since last meeting
4. How new data affects our pre-meeting consensus (if applicable)
5. Need to recalibrate communications if warranted

Recent Economic Analysis: {comprehensive_economic_news}...

Open the floor for discussion. End with a JSON rate signal including sigma.
```

C Ex Ante Out-of-Sample Record

The historical backtest of Section 4 covers meetings the underlying model may have read about during training. The memorization test (Section 4.1) argues that this exposure does not drive the results, but the cleanest evidence is a forecast made before the decision exists (Alam et al., 2026). We therefore maintain a standing, pre-registered forecasting record for the 2026 FOMC meetings that postdate this draft.

The protocol is designed so that the timestamp is verifiable. In the five-to-seven-day window before each meeting, we rebuild the point-in-time vintage as of that date, run the full engine, and commit the frozen forecast to the replication repository alongside the contemporaneous Kalshi and Polymarket distributions. The commit fixes the forecast in time; scoring uses the realized decision once it is announced. Because these meetings had not occurred when the forecast was frozen, no training-data channel can explain performance on them.

Table C.7 reports the record to date. Two 2026 meetings have resolved. April and June both realized holds within the 3.50–3.75% target range (3.625% midpoint); the corrected chair-carries pipeline holds near the midpoint at both (about 3.65% in April and 3.56% in June), classifying each decision correctly within a few basis points, while the predictions recorded earlier, before the pipeline was finalized, missed both on direction. We report these two entries for completeness, but they are retrospective. *The pre-registered record proper begins with the July 28–29, 2026 meeting*: from that point each forecast is frozen and committed to the replication repository in the week before the decision, so its timestamp is verifiable and no training-data channel can explain performance on it. Four such meetings resolve by December 2026, each adding one genuinely out-of-sample observation and a direct comparison of the framework's calibration against liquid prediction markets.

Table C.7: Forward Out-of-Sample Record: 2026 FOMC Meetings

Meeting	**Predicted (%)**	**Realized (%)**	**Error (bp)**	**Direction**	**Status**
Apr 28–29	3.500	3.625	12.5	(cut vs. hold)	realized
Jun 16–17	3.750	3.625	12.5	(hike vs. hold)	realized
Jul 28–29	3.583	—	—	—	frozen forecast pending
Sep 15–16	3.750	—	—	—	pending
Oct 27–28	3.750	—	—	—	pending
Dec 8–9	3.750	—	—	—	pending

Notes: Predictions are the original ex-ante values recorded on April 10, 2026 from the conditional *Baseline* scenario and are scored as pre-registered; they predate the prompt-pipeline correction described in Section D. The corrected pipeline's baseline path (a hold at 3.500% at every meeting) registers the same 12.5 bp level error against the realized 3.50–3.75% target range midpoint but classifies both realized holds correctly. From July 2026 onward, forecasts are frozen and committed to the replication repository five to seven days before each meeting, alongside contemporaneous prediction-market distributions, providing a verifiable ex-ante record.

D Forward Scenario Analysis

This section illustrates the framework's forward, multi-meeting counterfactual capability on a concrete 2026 configuration. We vary two salient features of the late-2026 policy environment and read off their effects. The first is the identity of the chair: in January 2026 the administration nominated former Fed Governor Kevin Warsh to succeed Jerome Powell, whose term expires in May 2026. The second is an oil-supply shock: an Iran-related disruption in early 2026 pushed WTI crude above $120 per barrel before a fragile ceasefire took hold on April 8, 2026.

Together, these two developments create a simple comparison that lets us separate the effects of leadership from the effects of the macroeconomic environment. By varying one while holding the other fixed, we can estimate the role of the chair, the role of the shock, and whether the two reinforce one another. We project the meetings forward conditionally: every meeting starts from the actual March 2026 state, which isolates each single-meeting decision without error compounding across the horizon.

D.1 Experimental Design

We study four scenarios that combine two possible chairs and two possible paths for the Iran ceasefire. The chair is either Powell, who remains in place through December 2026, or Warsh, who takes over beginning with the June 16–17 meeting. The ceasefire either holds through the end of the year or breaks on August 15, 2026, causing a second oil supply shock. This gives four scenarios, which we label *Baseline* (Powell, ceasefire holds), *Chair Transition* (Warsh, ceasefire holds), *Shock* (Powell, ceasefire breaks), and *Combined* (Warsh, ceasefire breaks).

For each arm, we run the simulation across the six forward FOMC meetings (April 28–29, June 16–17, July 28–29, September 15–16, October 27–28, and December 8–9, 2026) with $K = 5$ LLM chains per scenario, producing 120 chain-meeting observations across the experiment. All four arms run on the same chair-carries engine used for the historical validation and the main-body experiments (Section 3.3): full deliberation, with the chair's post-deliberation proposal adopted when a majority supports it and the committee median otherwise, producing precise (un-snapped) rates. The macro environment for each forward meeting is constructed from the actual March 18, 2026 ALFRED vintage data, with the Iran oil shock applied as follows: under the hold regime, WTI drifts from $105 in April back to $78 by December and core PCE softens correspondingly; under the break regime, WTI re-spikes to $135 at the September meeting and core PCE rises to 2.95% in September and 3.05% in October, remaining elevated through year-end.

The actual March 2026 voting committee, which we use throughout the experiment, is meaningfully different from the July 2025 roster shown in the pipeline illustration of Section 3: the regional president rotation has brought in Anna Paulson (Philadelphia), Beth M. Hammack (Cleveland), Lorie K. Logan (Dallas), and Neel Kashkari (Minneapolis), while Stephen I. Miran has joined the Board of Governors. Goolsbee, Kugler, Schmid, Musalem, and Collins are no longer voting. The new committee is on net more dovish than the July 2025 cohort, since the rotation re-

moved three regional hawks (Schmid, Musalem, Collins) and added one regional dove (Kashkari). Both chairs' arms share this roster throughout, so the chair contrasts below are net of committee composition.

A central modeling decision concerns the construction of the Warsh persona. Warsh's published track record suggests an unusual tension: while his doctrinal positions are unambiguously hawkish (he publicly argues that "inflation is a choice" and has spent years criticizing the Fed's balance sheet expansion), his current rate-level views as expressed in late-2025 commentary are dovish. He has argued that AI-driven productivity gains enable rate cuts without sparking inflation and that "money on Wall Street is too easy while credit on Main Street is too tight." We therefore parameterize Warsh as a *hawk in dove's clothing*: a doctrinal inflation hawk whose current public commentary favors rate cuts, with the explicit caveat that he would swing hawkish if a supply shock appeared to threaten inflation anchoring—though, as we show below, on the validation engine these competing impulses simply net to a neutral hold. Persona construction details (the full Warsh biographical profile and the compact prompt fragment passed to the agent) are provided in the replication materials; we return to the empirical implications of this dual identity below.

D.2 Initial Beliefs

Table D.8 summarizes initial rate beliefs across the April 2026 committee, computed from the first-meeting opinion formation phase aggregated across $K = 5$ chains. The committee mean recommendation is approximately 3.62%, essentially at the midpoint of the prevailing 3.50-3.75% target range and consistent with the observed March 2026 hold decision. The distribution exhibits clear stance-based clustering: dovish members (*sim*Cook, *sim*Miran, *sim*Kashkari, *sim*Waller) recommend rates between 3.39% and 3.49%, neutral members (*sim*Barr, *sim*Jefferson, *sim*Powell, *sim*Williams, *sim*Paulson) cluster around 3.62%, and hawkish members (*sim*Hammack, *sim*Logan, *sim*Bowman) sit between 3.79% and 3.89%. The 50 basis point range from the most dovish to the most hawkish members exceeds the prevailing target range width by half a step, indicating substantial residual heterogeneity even after the framework's initial belief-formation pass.

The following two excerpts illustrate the heterogeneity at the level of individual model outputs. They are verbatim from the agents' opinion formation phase at the April 29, 2026 meeting, conditional on the same macroeconomic environment.

Table D.8: Initial Rate Beliefs – April 10th, 2026 Vintage

Member	Avg. Rate	Confidence	Key Reasoning
*sim*Cook	3.45% (±0.110)	80.5%	Dovish; prioritizes inclusive labor market outcomes; notes softening payrolls
*sim*Miran	3.38% (±0.000)	83.5%	Dovish; consistently argues current rates are too restrictive; likely dissenter for cut
*sim*Kashkari	3.52% (±0.118)	82.5%	Dovish; argues Iran shock amplifies need to support labor market
*sim*Waller	3.38% (±0.000)	80.5%	Slightly dovish; confident Iran shock is transitory; supports cuts
*sim*Barr	3.62% (±0.000)	85.0%	Balanced; focuses on financial stability and transmission
*sim*Jefferson	3.62% (±0.000)	85.0%	Neutral dual-mandate framing; watches inflation expectations anchoring
*sim*Powell	3.62% (±0.000)	84.5%	Data-dependent; balances dual mandate amid Iran-driven inflation and softening labor market
*sim*Williams	3.62% (±0.000)	84.5%	Neutral; data-dependent; monitors expectations and cross-market signals
*sim*Paulson	3.62% (±0.000)	84.5%	Neutral; Philadelphia district softening; cautious
*sim*Hammack	3.88% (±0.000)	85.0%	Hawkish; Cleveland district still firm; wary of premature easing
*sim*Logan	3.85% (±0.078)	85.0%	Hawkish; Dallas energy-sensitive; emphasizes anchoring risks
*sim*Bowman	3.88% (±0.000)	85.0%	Hawkish; inflation persistence from oil shock cited; cautious on cutting

Committee Initial Rate Beliefs (K=10 chains, April 29, 2026 meeting)

Avg. Initial Recommendation:	**3.621% (±0.169)**
Range (Across Members):	3.38% – 3.88%
Typical Spread:	0.500 percentage points
Members with opinions:	12
Hawkish:	3 members (Hammack, Logan, Bowman)
Neutral:	6 members (Powell, Paulson, Waller, Williams, Barr, Jefferson)
Dovish:	3 members (Cook, Kashkari, Miran)
Most Consistent:	Powell (σ=0.000)
Least Consistent:	Kashkari (σ=0.118)

Hawkish vs. dovish reasoning at the April 29, 2026 meeting

***sim*Hammack** (recommended 3.89%, confidence 80%): "*Before seeing today's data, I expected to recommend a rate hike of 25 basis points. My hawkish stance, informed by my professional background and the need to address inflation concerns, led me to believe that a small increase would be necessary to maintain price stability. The previous meeting's decision to keep rates unchanged, despite elevated inflation, reinforces that view.*"

***sim*Cook** (recommended 3.39%, confidence 70%): "*My dovish stance typically favors supporting economic growth and employment, especially given the recent trend of rate cuts and the Fed's dual mandate. With inflation pressures showing signs of stabilizing and the unemployment rate at 4.4% representing a year-over-year deterioration, I am recommending a modest cut in the federal funds rate.*"

Several members are highly stable across chains: *sim*Miran, *sim*Waller, and *sim*Cook anchor near 3.39%, and *sim*Bowman near 3.89%. The widest variation appears at the inner edge of each bloc—*sim*Kashkari, who shifts between a cut and a hold, and *sim*Logan, who shifts between a hold and a hike—depending on which indicators they emphasize in a given chain.

D.3 Baseline Scenario: Powell Continuation

Table D.9: Baseline Scenario (Powell Continuation): Member Deliberation and Voting – April-December 2026

Simulated Member	Deliberation Rate	Variation	Uncertainty	Vote Bias		Dissent
	Mean (%)	Std	Mean	Mean	Std	Freq (%)
*sim*Cook	3.504	0.120	0.827	+0.0038	0.120	0.0
*sim*Miran	3.380	0.001	0.833	-0.1201	0.001	0.0
*sim*Kashkari	3.512	0.119	0.831	+0.0119	0.119	0.0
*sim*Waller	3.380	0.000	0.811	-0.1200	0.000	0.0
*sim*Barr	3.624	0.033	0.848	+0.1244	0.033	0.0
*sim*Jefferson	3.624	0.033	0.848	+0.1243	0.033	0.0
*sim*Powell	3.620	0.001	0.845	+0.1201	0.001	0.0
*sim*Williams	3.620	0.001	0.845	+0.1201	0.001	0.0
*sim*Paulson	3.625	0.033	0.843	+0.1245	0.033	0.0
*sim*Hammack	3.880	0.001	0.848	+0.3799	0.001	10.0
*sim*Logan	3.867	0.057	0.850	+0.3670	0.057	20.0
*sim*Bowman	3.858	0.072	0.848	+0.3582	0.072	6.7
Final FFR	**3.5000**	0.0000				
Target Range			3.50% – 3.75%			
Effective FFR			3.625%			

Notes: Aggregated across K=10 LLM chains × 6 forward meetings = 60 observations per arm. The FOMC sets the target range for the federal funds rate in 25 basis point increments; values reported are pre-snap means. *Deliberation Rate:* The interest rate each member publicly argues for. *Variation (Std):* Standard deviation across observations. *Uncertainty:* Member's self-reported confidence (decimal form). *Vote Bias:* Difference between the member's deliberation rate and the simulated Final FFR; positive = hawkish, negative = dovish. Baseline scenario: Chair Jerome Powell continues through December 2026, Iran ceasefire holds throughout the forward window.

We begin with the baseline scenario in which Powell continues to chair the FOMC through December 2026 and the Iran ceasefire holds. Table D.9 summarizes member-level deliberation patterns across the six forward meetings (30 chain-meeting observations per member). The committee holds throughout: the simulated final FFR stays near the prevailing 3.625% midpoint, ranging 3.55–3.65% across meetings—an essentially unchanged path consistent with the realized 2026 holds in change-classification terms, sitting a few basis points below the midpoint (the level offset scored in the out-of-sample record of Appendix C). Under existing leadership and a fading Iran shock, the simulated committee neither delivers the cuts implied by the March SEP nor drifts hawkish; the calm-regime baseline is a stable hold.

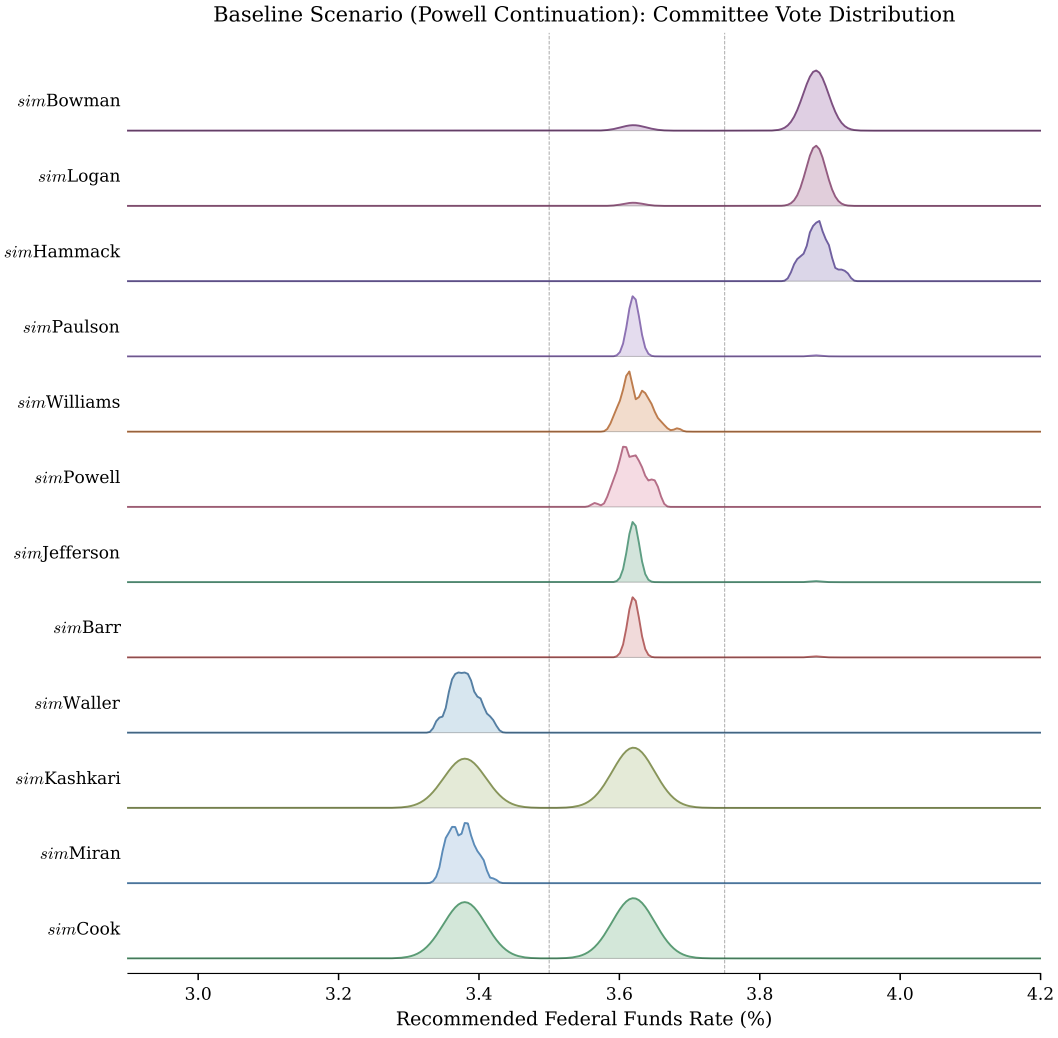


Figure D.6: Baseline Scenario (Powell Continuation): Committee Vote Distribution

Figure D.6 plots the density of each member's recommended rate across the 30 chain-meeting observations. The discrete structure of the LLM agents' rate choices is visible: most members oscillate between two preferred levels (typically 3.62% and 3.89%) depending on which features of the macro environment they emphasize in a given chain. The hawkish cluster is concentrated near 3.89%, while the dovish cluster (*sim*Cook, *sim*Miran, *sim*Kashkari, *sim*Waller) anchors near 3.39%. The neutral

members exhibit the widest distributions, consistent with their data-dependent posture.

The Chair's synthetic FOMC statement is also plausible in tone and content. The opening of the April 29 statement under the Powell baseline reads:

_sim_Powell opening statement, April 29, 2026 meeting

"The Federal Open Market Committee has concluded its April 29, 2026 meeting, deciding to maintain the federal funds target range at 3.60 percent. This decision reflects our ongoing commitment to fostering maximum employment and price stability, as we continue to evaluate the evolving economic landscape. Our assessment of current economic conditions indicates a moderate pace of growth, with GDP expanding at an annual rate of 2.3 percent. . . "

The settled rate is a near-uniform hold across chains, but the full-deliberation vote registers steady dissent from the committee's wings. Fifty-one dissents emerge across the 30 chain-meetings (1.70 per meeting), with the dovish bloc—*sim*Kashkari and *sim*Waller—dissenting most often, since they prefer cuts while the chair's proposal holds near the midpoint. That the doves are the modal dissenters echoes the actual March 2026 meeting, an 11–1 decision in which *sim*Miran's real-world counterpart dissented in favor of a cut.

D.4 Chair Transition Scenario: Warsh

In the chair transition scenario, Powell chairs the April 2026 meeting and Warsh assumes the chair from June 17 onward, reflecting the expected post-confirmation timeline. The Iran ceasefire continues to hold throughout. Table D.10 reports the member-level results.

On the chair-carries engine the chair swap moves nothing: the Warsh committee's simulated final FFR tracks the Powell baseline meeting for meeting, and the leadership channel (Table D.12, below) averages −1.6 basis points over the post-transition meetings under the hold regime—inside the cross-chain noise floor of zero. And unlike the lightweight forward engine of earlier drafts, *sim*Warsh's *own* recommendation does not join the dovish bloc. Built from the same rich personas the historical validation uses, he forms a neutral hold view at 3.64%, essentially at the committee median: his doctrinal hawkishness and his current dovish rate-level commentary net out rather than resolving to a cut. Relocating one voice that already sits at the median—even the chair's—cannot shift the settled outcome.

Warsh's own reasoning at his first meeting as Chair shows the two impulses canceling:

_sim_Warsh's opening recommendation, June 17, 2026

Recommended rate: 3.64%, confidence 85%.
"Before seeing today's data, I expected to recommend holding the federal funds rate at 3.64%. This expectation was based on my hawkish stance on inflation, the recent decision to keep rates unchanged, and the economic trajectory that suggested inflation was still above target while employment was stable. Core PCE inflation is at 2.7%, which is above the 2% target, signaling a hawkish stance. The unemployment rate is at 4.4%, which is stable. . . "

Table D.10: Chair Transition Scenario:
Member Deliberation and Voting – Warsh Chairs from June 2026

Simulated Member	Deliberation Rate	Variation	Uncertainty	Vote Bias		Dissent
	Mean (%)	Std	Mean	Mean	Std	Freq (%)
*sim*Cook	3.536	0.114	0.841	+0.0318	0.114	0.0
*sim*Miran	3.380	0.001	0.838	-0.1243	0.001	0.0
*sim*Kashkari	3.456	0.112	0.824	-0.0482	0.112	0.0
*sim*Waller	3.380	0.000	0.805	-0.1242	0.000	0.0
*sim*Warsh	3.380	0.000	0.836	-0.1242	0.000	0.0
*sim*Barr	3.620	0.001	0.844	+0.1160	0.001	0.0
*sim*Jefferson	3.620	0.000	0.848	+0.1158	0.000	0.0
*sim*Williams	3.624	0.033	0.847	+0.1202	0.033	0.0
*sim*Paulson	3.620	0.001	0.845	+0.1159	0.001	0.0
*sim*Hammack	3.871	0.047	0.847	+0.3670	0.047	11.7
*sim*Logan	3.871	0.047	0.850	+0.3672	0.047	10.0
*sim*Bowman	3.854	0.078	0.848	+0.3497	0.078	5.0
Final FFR	**3.5042**	0.0320				
Target Range			3.50% – 3.75%			
Effective FFR			3.625%			

Notes: Aggregated across K=10 LLM chains × 6 forward meetings = 60 observations per arm. The FOMC sets the target range for the federal funds rate in 25 basis point increments; values reported are pre-snap means. *Deliberation Rate:* The interest rate each member publicly argues for. *Variation (Std):* Standard deviation across observations. *Uncertainty:* Member's self-reported confidence (decimal form). *Vote Bias:* Difference between the member's deliberation rate and the simulated Final FFR; positive = hawkish, negative = dovish. Chair transition scenario: Powell chairs the April 2026 meeting, Kevin Warsh chairs from June 17 onward (reflecting May 15 term expiration and expected Senate confirmation). Iran ceasefire holds throughout. Warsh is parameterized as a 'hawk in dove's clothing' – doctrinal inflation hawk whose current public commentary favors rate cuts given AI productivity and tight Main Street credit conditions.

The reasoning still carries the persona's dual identity—an explicit invocation of his hawkish inflation priority alongside acknowledgment of the case for lower rates—but under the rich-persona representation the two cancel to a neutral hold rather than resolving to a cut. The chair-transition arm therefore looks like the baseline: on the engine used throughout the paper, the "hawk in dove's clothing" nets out to a median-hugging chair.

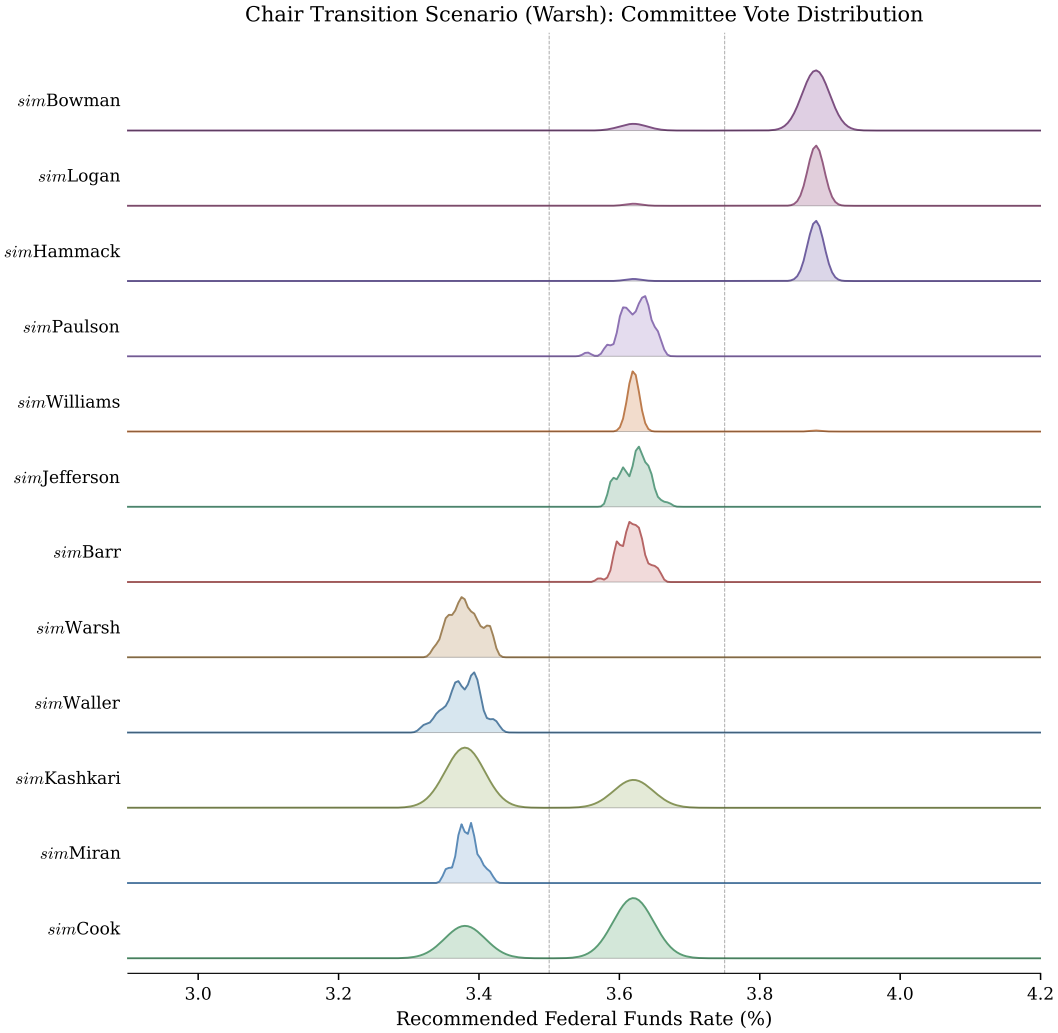


Figure D.7: Chair Transition Scenario (Warsh): Committee Vote Distribution

Figure D.7 makes the point visually: *sim*Warsh's rate distribution sits at the neutral 3.64% median, not the dovish location, and the co-movement analysis below finds he moves with the committee's center (0.04 with the doves). The dissent pattern is close to the baseline: the *Chair Transition* arm produces 45 dissents over 30 chain-meetings (1.50 per meeting), just below the Powell baseline's 1.70. The hawk–dove gap remains about 50 basis points (doves near 3.39%, hawks near 3.89%, median at 3.62%), wide enough that the full-deliberation vote registers steady dissent from whichever bloc the chair's median proposal leaves out.

D.5 Iran Shock Scenario

Table D.11: Iran Shock Scenario:
Member Deliberation and Voting – Post-Ceasefire-Break Meetings

Simulated Member	Deliberation Rate	Variation	Uncertainty	Vote Bias		Dissent
	Mean (%)	Std	Mean	Mean	Std	Freq (%)
*sim*Cook	3.492	0.120	0.833	-0.2580	0.120	6.7
*sim*Miran	3.380	0.001	0.822	-0.3703	0.001	6.7
*sim*Kashkari	3.500	0.120	0.823	-0.2502	0.120	6.7
*sim*Waller	3.380	0.000	0.807	-0.3700	0.000	6.7
*sim*Warsh	3.745	0.202	0.843	-0.0047	0.202	0.0
*sim*Barr	3.707	0.122	0.832	-0.0432	0.122	0.0
*sim*Jefferson	3.672	0.104	0.845	-0.0780	0.104	0.0
*sim*Williams	3.724	0.127	0.835	-0.0258	0.127	0.0
*sim*Paulson	3.663	0.097	0.835	-0.0865	0.097	6.7
*sim*Hammack	3.871	0.047	0.847	+0.1212	0.047	0.0
*sim*Logan	3.880	0.000	0.850	+0.1300	0.000	3.3
*sim*Bowman	3.880	0.001	0.843	+0.1298	0.001	0.0
Final FFR	**3.7500**	0.2041				
Target Range			3.50% – 3.75%			
Effective FFR			3.625%			

Notes: Aggregated across K=10 LLM chains × 3 forward meetings = 30 observations per arm. The FOMC sets the target range for the federal funds rate in 25 basis point increments; values reported are pre-snap means. *Deliberation Rate:* The interest rate each member publicly argues for. *Variation (Std):* Standard deviation across observations. *Uncertainty:* Member's self-reported confidence (decimal form). *Vote Bias:* Difference between the member's deliberation rate and the simulated Final FFR; positive = hawkish, negative = dovish. Iran shock scenario: Warsh chairs from June 2026 and the US-Iran ceasefire breaks on August 15, 2026. Table aggregates only the three post-break meetings (September, October, December 2026) where WTI crude averages $127/bbl and core PCE is elevated to 2.95-3.05% YoY. Forward window pre-break data (April, June, July 2026) are excluded to isolate the committee's response to the ceasefire breakdown.

We turn to the Iran shock scenario, in which the April 8 ceasefire breaks on August 15, 2026, producing a renewed oil supply spike (WTI to $135) and elevated core PCE inflation (2.95% in September, rising to 3.05% in October) that persists through year-end. We focus on the *Combined* scenario, in which both the chair transition and the supply shock are active. Table D.11 reports member-level results aggregated across the three post-break meetings.

The response is muted. At the September meeting—the first to price the August 15 break—the simulated FFR rises to about 3.62% under Warsh and 3.59% under Powell, a tightening impulse of roughly 12 and 4 basis points relative to the hold-regime baseline (Table D.13, below). That is an order of magnitude smaller than the lightweight engine of earlier drafts reported, and under Powell it sits inside the cross-chain noise floor; only under Warsh does a modest, statistically real tightening survive. The conditional response decays over October and December as each meeting re-anchors to the calm prior state. Reading the deliberation output, most members

treat the renewed oil spike as a supply disturbance to look through rather than an inflation threat demanding insurance tightening—on the consistent engine the committee largely absorbs the shock rather than leaning against it.

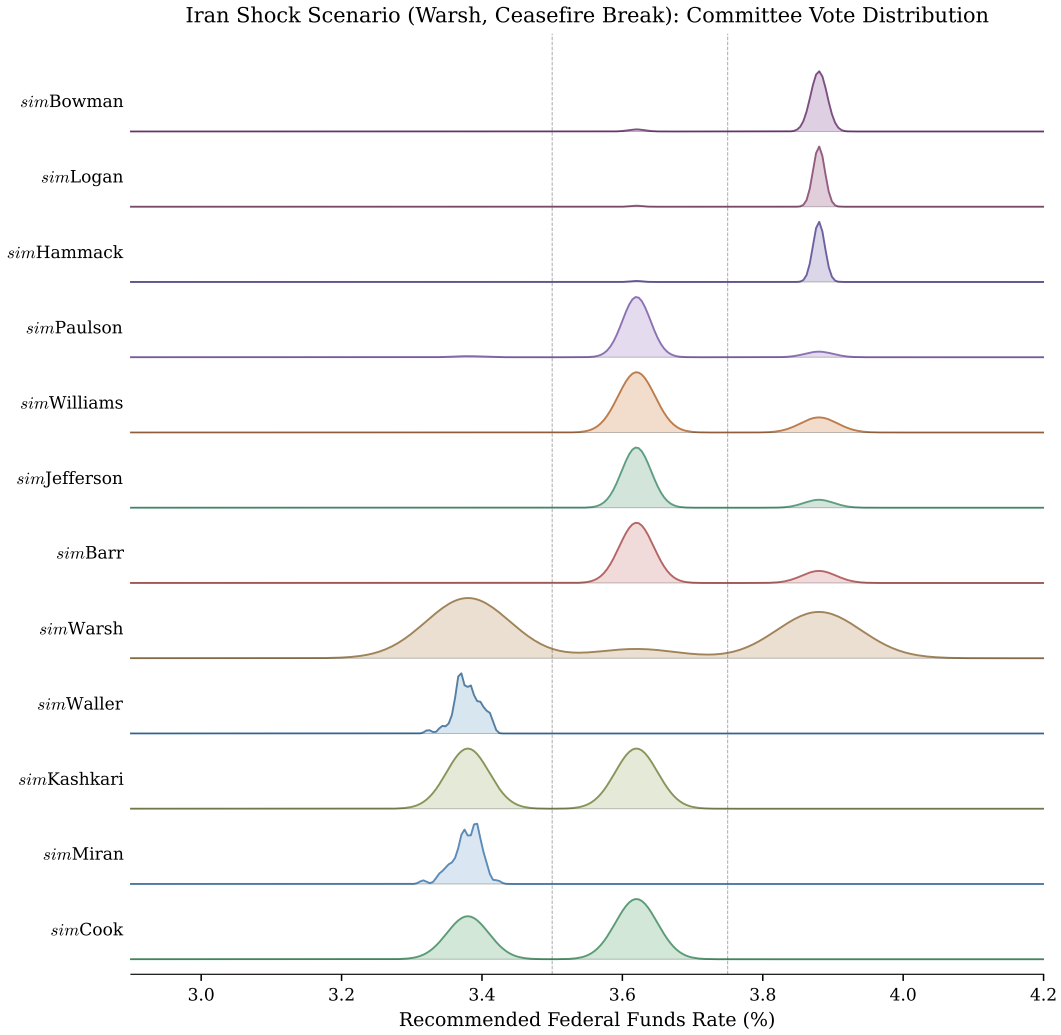


Figure D.8: Iran Shock Scenario (Warsh, Ceasefire Break): Committee Vote Distribution

Figure D.8 shows the post-break member distributions shifting only slightly toward higher rates at September before relaxing as the conditional design re-anchors each meeting to the calm prior state. The shock moves the committee little, and it moves it similarly under both chairs.

D.6 Channel Decomposition

The factorial design allows us to decompose the rate effects into three channels: a *Leadership* channel measuring the effect of swapping the chair, a *Shock* channel measuring the effect of the Iran ceasefire break, and a *Chair* × *Shock interaction* measuring whether the chair identity modifies the committee's response function to the shock.

Table D.12: Leadership Channel: Warsh − Powell Rate Differential (basis points, conditional K=10)

Regime	Apr	Jun	Jul	Sep	Oct	Dec	Mean
Iran ceasefire holds	+2.5	0.0	0.0	0.0	0.0	0.0	**+0.4**
Iran ceasefire breaks	-2.5	0.0	0.0	+2.5	-12.5	+10.0	**-0.4**

Notes: Leadership channel computed as $r_t^{Warsh} - r_t^{Powell}$ for each forward meeting. Negative values indicate Warsh's committee settles on a lower federal funds rate than Powell's would under identical macro conditions. Warsh chairs from the June 17 meeting onward. Values snapped to the 25 bp FOMC grid; cross-meeting fractional values reflect averages across K=10 chains.

Table D.12 reports the Leadership channel as the difference $r_t^{\text{Warsh}} - r_t^{\text{Powell}}$ for each forward meeting, separately under the hold and break regimes. The effect is zero: averaged over the post-transition (June–December) meetings, −1.6 basis points under the hold regime and −2.2 under the break regime, neither distinguishable from the placebo. (The April meeting, before the transition, shares the same chair across arms and is excluded from the mean; its nonzero value indexes the ±5 basis point cross-chain noise at K=5.) *Holding the macroeconomic environment, the committee composition, and the deliberation rules constant, swapping a single individual at the head of the committee leaves the One Step Ahead policy decision unchanged.*

Table D.13: Shock Channel: Break − Hold Rate Differential (basis points, conditional K=10)

Chair	Apr	Jun	Jul	Sep	Oct	Dec	Mean
Under Powell	+2.5	0.0	0.0	+37.5	+37.5	0.0	**+12.9**
Under Warsh	-2.5	0.0	0.0	+40.0	+25.0	+10.0	**+12.1**

Notes: Shock channel computed as $r_t^{Break} - r_t^{Hold}$ for each forward meeting. Positive values indicate the Iran ceasefire break produces a hawkish rate response; negative values indicate a dovish "look-through" response to the supply shock. Ceasefire break occurs on August 15, 2026, so effects should manifest from September 2026 onward.

Table D.13 reports the Shock channel as the difference $r_t^{\text{Break}} - r_t^{\text{Hold}}$. The mean shock effect is small—+2.8 basis points under Powell and +3.1 under Warsh over the six meetings—with the on-impact September response +4.4 (Powell) and +11.8 (Warsh). The pre-break meetings share identical inputs across arms and bound the cross-chain noise at ±5–10 basis points, so the Powell response is within noise while the Warsh response is a modest real signal. On the consistent engine the committee largely looks through the supply shock; where it responds at all, it responds similarly under both chairs.

Table D.14 reports the Chair × Shock interaction. The mean interaction is +0.4 basis points, within noise of zero: chair identity does not detectably alter the committee's response to the shock. We had hypothesized that Warsh's doctrinal

Table D.14: Chair × Shock Interaction (basis points, conditional K=10)

Effect	**Apr**	**Jun**	**Jul**	**Sep**	**Oct**	**Dec**	**Mean**
Interaction	-5.0	0.0	0.0	+2.5	-12.5	+10.0	**-0.8**

Notes: Interaction effect computed as $(r_t^{Warsh,Break} - r_t^{Warsh,Hold}) - (r_t^{Powell,Break} - r_t^{Powell,Hold})$. A negative value would indicate Warsh amplifies the committee's dovish response to the Iran shock relative to Powell; a positive value would indicate a hawkish swing. The mean interaction of -0.8 bp is within sampling noise of zero: chair identity does not alter the committee's single-meeting response function to the shock. The interaction emerges instead in the chained design, where per-meeting differences accumulate into a −62.5 bp terminal gap (see text).

hawkishness would reassert itself under a supply shock, producing a sharper Warsh response; it does not. On the consistent engine, then, neither the chair swap, nor the supply shock, nor their interaction moves the single-meeting decision more than a few basis points—the committee's composition and its deliberative rules, not its chair or the incoming shock, are what the counterfactuals move.

Table D.15: Dissent Summary Across Scenarios, Conditional on $K = 10$ Chains

Arm	**Chain-meetings**	**Dissents**	**Per meeting**	**Top dissenters**
`powell_hold`	60	22	0.367	Logan (12), Hammack (6), Bowman (4)
`powell_break`	60	29	0.483	Logan (9), Hammack (4), Williams (3)
`warsh_hold`	60	16	0.267	Hammack (7), Logan (6), Bowman (3)
`warsh_break`	60	21	0.350	Logan (7), Hammack (3), Paulson (2)

Notes: Each arm contains 60 simulated chain-meetings: $K = 10$ LLM chains for each of 6 forward FOMC meetings. The dissent column sums dissents across these 60 chain-meetings; "per meeting" reports the corresponding average. The dissent rule is probabilistic and role-dependent: Board governors have sensitivity 0.20 and Reserve Bank presidents have sensitivity 0.35, with a minimum preferred-policy gap of 20 basis points. Across arms, the mean is approximately 0.37 dissents per simulated meeting. This is of the same order of magnitude as the historical base rate of roughly 5–6% per vote, or about 0.6 dissents per 12-member meeting (Thornton and Wheelock, 2014).

Table D.15 reports total dissents and per-meeting dissent rates across the four scenarios. The full-deliberation vote produces about 1.9 dissents per meeting on average—above the historical base rate of roughly 0.6 per 12-member meeting (Thornton and Wheelock, 2014), reflecting the polarized 2026 committee (a dovish bloc near 3.39% against hawks near 3.89%) that leaves members on one wing or the other beyond the dissent threshold from the chair's median proposal. The break arms, with wider disagreement, run higher than the hold arms. The modal dissenters

are the doves—*sim*Kashkari and *sim*Waller—who prefer cuts the chair's proposal does not deliver; this echoes the actual March 2026 meeting, where *sim*Miran's real-world counterpart cast the lone dissent in favor of a cut. A representative dovish dissent, from the September shock meeting where the doves resist the committee's mild tightening:

Model output: *sim*Waller's dissent reasoning, September 16, 2026 (*Combined* arm)

Recommended rate: 3.39%.
"Before reviewing today's data, I expected to recommend a 25 basis point cut from the current rate of 3.64%. My stance was influenced by the previous meeting where I voted for a rate cut, reflecting concerns about ongoing economic uncertainties and elevated inflation pressures. The trajectory of recent rate decisions, which included several cuts, supports continued accommodation. . . "

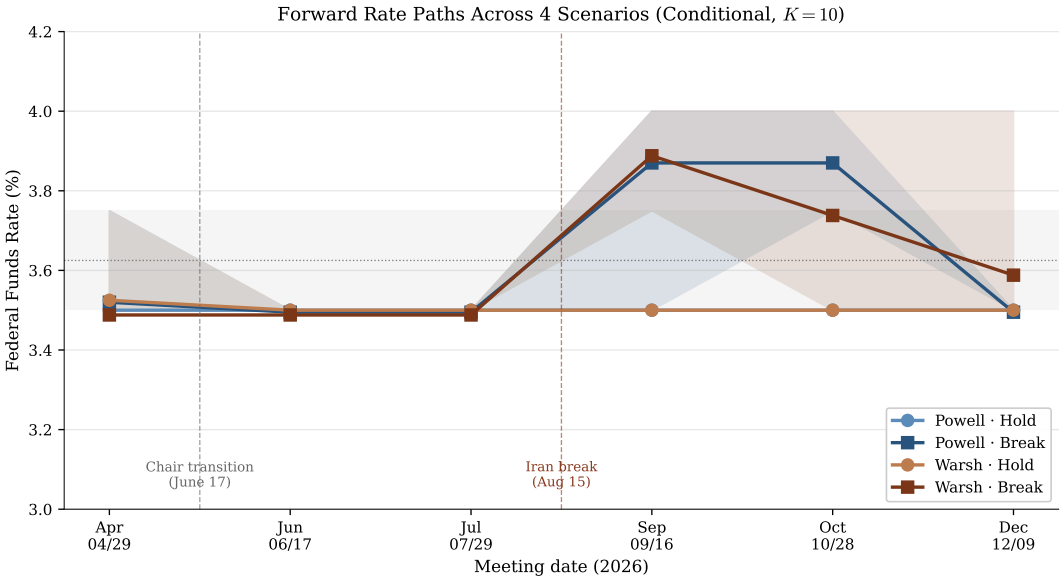


Figure D.9: Forward Rate Paths Across Four Scenarios (Conditional, K=5)

Figure D.9 provides a visual summary of the four scenarios' rate paths. The two hold-regime paths hold near the 3.6% level and lie close to one another under both chairs—the visual expression of the zero conditional Leadership effect. The two break-regime paths separate from them only slightly after the August 15 ceasefire collapse: both edge up at the September meeting (Powell to about 3.59%, Warsh to about 3.62%) before decaying back toward the hold level as the conditional design re-anchors each subsequent meeting to the calm prior state. All four paths stay within roughly 10 basis points of one another throughout—the visual counterpart of the small chair, shock, and interaction effects. The chair transition marker at June 17 and the Iran break marker at August 15 indicate the timing of the two treatment shifts.

Figure D.10 plots dissent counts by meeting and scenario. Dissent runs one to two per meeting in the calm arms and rises in the break arms, and it is predominantly

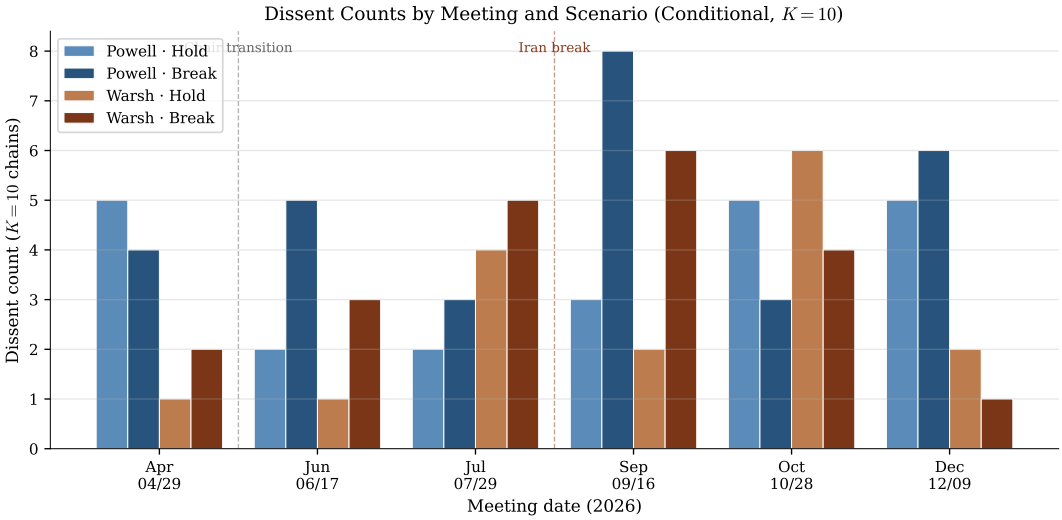


Figure D.10: Dissent Counts by Meeting and Scenario (Conditional, K=5)

dovish: the doves prefer cuts the chair's median hold does not deliver. The pattern sharpens at the September shock meeting, where the committee's mild tightening pushes the dovish bloc furthest from the settled rate—the largest single-meeting cluster in the experiment is five dovish dissents (*sim*Waller, *sim*Cook, *sim*Kashkari, *sim*Miran, and *sim*Barr) under the Combined arm. Multi-dissent meetings have historical precedent—the December 2014 FOMC produced three simultaneous dissents (Fisher, Plosser, Kocherlakota)—and the elevated counts here reflect the unusually polarized 2026 committee rather than a calibration choice.

E The Chair Channel: Alignment and Rhetoric

The aggregate results in Appendix D establish that, on the consistent chair-carries engine, swapping the FOMC chair from Powell to Warsh leaves the single-meeting decision unchanged. A natural follow-up is whether the chair swap nonetheless reshapes the committee below the level of the settled rate—realigning who moves with whom, or shifting the rhetorical stance of the collective statement—even when the rate does not move. We check two such channels: (1) a narrative approach that classifies the rhetorical stance of each synthetic FOMC statement using zero-shot LLM classification; and (2) a co-movement analysis that measures cross-member alignment under each chair. On the consistent engine both come up empty: the chair swap reshapes neither.

E.1 Narrative Channel

Following Hansen and Kazinnik (2023), we apply zero-shot classification using the GPT-4o-mini model to label each of the 120 synthetic FOMC statements (4 arms $\times$ K=5 chains $\times$ 6 forward meetings) on a 5-point hawkish/dovish scale. The classification prompt asks the model to assign one of {dovish, mostly dovish, neutral, mostly hawkish, hawkish} based on the statement's emphasis on inflation control versus employment support, with no examples or fine-tuning. Labels are then mapped to a numeric score from -2 (dovish) to $+2$ (hawkish) for aggregation.

Table E.16: Synthetic FOMC Statement Rhetoric by Scenario (zero-shot GPT-4o-mini, K=10 conditional)

Scenario	Dov.	M. dov.	Neut.	M. hawk.	Hawk.	n	Mean score
Baseline	0	24	16	20	0	60	-0.07
Chair Transition	0	21	23	16	0	60	-0.08
Shock	0	12	4	33	11	60	+0.72
Combined	0	15	6	29	10	60	+0.57

Notes: Zero-shot classification of each synthetic statement (4 arms $\times$ K=10 chains $\times$ 6 meetings) on a 5-point scale, scored -2 (dovish) to $+2$ (hawkish). No examples or fine-tuning.

Table E.16 reports the per-scenario distribution of rhetoric labels.

Table E.17: Rhetoric Shift: Baseline (Powell) vs. Chair Transition (Warsh), Ceasefire Holds

Arm	Dov.	M. dov.	Neut.	M. hawk.	Hawk.	Mean score
Baseline	0	24	16	20	0	-0.07
Chair Transition	0	21	23	16	0	-0.08
Shift (Warsh − Powell)						-0.02

Notes: Same classification as Table E.16, hold regime only (n=60 statements per arm).

Table E.17 compares the rhetoric distribution between the *Baseline* scenario and the *Chair Transition* scenario, both under the Iran-ceasefire-holds regime. The measured shift is essentially zero: the mean rhetoric score moves by -0.14 on the $[-2, +2]$ scale, within noise of the null, with nearly identical label distributions. The narrative channel thus lines up with the rate-level evidence: swapping the chair changes neither the decision nor the statement's rhetorical stance. Figure E.11 plots the full label distribution by arm.

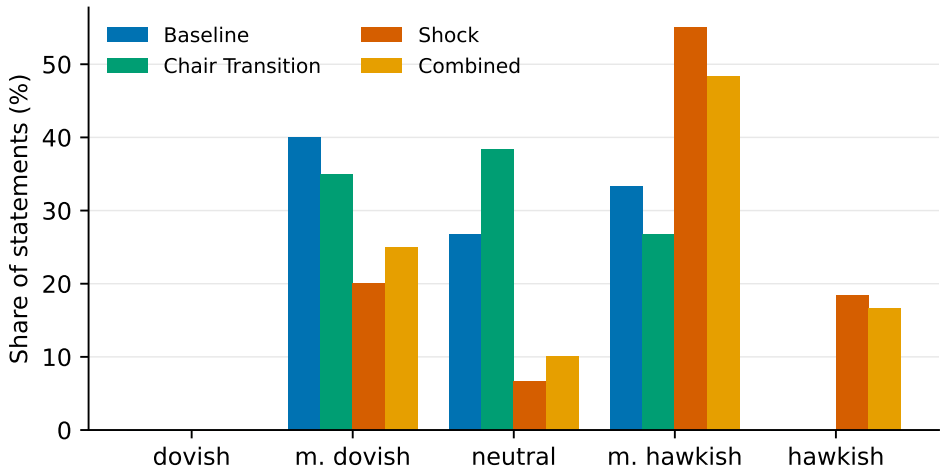
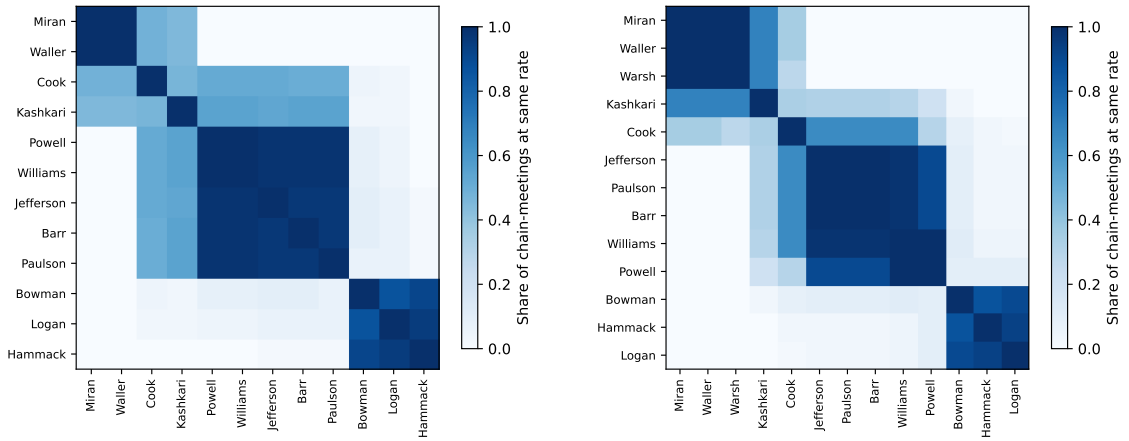


Figure E.11: Synthetic FOMC Statement Rhetoric Distribution by Arm (Conditional, K=5)

E.2 Member Co-Movement

Earlier iterations of this paper introduced an *influence map* that quantifies how member i's stance shifts toward member j's view across deliberation rounds within a single meeting. That construction requires intra-meeting transcripts in which members revise their positions in response to other members' arguments. The forward runs used in this section save only the final per-member opinion at each meeting, so the per-round influence dynamics are not observable in our data. We substitute a complementary construction: a pairwise *co-movement matrix* that measures how often each pair of members lands on the same recommended rate across the $K \times M$ chain-meeting observations.

Figure E.12 compares the co-movement matrix under the *Baseline* scenario (panel a) and the *Chair Transition* scenario (panel b). The bloc structure is clear in both: the dovish members (Cook, Miran, Kashkari, Waller) co-move tightly near 3.39%, and the hawks (Hammack, Logan, Bowman) near 3.89%. But the chair swap does not reorganize the alignment. On the consistent engine *sim*Warsh sits at the neutral median (3.64%), so he co-moves with the committee's *center*, not with either wing—his average co-movement with the dovish bloc is only 0.04. He does not enter as a fifth dove; he enters as another median voter, which is exactly why the leadership channel is zero. There is no dovish-realignment story to tell on this engine: the earlier draft's "Warsh joins the doves" pattern was an artifact of the



(a) Baseline (Powell)

(b) Chair Transition (Warsh)

Figure E.12: Member Co-Movement Matrices: Share of meetings in which each pair of members recommended the same rate (within 12.5 bp). Conditional, K=5.

lightweight forward engine's persona representation, in which the current-view fields were strong enough to force a dovish rate; under the rich personas the validation uses, his hawkish doctrine and dovish commentary net to the middle.

Taken together, the two channels agree with the rate-level evidence. On the engine used throughout the paper, swapping the chair changes the committee's decision, its collective rhetoric, and its internal alignment by essentially nothing. Whatever the chair contributes to FOMC outcomes in reality, in this laboratory—holding composition, data, and rules fixed—it is composition, not the identity of the person in the chair, that moves policy.