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

Sophia Kazinnik* Tara M. Sinclair†

March 26, 2026

**Abstract**

We develop an LLM multi-agent simulation of the Federal Open Market Committee that models policy deliberation, disagreement, and consensus formation from member specific priors grounded in real time data. The framework supports counterfactual analysis of how policy outcomes change under alternative economic conditions, political pressure, and revisions to incoming information. Across the applications we study, political pressure increases dissent and dispersion, while weaker labor market data shifts outcomes modestly in a more dovish direction. We validate the framework against the historical record: in a conditional backtest over 215 FOMC meetings from 2000 to 2026, it predicts the policy rate within 25 basis points in 93% of meetings, with a mean absolute error of 9.1 basis points. In a longer horizon trajectory exercise, it tracks the broad evolution of the federal funds rate more closely than Taylor rule and other rule based benchmarks. Comparison to a fixed voting benchmark helps identify how deliberation and institutional frictions shape policy outcomes, creating an *in silico* framework for monetary policy experimentation.

**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, Thomas R. Cook, Stephen Hansen, Federic Holm-Hadulla (discussant), Marek Jarociński, Seung J. Lee, Andrew Martinez, Daniela Puzzello, Barbara Rossi, and Eric Zivot, as well as the participants at the Stanford Digital Economy Lab seminar series, the Stanford AFTLab seminar, the 26th Federal Forecasters Conference, the Applied Machine Learning, Economics, and Data Science (AMLEDS) webinar, the Bank of England Monetary Policy Committee meeting, the University of Washington Econometrics group, and 13th ECB Conference on Forecasting Techniques for useful comments and feedback. We are grateful to the Stanford Digital Economy Lab for funding and support. 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 deliberation and the final rate decision are separated by a "shadow" of persuasion, institutional norms, and social dynamics. Federal Open Market Committee (FOMC) decisions reflect more than macroeconomic data and formal models; they also depend on members' experience, behavioral biases, and the way information is communicated within the committee (e.g., Hughes and Fehrler, 2015; Haldane, 2018; Maih et al., 2021). As a result, *who* participates in deliberation and *how* the discussion unfolds can shape policy outcomes beyond the underlying economic fundamentals.

Traditional models of committee decision-making capture important strategic features, but they often abstract from these interpersonal and behavioral frictions. At the same time, recent computational approaches, including agent-based models, reinforcement learning, and large language model (LLM) agents, offer richer representations of heterogeneity, adaptation, and communication (Dwarakanath et al., 2024; Hinterlang and Tänzer, 2021; Park et al., 2022; Zhang et al., 2024; Horton, 2023; Korinek, 2024). Yet existing approaches still do not provide a common framework in which rational benchmark voting and behaviorally rich deliberation can be studied under the same information set.

We address this gap with a dual-track framework that combines an LLM-based simulation of FOMC deliberation with a Bayesian voting benchmark implemented via Monte Carlo (MC) simulation. In the rational track, policymakers observe macroeconomic information, update beliefs, and vote within a formal strategic model. In the behavioral track, LLM agents receive the same underlying information but deliberate in natural language, allowing outcomes to reflect persona heterogeneity, persuasion, institutional roles, and social pressure. Comparing the two tracks allows us to isolate the contribution of deliberation and other behavioral frictions by identifying where LLM outcomes depart from the rule-based benchmark.

The framework works as follows. First, we automatically ingest real time macroeconomic data, recent FOMC communications, information at the district level , speeches, and financial news to reconstruct the current policy environment. Second, we build detailed personas for each committee member using biographies, historical policy leanings, and recent public communication. Third, conditional on the same environment and member-specific priors, we simulate a meeting under two alternative mechanisms: an LLM-based deliberation process and an MC benchmark that excludes conversational and institutional dynamics. Outputs include initial beliefs, posterior beliefs, vote records, dissent patterns, and full deliberation logs.

We run a historical validation of this framework. Because the framework is designed not only to generate plausible conversations but also to recover meaningful policy behavior, we evaluate it quantitatively against both benchmark models and the historical record. In a conditional backtest covering 215 FOMC meetings from 2000 to 2026, the framework predicts the policy rate with a mean absolute error of 9.1 basis points and an RMSE of 15.9 basis points, matches the realized decision exactly in 54% of meetings, and falls within 25 basis points in 93% and within 50 basis points in 99.1% of cases. It also improves materially on the Monte Carlo benchmark, reducing mean absolute error from 12.8 to 9.1 basis points and raising directional accuracy from 66.5% to 77.2%. Performance is especially strong in periods that are difficult for simple policy rules to capture: in the conditional exercise, the framework predicts the zero lower bound era perfectly, achieves 81% directional accuracy during the Financial Crisis, and 95% during the COVID period. In a longer-horizon chained validation, it also tracks the broad path of the federal funds rate more closely than Taylor-rule and rule-based RRM benchmarks, with lower RMSE, lower cumulative error, and substantially better classification of hold decisions. These results help distinguish simulations that merely sound realistic from simulations that recover the structure of actual monetary policy behavior.

To illustrate the framework, we study three environments: a baseline setting, a political pressure scenario, and a negative labor market revision shock. In the political pressure treatment, we model both direct pressure on the Chair and forward-looking career incentives for other members under an anticipated leadership change. In the

labor market treatment, we introduce a sudden negative revision to employment data, allowing us to study how real time information shocks alter discussion, belief updating, and final votes. These counterfactuals let us examine not only the level of the final policy recommendation, but also how institutional pressure and uncertainty affect disagreement, dissent, and convergence.

Our results show that deliberation matters even when all agents begin from the same macroeconomic environment. In the baseline case, the LLM committee exhibits a modest hawkish drift relative to the MC benchmark, suggesting that discussion and persuasion can move the committee even when the final target range is unchanged. Under political pressure, outcomes become more polarized and dissent becomes more frequent, with both simulation tracks indicating downward pressure on the committee's center of gravity. When a negative labor market shock is layered on top of that political environment, the simulated committee moves further in a dovish direction, though still within a constrained policy band. Taken together, these results show how communication, institutional authority, and real time uncertainty can interact to reshape policy dynamics without necessarily producing dramatic changes in the headline rate decision.

The paper contributes to the following strands of 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., Gerlach-Kristen, 2006; Riboni and Ruge-Murcia, 2010; Kamenica and Gentzkow, 2011; Holmström, 1999). 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).

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 not simply to simulate conversation, but to build a tool for disciplined counterfactual analysis of monetary policy deliberation.

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 3.2 presents the game-theoretic implementation, and Section 3.3 describes the Monte Carlo benchmark. Section 4 presents the counterfactual results, and Section 7 concludes.

# 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[1]

FOMC meetings are highly structured deliberations that anchor U.S. monetary 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.[2] 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

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

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

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 can easily extensible to include non-voting participants, other attendees, or hypothetical members. We describe the simulation pipeline next.

# 3 LLM Simulation Framework

Standard models of committee voting assume that members hold rational expectations, update beliefs via Bayes' rule, and vote to minimize well-defined loss functions. These assumptions yield sharp predictions for voting patterns, dissent, and policy outcomes. Empirically, however, FOMC behavior looks quite different. Between 1978 and 2023, fewer than 6% of all recorded ballots were dissents (Thornton and Wheelock, 2014), and from 2005 to 2022 not a single Board Governor dissented (Bobrov et al., 2025). Archival transcripts and insider accounts show that members receive extensive staff briefings and engage in private discussions in the week before each meeting, typically converging on the Chair's proposal well before the public vote (Bernanke, 2022). Dissent costs are also role-specific: Reserve Bank presidents, who face weaker career constraints, cast the vast majority of dissents, whereas D.C.-based Governors almost always vote with the Chair (Chappell and McGregor, 2013). External pressures further shape decisions: political jawboning measurably affects interest rate futures (Bianchi et al., 2023), and heightened transparency can impair information aggregation and induce conformity in committees (Hughes and Fehrler, 2015). In short, pre-meeting coordination, consensus norms, career incentives, and political or market pressures all interact to generate outcomes that depart from textbook voting models. Capturing these informal frictions is therefore central to any

realistic simulation of FOMC decision-making. Throughout the paper, we treat the LLM-based committee as a *behavioral* benchmark, and the Bayesian voting model as the *rational expectations* benchmark against which we measure these behavioral effects.

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. Because LLMs process and generate natural language, they provide a convenient way to represent the information sharing, persuasion, and strategic positioning that turn individual views into collective monetary decisions. 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. The simulation pipeline (Figure 1) unfolds in 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. We describe this pipeline in detail in the next section.[3]

## Phase 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.[4] Specifically, we pull the latest FOMC policy statement (in this case from June 18, 2025), key macroeconomic indicators, and retrieve recent speeches from each Governor. The core indicators summarized in Table 1 are obtained via the FRED and BLS APIs, providing a consistent economic backdrop for the simulation.

In addition to the raw data, the simulation generates a synthetic, up to date version of the Federal Reserve's Beige Book. To do so, we combine the most recent official

[3]The simulation employs OpenAI's GPT family of models with differentiated configurations across pipeline stages to optimize for specific tasks. We provide more details in the Online Appendix.

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

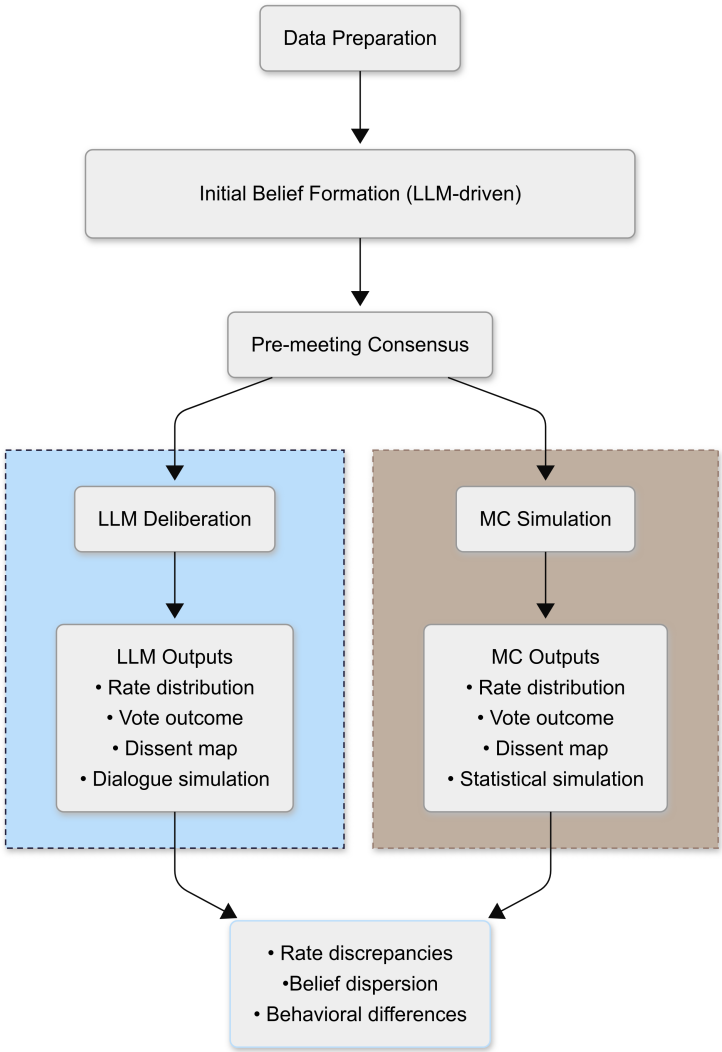



Figure 1: Simulation Framework

*Notes.* Figure 1 shows the simulation workflow. We start from real time economic inputs and agent personas, then run two parallel tracks: (i) a behaviorally rich LLM deliberation and (ii) a Monte Carlo (MC) benchmark. The design mirrors the FOMC process and lets us switch on counterfactuals to see how communication frictions and political pressure translate into policy drift.
**Phase 1 (Data & Priors):** produce member priors $(\mu_i, \sigma_i^2)$.
**Phase 2 (Consensus):** update toward the committee mean with $\alpha_i$ and compress variance by $\gamma$.
**Phase 3 (Agenda & Vote):** the Chair proposes $r_c = r_{\text{base}} + \Delta$; the final rate is $r_f = w_a r_c + (1 - w_a)\ \text{median}_i(\mu_i^{\text{post}})$; dissent propensity depends on the role-based cost $\kappa_i$.
**Political pressure (treatment):** lower $w_a$ (weaken Chair authority) and apply appointment-weighted shifts to selected members.
**Data revision (treatment):** incorporate ex-post data updates by re-initializing priors from the revised dataset: $(\mu_i, \sigma_i^2) \leftarrow (\mu_i + \delta_i,\ \rho_i\, \sigma_i^2)$

| | |
|---|---|
| $\mu_i$ | member $i$'s belief (rate) |
| $\sigma_i^2$ | uncertainty (variance) |
| $\alpha_i,\ \gamma$ | consensus pull and variance compression (Phase 2) |
| $r_c,\ \Delta$ | agenda rate and Chair bias (Phase 3) |
| $w_a$ | weight on agenda in $r_f$ (Chair authority) |
| $\kappa_i$ | role-based dissent cost |

Table 1: Key U.S. Economic Indicators – 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 % | −75.0 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 |
| U.S. Dollar Index (DXY) | 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 $\sigma$ 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.

(*actual*) Federal Reserve Beige Book content with real time web intelligence to create a comprehensive, current economic assessment. Drawing on data from government agencies, trade groups, regional media, and research centers, the system covers all twelve Federal Reserve districts. The primary method for this involves an LLM-powered `GPTResearcher` that acts like a team of specialized economic analysts, with each one focusing on one of the twelve Federal Reserve districts.[5] Each digital analyst constructs targeted queries about their assigned region. These "researchers" work in parallel, gathering real time economic data on employment trends, business activity, inflation pressures, and sector-specific developments across all fifty states. The system then synthesizes these regional reports into a national economic summary, complete with a structured district comparison table that highlights regional disparities and emerging patterns.

Table 2 compares the most recent official Beige Book (published on July 16th, 2025) with our synthetic version, produced on July 30th, 2025[6]:

Table 2: District Economic Conditions:
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 $\leq 1$ step; Shifted = swing of $\geq 2$ steps.
**Trend arrows**: ↑ improving, → stable, ↓ weakening.

[5]Please see Online Appendix for more details.
[6]Full document is available in the Online Appendix.

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.

Next, we build an executive summary of recent economic news by running multiple search queries across relevant economic domains. Each article discovered through this process is scored by an algorithm that evaluates source credibility and economic indicator relevance. The highest scoring articles are then categorized by theme and summarized. The summary highlights overall economic momentum, key risks and opportunities, and direct implications for monetary policy decisions.[7] Finally, we retrieve and summarize recent speeches made by the *actual* FOMC participants, as well as the most recent FOMC announcement.

## Phase 2: Individual Policy Formation

### Agents

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-world characteristics of its corresponding FOMC voting member. Personas are built from biographical details, recent public statements, inferred policy leanings, and district-specific economic context. Each agent receives:

a) a role description and career biography[8];

[7] Analyses draw on nearly 190 unique articles over a two-week collection window ending July 31, 2025. The highest relevance pieces are clustered around labor market trends (10 articles).

[8] Details provided in the Online Appendix.

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 synthetic Beige Book summary, recent news, and the member's own speeches and interviews.

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, $\mu_i^{\text{initial}}$. This output serves as input to the numeric vote engine, detailed in Section 3.1.

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.[9] 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

A unique aspect of the simulation is the modeling of the FOMC Chair, who carries a dual role. First, the Chair acts as a *deliberating committee member*, i.e., an LLM agent who participates in the discussion with their own dynamically updated beliefs. Second, the Chair serves as the *procedural agenda-setter*, a role governed not by the LLM but by a deterministic rule we set. This separation is a deliberate design choice to transparently model both the Chair's intellectual contribution and their formal institutional function.

As a deliberating member, the Chair agent (e.g., *sim*Powell) participates in the simulation identically to all other members. The agent forms its own private opinion, engages in dialogue, and its internal policy belief, $\mu_{\text{chair}}$, is updated throughout the meeting via the same Bayesian mechanism as its peers. In this capacity, the Chair is simply one of twelve voices contributing to the discussion.

[9] 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.

In its procedural capacity as agenda-setter, the Chair's role is to formulate the proposed interest rate, denoted as $r_c$, which serves as a critical anchor for the final vote. This process is governed by a deterministic, rule-based function to ensure transparency and reproducibility. The function systematically translates the final, updated beliefs of all committee members, including the Chair's own private belief, into a single, coherent policy proposal. The proposal generation is as follows. First, the function gathers the latest set of policy beliefs from all participating governor agents. For each of the 12 members on the committee, it extracts their current preferred interest rate, or belief mean, $\mu_i$. Next, the function aggregates this set of beliefs into a single baseline proposal, $r_{\text{base}}$, using one of two distinct strategies, which is a configurable parameter of the simulation. The agent's policy belief has two key components:

- **Mean:** Agent's current best estimate of the appropriate policy rate (e.g., 4.5%). This is their "belief."
- **Variance:** Agent's uncertainty about their belief. A high variance implies low confidence and greater openness to persuasion, and vice versa.

The default approach calculates the median of the members' beliefs.[10] This method is robust as it is not excessively influenced by outlier policy views:

$$r_{\text{base}} = \text{median}\Big(\{\mu_i\}_{i=1}^{N}\Big)$$

The model also allows for the application of an optional bias, denoted by $\Delta$. This parameter is designed to model a chair who may have an intrinsic hawkish (positive $\Delta$) or dovish (negative $\Delta$) inclination. This bias is applied as a simple additive shift to the baseline proposal. The final proposed rate, $r_c$, is the result of combining the baseline proposal with the optional chair bias:

$$r_c = r_{\text{base}} + \Delta$$

These two roles are reconciled during the final vote. The simulation models the

[10] Median is the default in our implementation. Other options include mean and status quo.

Chair's institutional responsibility through extremely high dissent costs rather than an explicit override. The voting logic assigns the Chair agent a dissent cost approximately four times higher than other members, making it extremely unlikely (though not impossible) for the Chair to dissent from their own proposal $r_c$, even if it diverges from their private, deliberated belief $\mu_{\text{chair}}$. This probabilistic mechanism models the Chair's responsibility to lead the committee to a consensus while preserving the possibility of rare institutional dissent.

The framework's flexibility allows for exploring alternative scenarios by adjusting both the aggregation strategy and the bias term. These are discussed in more detail in Section 4.

## 3.1 Phase 3: Committee Deliberation and Voting

The deliberation stage replicates the procedural and behavioral dynamics of real FOMC meetings. Rather than modeling policy decisions in isolation, it simulates the full decision-making pipeline: information flow, informal pre-meeting coordination, structured debate moderated by the Chair, and institutional pressures that influence final votes. By capturing both informational inputs and member incentives, the system reflects how private priors, public reasoning, and consensus building interact to shape collective decisions.

Deliberation begins with the integration of all economic intelligence gathered earlier, including macro indicators, regional assessments, recent speeches, and the latest FOMC statement, along with each member's initial policy stance. Before formal debate, the system includes an optional consensus building phase, simulating the informal coordination typical of FOMC culture (Bernanke, 2022).[11]

[11]To model this, each agent's initial belief $\mu_{i,\text{old}}$ is iteratively nudged toward a consensus target $\mu_{\text{consensus}}$, defined as a weighted average of the committee's median and mean beliefs:

$$\mu_{i,\text{new}} = (1 - \omega_i)\mu_{i,\text{old}} + \omega_i \mu_{\text{consensus}},$$

where $\omega_i = 0.15 \times (\sigma_i^2/0.1)$ is a variance-dependent convergence rate that makes more uncertain members more susceptible to consensus pressure. Shared information in this stage also increases confidence by reducing each agent's variance through a confidence boost multiplier $\gamma < 1$:

$$\sigma^2_{i,\text{new}} = \gamma \cdot \sigma^2_{i,\text{old}}.$$

This ensures agents enter the formal meeting with more aligned beliefs and slightly higher conviction

Once consensus building concludes, the system establishes the meeting context by presenting three key pieces of information. It begins by reviewing the results and decisions from the previous meeting to provide continuity in policy reasoning. It then extracts key language and forward guidance from the most recent FOMC statement, preserving institutional memory and highlighting any prior commitments. Finally, it presents an updated snapshot of current economic conditions, detailing changes in inflation, employment, growth, and regional dynamics since the last meeting.

The formal meeting simulation begins with the Chair opening the session, delivering an economic briefing, highlighting major developments, and posing the central policy question, i.e., whether to maintain the current federal funds rate or adjust it. Deliberation proceeds in structured rounds that mirror real FOMC debates. In the first round, selected members present policy recommendations with brief justifications based on their economic philosophy, regional context, and interpretation of current conditions. The second round shifts to a discussion moderated by the Chair, focused on the employment–inflation trade off, with dovish members emphasizing labor markets and hawkish members prioritizing price stability, both citing data. A third round enables direct engagement, as members challenge one another's arguments and refine their views. The Chair then concludes by summarizing key points, noting areas of consensus and disagreement, and proposing an agenda rate based on the discussion.

The decision phase then follows. The agenda rate is used as an anchor, and the final rate is calculated as a weighted average of the agenda rate (30%) and the committee median (70%).[12] Institutional pressure and dissent costs are applied, reflecting the dynamics of consensus seeking behavior within the FOMC. Members' institutional roles affect these dynamics: the Chair has strong agenda-setting power and very high dissent costs, while Governors have more independence. The dissent decision is modeled as a probabilistic rule tied to distance from the adopted rate,

than their private priors.

[12]Formally, the Fed Chair has one vote like any other FOMC voter. However, the Chair's influence is far greater than a 1/12 vote share due to control of procedure and communications. Prior empirical studies have found that during the Arthur Burns era (1970–78), for example, the FOMC Chair effectively carried 40–50% of the voting weight in policy decisions (Chappell Jr et al., 2004).

role-specific dissent costs, and each member's uncertainty. For member $i$, the dissent propensity is

$$P_i^{\text{dissent}} = \left|\mu_i - r^{\text{final}}\right| \times \frac{1}{\kappa_i} \times \left(1 + \text{Var}_i\right),$$

where $\mu_i$ is the member's preferred rate, $r^{\text{final}}$ is the adopted rate, $\text{Var}_i$ is the member's belief variance, and $\kappa_i$ is a role-based dissent cost. In baseline settings,

$$\kappa_{\text{Chair}} = 10.0, \qquad \kappa_{\text{Regional}} = 2.5, \qquad \kappa_{\text{Board}} = 2.5,$$

with the Chair's cost reduced to 5.0 under political pressure "reduced authority." A dissent is realized if:

$$P_i^{\text{dissent}} > 0.5^{[13]} \quad \text{and} \quad \left|\mu_i - r^{\text{final}}\right| > 0.25 \text{ pp}.$$

Political pressure modifies $\kappa_i$ asymmetrically using appointment-likelihood weights: dovish dissent becomes cheaper and hawkish dissent more costly for high-likelihood candidates. The activation thresholds are relaxed for non-chair members with high appointment likelihood (effective cutoff lowered to 0.3 and minimum distance to 0.15 pp).

## 3.2 Implementation Framework

We implement the Riboni and Ruge-Murcia (2010) framework with three key deviations for a more appropriate comparison against the LLM-based environment: (i) private signals are independent across members to model true heterogeneity[14]; (ii) the Chair proposes the agenda via a fixed *rule* (median, mean, or status-quo) to isolate

[13] We calibrate the dissent indicator threshold $\tau$ in the rule $p_{\text{dissent}} = \Delta\,(1/\kappa)\,(1+\sigma)$—with $\Delta$ the absolute gap between a member's preferred rate and the final rate, $\kappa$ the institutional dissent cost, and $\sigma$ residual uncertainty, so that realized dissent frequencies align with the FOMC's historical range (about 5–15%), with $\tau = 0.5$ landing around 8–10% in our baseline runs.

[14] Riboni and Ruge-Murcia (2010) allow for correlated signals, whereas we assume independence. The rationale is pragmatic since correlation parameters are not separately identified in historical voting data, and heterogeneity at the FOMC is often driven by district specific, idiosyncratic information. Empirically, Chappell and McGregor (2013) show that cross-member forecast errors display weak correlation once common macro factors are removed. Independent signals therefore approximate the "residual" heterogeneity relevant for persuasion.

committee dynamics, rather than by optimizing their own loss[15]; (iii) if a proposal fails, the committee adopts the *median of the ballots* as a realistic compromise, not the previous-meeting rate.[16] These modifications retain the logic of Bayesian voting yet add empirically grounded institutional detail.

Each committee member $i$ begins with a normal prior belief about the appropriate policy rate, $r \sim \mathcal{N}(\mu_i^0, \sigma_i^{0\,2})$. The agent then receives an independent private signal, $s_i$, drawn from a distribution centered around their prior:

$$s_i \;=\; \mu_i^0 + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

where $\sigma^2$ is the globally set signal noise. After receiving this signal, the agent forms a posterior belief $\mathcal{N}(\mu_i^{\text{post}}, \sigma_i^{\text{post}\,2})$ via the conjugate normal–normal updating rule. This posterior serves as the agent's initial belief entering the meeting phase.

The LLM simulation employs a consensus convergence mechanism that models realistic committee dynamics. During pre-meeting consensus building, beliefs evolve according to:

$$\mu_{i,t+1} = \mu_{i,t} + \alpha_i \cdot (\bar{\mu}_t - \mu_{i,t}), \qquad \sigma^2_{i,t+1} = 0.95 \cdot \sigma^2_{i,t},$$

where $\bar{\mu}_t$ is the committee average belief, $\alpha_i = 0.15 \cdot (\sigma^2_{i,t}/0.1)$ captures convergence strength based on uncertainty, and variance reduces by 5% per round to model confidence building through information sharing. This mechanism captures the realistic tendency for committee members to converge toward consensus through informal coordination, rather than performing precise mathematical calculations.

The Chair applies a deterministic rule to the final set of posterior beliefs, $\{\mu_i^{\text{post}}\}$,

[15] Riboni and Ruge-Murcia (2010) model the Chair as minimizing her own loss. In practice the Fed Chair emphasizes consensus and signaling credibility more than personal utility. Transcripts reveal that Chairs routinely propose "somewhere around the middle" of expressed views (see Bernanke, 2022). We therefore implement an institutional rule, $r_c = \text{median}\{\mu_i\}$ by default, and treat alternative rules (mean or status quo) as robustness cases. This choice converts the benchmark from a chair-dominance regime to a median voter regime, which is the modal empirical pattern documented by Thornton and Wheelock (2014).

[16] Historical minutes show that when the committee cannot agree on the Chair's proposal it typically negotiates a mid-point compromise rather than mechanically retaining the prior setting (e.g. the August 1989 and June 2003 meetings). This median fallback matches that practice and avoids artificially inflating policy inertia.

to generate the proposed agenda, $r_c$:

$$r_c = \begin{cases} \text{median}_i\, \mu_i^{\text{post}} & \text{baseline strategy,} \\ \text{mean}_i\, \mu_i^{\text{post}} & \text{alternative strategy,} \\ r_{\text{previous}} & \text{status-quo strategy.} \end{cases}$$

The voting mechanism incorporates institutional pressures and role-based dissent costs. Member $i$ faces a dissent cost $C_i$ that varies by institutional role, mirroring the LLM set up. The final rate reflects both agenda influence and committee preference through a weighted combination: $r_f = w_a \cdot r_c + (1 - w_a) \cdot \text{median}_i\, \mu_i^{\text{post}}$, where $w_a = 0.3$ under normal conditions and $w_a = 0.1$ under political pressure, modeling reduced Chair authority. If the proposal fails in either implementation, the committee implements $r_f = \text{median}_i\, b_i$, the median of all submitted ballots.

## 3.3 Monte Carlo Simulation

The Monte Carlo (MC) analysis implements the mathematical core of the Riboni and Ruge-Murcia (2010) framework with enhanced methodological features to enable rigorous statistical comparison with LLM simulation results. This approach abstracts away from natural language deliberation to focus on formal game theory mechanisms, allowing us to systematically vary institutional settings while maintaining strict comparability between simulation approaches.

Each MC draw represents a single realization of the voting game. The simulation begins with initial policy beliefs sourced from the LLM opinion formation stage. In both tracks, the initial cross-section of beliefs is obtained from the LLM-based opinion formation stage, which translates real time macro data and member personas into heterogeneous prior means and variances. The Monte Carlo implementation should therefore be interpreted as a rational Bayesian voting model *conditional* on this empirically generated prior, not as a fully structural model with independently specified priors. Any misspecification or artifacts in the LLM priors are common to both tracks; our "deliberation wedge" is identified as the difference in outcomes holding this prior fixed while switching from a rule-based voting environment to a

conversational one.

For each committee member $i$, their belief is represented as a normal distribution:

$$\text{Belief}_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$$

where $\mu_i$ is the agent's preferred policy rate from LLM opinion formation and $\sigma_i^2$ represents their baseline uncertainty derived from confidence levels. Belief means are sampled from agents' own reported confidence distributions:

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

where the confidence-based variance is derived from self-reported confidence levels:

$$\sigma_i^* = \kappa \cdot \frac{1 - c_i}{c_i}$$

Here, $c_i \in [0, 1]$ represents agent $i$'s confidence level and $\kappa = 25$ basis points is a calibration parameter ensuring that agents reporting 50% confidence have a standard deviation of $\kappa$.

Following belief initialization, each agent receives an independent private signal $s_i$ and updates their belief using conjugate normal-normal Bayesian updating:

$$\tau_i \equiv \sigma_i^{-2}, \qquad \tau_{\text{signal}} \equiv \sigma_{\text{signal}}^{-2}, \tag{1}$$

$$\mu_i^{\text{post}} = \frac{\tau_i \, \mu_i' + \tau_{\text{signal}} \, s_i}{\tau_i + \tau_{\text{signal}}}, \tag{2}$$

$$\sigma_i^{2,\text{post}} = \frac{1}{\tau_i + \tau_{\text{signal}}} \tag{3}$$

$$\text{Belief}_i^{\text{updated}} \sim \mathcal{N}\left(\mu_i^{\text{updated}}, \sigma_i^{2,\text{updated}}\right) \tag{4}$$

where $\sigma_{\text{signal}}^2$ represents the precision of private information. The simulation also includes a pre-meeting consensus-building phase designed to replicate institutional coordination mirroring the LLM simulation.

The Chair applies a deterministic rule to the post-consensus belief set $\{\mu_i^{\text{post}}\}$ to

generate the agenda proposal:

$$r_{\text{base}} = \begin{cases} \text{median}(\{\mu_i^{\text{post}}\}) & \text{(baseline strategy)} \\ \frac{1}{N}\sum_{i=1}^{N} \mu_i^{\text{post}} & \text{(mean strategy)} \\ r_{\text{previous}} & \text{(status-quo strategy)} \end{cases}$$

The final agenda also incorporates Chair bias, if present: $r_c = r_{\text{base}} + \Delta$, where $\Delta$ represents hawkish ($\Delta > 0$) or dovish ($\Delta < 0$) tendencies.

Institutional voting follows individual dissent cost calculations. Member $i$ votes YES on proposal $r_c$ if:

$$\left|\mu_i^{\text{post}} - r_c\right| \leq \epsilon \cdot \kappa_i$$

where $\epsilon = 0.25$ percentage points represents the tolerance band and $\kappa_i$ captures member-specific dissent costs varying by institutional role and political conditions, calibrated to match LLM simulation patterns.

The proposal passes if the weighted vote share exceeds threshold $\vartheta = 0.5$:

$$\frac{\sum_i w_i \mathbf{1}_{\text{YES}}}{\sum_i w_i} \geq \vartheta$$

where standard voting weights are $w_i = 1.0$ for all members. Failed proposals trigger implementation of the median submitted ballot: $r_f = \text{median}(\{b_i\})$.

Each Monte Carlo draw records key variables for policy analysis: the Chair's proposed agenda rate ($r_c$), the final implemented policy rate ($r_f$), and the set of dissenting members ($D$). By running 10,000 draws, the Monte Carlo simulation provides statistical grounding for LLM results.

# 4 Results

With these two simulation methods at hand, we now arrive at two complementary setups: a "control" and a "treatment." In the control case, the Monte Carlo simulations depict a world of purely rational, rule-based agents. They receive initial beliefs, undergo a parameter change, and vote strictly according to predefined rules. This

serves as the benchmark; what pure economic theory would predict under idealized conditions. The treatment, in contrast, introduces the LLM simulation, which reflects the messy and human-like dynamics of deliberation. Here, agents don't simply vote mechanically; they debate, argue, persuade, and react to one another. This layer of social interaction captures the complexity and unpredictability of real-world decision making.

In this section, we present the results of our primary counterfactual simulation. Across all scenarios, the underlying economic environment, such as inflation, unemployment, and other key indicators, is held fixed at U.S. conditions in late July of 2025. This ensures that any differences in policy outcomes arise solely from changes in committee composition or decision-making rules, rather than from varying macroeconomic shocks.

However, to meaningfully evaluate the effects of political pressure, we must first understand the system's behavior in the absence of such distortions. We therefore begin with a baseline scenario, establishing a reference point under normal conditions.

## 4.1 Initial Beliefs

Table 3 summarizes the multi-iteration analysis ($\sim$ 100 runs) of rate preferences, modeling high dimensional and complex belief formation. Confidence here is the agent's self-assessed certainty (percent) about its initial rate view. We transform this into the member's belief variance; higher confidence yields tighter distributions, which affects deliberation dynamics and posterior sampling in the MC mirror.

The results reveals distinct and persistent themes in the reasoning patterns of each FOMC member. On the dovish end of the spectrum, *sim*Kugler, *sim*Cook, and *sim*Goolsbee consistently emphasize the importance of supporting the labor market and sustaining the economic expansion in the face of gradually easing inflation. These members view monetary policy as a tool to ensure continued employment gains, and frequently cite the deceleration of inflationary pressures as justification for maintaining or even lowering interest rates. Their focus leans heavily toward protecting growth and minimizing the risks of premature tightening.

Simulated Powell, Williams, and Jefferson adopt a more balanced and flexible

Table 3: Initial Rate Beliefs – July 31st, 2025 Vintage

| Member | Avg. Rate | Confidence | Key Reasoning |
|---|---|---|---|
| *sim*Powell | 4.44% (±0.080) | 84.8% | Emphasizes balance of risks; data-driven neutrality amid persistent inflation and solid growth |
| *sim*Cook | 4.24% (±0.025) | 82.0% | Inflation shows signs of easing; consistent support for dovish policy to aid labor inclusivity |
| *sim*Goolsbee | 4.24% (±0.036) | 82.2% | Supports growth while acknowledging inflation; balances expansion with caution |
| *sim*Kugler | 4.25% (±0.019) | 81.8% | Prioritizes employment stability; consistent preference for accommodative policy amid softening inflation |
| *sim*Williams | 4.43% (±0.086) | 85.0% | Neutral stance grounded in strong fundamentals and declining inflation risks; prefers steady path |
| *sim*Jefferson | 4.35% (±0.060) | 84.8% | Dual mandate balance; favors slightly below-neutral rate to safeguard employment and prices |
| *sim*Barr | 4.41% (±0.088) | 84.8% | Persistent inflation cited as key concern; moderately hawkish stance to maintain progress |
| *sim*Waller | 4.18% (±0.088) | 77.0% | Concerned about sticky core inflation, but leans dovish with focus on maintaining momentum |
| *sim*Bowman | 4.22% (±0.208) | 83.0% | Highlights volatility and uncertainty in inflation data; cautious with occasional dovish signals |
| *sim*Schmid | 4.58% (±0.117) | 84.9% | Strongly hawkish; inflation pressures and robust growth warrant upper-bound positioning |
| *sim*Musalem | 4.62% (±0.123) | 84.6% | Focused on anchoring inflation expectations; prefers tighter policy to guard credibility |
| *sim*Collins | 4.51% (±0.036) | 85.0% | Sustained demand and resilient inflation support firm, hawkish positioning |

**Committee Initial Rate Beliefs (100 Iterations)**

| | |
|---|---|
| **Avg. Initial Recommendation:** | **4.375% (±0.026)** |
| **Range (Across Iterations):** | 4.33% – 4.44% |
| **Typical Spread:** | 0.552 percentage points (±0.150) |
| **Spread Range:** | 0.25pp – 0.75pp |
| **Members with opinions:** | 12 |
| **Hawkish (≥4.50%):** | 3 members (Schmid, Musalem, Collins) |
| **Moderate (4.41–4.44%):** | 3 members (Powell, Williams, Barr) |
| **Leaning Dovish (4.24–4.35%):** | 4 members (Goolsbee, Kugler, Jefferson, Cook) |
| **Dovish (<4.24%):** | 2 members (Waller, Bowman) |
| **Most Consistent:** | Adriana D. Kugler ($\sigma$=0.019) |
| **Least Consistent:** | Michelle W. Bowman ($\sigma$=0.208) |

approach. These members consistently express a desire to remain data-dependent, neither aggressively tightening nor easing policy without clear justification. Their reasoning reflects a concern for maintaining economic stability, ensuring continued labor market strength, and watching closely for any resurgence of inflation. These policymakers tend to advocate for holding rates steady unless compelling evidence warrants a shift, favoring optionality and responsiveness to evolving macroeconomic conditions.

Meanwhile, on the hawkish end of the committee, *sim*Schmid, *sim*Musalem, and *sim*Collins personas are more resolute in their call for tighter monetary policy. Their reasoning centers on the persistence of inflation and the need to anchor expectations through strong, pre-emptive rate stances. These members frequently point to robust consumer demand, strong labor market data, and signs of underlying inflation resilience as justification for maintaining elevated policy rates. They view any premature easing as a threat to the Fed's credibility and the longer term goal of price stability.

The most ideologically consistent voices emerge at the extremes. For example, *sim*Kugler persona, on the dovish side, and *sim*Schmid persona, on the hawkish side, exhibit the most stable reasoning across the 100 iterations, with minimal variation in their underlying justifications. Their positions are deeply anchored in their respective macroeconomic priorities: employment for *sim*Kugler, and inflation control *sim*Schmid. In contrast, *sim*Bowman shows a more variable perspective; *sim*Bowman is alert to inflation but, with mixed signals, her stance changes more than her peers'.

## 4.2 Baseline Scenario

The simulation begins with launching an LLM-driven meeting of the FOMC. The twelve autonomous agents representing the committee members are not blank slates; they enter with structured priors derived from current macroeconomic data. These agents are also governed by a set of institutional rules that mirror the culture of the real Federal Reserve. The synthetic FOMC members are deliberating against a backdrop of solid economic growth (GDP growth 3.0%), inflation above target,

and a stable labor market (unemployment 4.1%, modest payroll growth). Table .15 summarizes the results:

Table 4: Baseline Scenario:
Member Deliberation and Voting – July 31st, 2025

| Simulated Member | Deliberation Rate | Variation | Uncertainty | Vote Bias | | Dissent |
|---|---|---|---|---|---|---|
| | Mean (%) | Std | Mean | Mean | Std | Freq (%) |
| *sim*Goolsbee | 4.250 | 0.036 | 0.822 | -0.1704 | 0.0360 | 0.0 |
| *sim*Cook | 4.250 | 0.025 | 0.820 | -0.1704 | 0.0250 | 0.0 |
| *sim*Kugler | 4.250 | 0.019 | 0.818 | -0.1704 | 0.0190 | 0.0 |
| *sim*Barr | 4.405 | 0.088 | 0.848 | -0.0159 | 0.0880 | 0.0 |
| *sim*Jefferson | 4.349 | 0.060 | 0.848 | -0.0710 | 0.0600 | 0.0 |
| *sim*Williams | 4.424 | 0.086 | 0.850 | 0.0033 | 0.0860 | 0.0 |
| *sim*Powell | 4.439 | 0.080 | 0.848 | 0.0184 | 0.0800 | 0.0 |
| *sim*Waller | 4.202 | 0.088 | 0.770 | -0.2181 | 0.0880 | 0.0 |
| *sim*Bowman | 4.465 | 0.208 | 0.830 | 0.0449 | 0.2080 | 0.0 |
| *sim*Schmid | 4.582 | 0.117 | 0.849 | 0.1620 | 0.1170 | 0.0 |
| *sim*Musalem | 4.622 | 0.123 | 0.846 | 0.2017 | 0.1230 | 0.0 |
| *sim*Collins | 4.514 | 0.036 | 0.850 | 0.0934 | 0.0360 | 0.0 |
| **Baseline FFR** | **4.4204** | 0.0229 | | | | |
| **Target Range** | | | 4.25% – 4.50% | | | |
| **Effective FFR** | | | 4.33% | | | |

*Notes:* In practice, the FOMC sets the target range for the federal funds rate in 25 basis point increments. The simulated *Final FFR* of 4.42% falls within the 4.25%–4.50% range, corresponding closer to the midpoint of the range (4.375%) than to the current effective FFR (4.33%) . *Deliberation Rate:* The interest rate each member publicly argues for during the formal meeting. *Variation (Std):* Standard deviation of the deliberation rate across 100 simulations. *Uncertainty:* Member's self-reported confidence/uncertainty (decimal form). *Vote Bias:* Difference between the deliberation rate and the simulated final FFR; positive = more hawkish, negative = more dovish.

The baseline distribution shows a modest hawkish tilt. *sim*Bowman is near the 4.50% upper bound (4.46%), while *sim*Collins (4.51%), *sim*Schmid (4.58%), and *sim*Musalem (4.62%) sit above it; *sim*Powell is just below at 4.44%. On the dovish side, *sim*Goolsbee, *sim*Cook, and *sim*Kugler are at 4.25%, with *sim*Jefferson (4.35%) and *sim*Barr (4.40%) slightly higher. The mean recommended rate is 4.40% (vs. a simulated final FFR of 4.42%), and views span roughly 42 basis points. Notably, *sim*Waller is dovish in this simulation (4.20%).

In the simulated aggregation, the committee's *Baseline FFR* comes to 4.42%, which lies well within the prevailing 4.25%–4.50% target range. Given that the effective federal funds rate at the prior meeting was about 4.33%, this outcome implies an upward drift of roughly 9 basis points, effectively tightening policy within

the existing range rather than formally raising it by a full 25 basis points. Figure 2 provides an overview of member-level vote distribution.

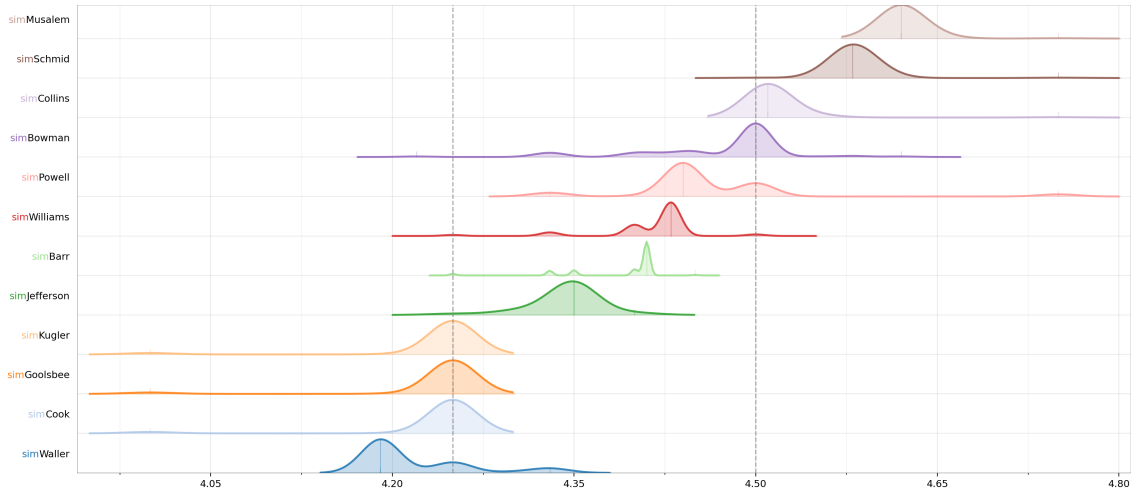



Figure 2: Baseline Scenario: Committee Vote Distribution

*Notes:* Each row plots a kernel density estimate of that member's votes (densities normalized within row). Dashed references at 4.25% and 4.50%.

Despite the heterogeneity in beliefs, no formal dissents emerge. The Chair synthesizes prevailing inflation concerns with growth risks, presenting the upward shift as a balanced response that all members ultimately accept. This outcome illustrates how the simulation captures both diversity in individual preferences and the institutional dynamics that narrow disagreement and support consensus in FOMC deliberations.

Next, we summarize outcomes from the baseline Monte Carlo simulation. In the baseline scenario for July 31, 2025, across 10,000 simulated meetings, the final policy rate clusters around the prior 4.25–4.50% target range, with a mean of 4.376%, a standard deviation of 7.4 basis points, and outcomes spanning 4.19% to 4.68%. The distribution is centered slightly above the midpoint (4.33%), implying an average upward drift of about 4–5 basis points in the effective federal funds rate. No formal dissents occur in any of the 10,000 runs, consistent with broad consensus.

Overall, both the LLM deliberation model and the Monte Carlo benchmark produce tightly clustered, consensus outcomes near the policy midpoint, but via different mechanisms. The LLM run offers a narrative of hawks and doves weighing conditions and converging on compromise. The Monte Carlo benchmark, relying

only on institutional voting rules and members' initial proposals, reduces dispersion and yields no dissents across 10,000 runs. This alignment is consistent with (and supports the plausibility of) the LLM-derived outcomes.

## 4.3 Political Pressure Scenario

> *"Leaving rates steady makes Powell a 'Total and Complete Moron.' Maybe, just maybe, I'll have to change my mind about firing him?"* – President Trump via Truth Social, June 20, 2025
>
> *"I've been nice, I've been neutral, and I've been nasty, and nice and neutral didn't work...Powell is a numbskull."* –President Trump via Truth Social, June 21, 2025
>
> *"I'd love it if he wants to resign.... The housing market would boom if this— you know—numbskull lowered the rates."* – President Trump in an interview, July 16, 2025.
>
> *"I think he's done a bad job, but he'll be out pretty soon anyway...People can't buy a house because this guy is a numbskull. He keeps rates too high, probably for political reasons."* – President Trump in a White House economic conference, July 23, 2025
>
> *"We're spending* ***hundreds of billions*** *because of one numbskull sitting over there - Jerome Powell."* – President Trump during pool spray with reporters, July 24, 2025

Recent White House remarks have reignited talk that President Trump might try to remove Chair Powell before his term ends.[17] Markets reacted immediately: fed funds futures now price in about 125 basis points of rate cuts by end of 2026, roughly 50 basis points more easing than the Fed's own dot plot suggests. Investors appear to be bracing for a more dovish successor[18], which in turn revives concerns about the Federal Reserve's ability to resist direct political pressure.

Political interference in central banking, while formally prohibited, is hardly rare. Using a new index of political pressure, Binder (2021) shows that even nominally independent central banks are often urged to loosen policy, with measurable effects on inflation expectations and actual decisions. History echoes this pattern: President

[17] See, e.g., "Trump says he will name new Fed chair 'a little bit earlier'," Reuters, Aug. 13, 2025; and "Trump weighs firing Powell's Fed, shows draft letter," Politico, July 16, 2025.

[18] Reuters, "Market bets on a more dovish Fed as Trump eyes Powell's replacement," June 27, 2025.

Nixon leaned on Chair Arthur Burns in the early 1970s to keep rates low, as documented by Meltzer (2009). Recent archival work by (Drechsel, 2025) shows that bursts of presidential pressure of similar magnitude raise the U.S. price level by roughly 7 percent over a decade without any lasting boost to real activity.

These episodes (past and present) inspire the political pressure channel in our model: diminished chair authority heightens the career incentives of other FOMC members to accommodate political demands, potentially skewing policy toward easier monetary conditions.[19]

Under political pressure, we shift authority away from the Chair and build in career-driven dovishness. Concretely, we cut the Chair's agenda weight from 0.3 to 0.1, so the policy rate becomes a weighted average of the Chair's agenda rate and the committee median: the lower weight means outcomes lean more toward the median. We also introduce a dovish adjustment: an appointment-linked bias equal to 50 basis points, scaled by appointment probability (only applied to members with high (0.6) appointment probability). Finally, pressure pulls the agenda itself below the median, setting the agenda rate 10 basis points under the committee median, so even before weighting, the Chair's target is lower. Together, these changes tilt decisions dovish overall and generate split behavior across members depending on their appointment odds.

Table 5 summarizes the results. Under the political pressure scenario, the Committee's internal balance of power shifts visibly. Public criticism from the White House erodes Chair's agenda setting weight and magnifies career incentives for those viewed as plausible successors. The political pressure scenario generates a more polarized distribution of preferences, with sharper dovish positioning among some members and persistent hawkish stances among others. *sim*Waller emerges as the most dovish, holding rates near 3.97% in every simulation and dissenting in 88% of runs. A small dovish bloc (including *sim*Waller, *sim*Bowman, and at times *sim*Barr) consistently pushes for levels below the lower bound of the prevailing 4.25%–4.50% range. On the hawkish side, *sim*Musalem (4.617%), *sim*Schmid (4.577%), and

[19]Academic and historical evidence supports the realism of a career related dovish bias, though it's usually discussed qualitatively rather than as a fixed formula.

Table 5: Political Pressure Scenario:
Member Deliberation and Voting

| Simulated Member | Deliberation Rate | Variation | Uncertainty | Vote Bias | | Dissent |
|---|---|---|---|---|---|---|
| | Mean (%) | Std | Mean | Mean | Std | Freq (%) |
| *sim*Waller | 3.969 | 0.053 | 0.700 | -0.4063 | 0.0000 | 88.0 |
| *sim*Bowman | 4.463 | 0.208 | 0.830 | 0.0873 | 0.2080 | 0.0 |
| *sim*Barr | 4.373 | 0.088 | 0.848 | -0.0026 | 0.0880 | 0.0 |
| *sim*Williams | 4.389 | 0.086 | 0.850 | 0.0138 | 0.0860 | 0.0 |
| *sim*Cook | 4.210 | 0.025 | 0.820 | -0.1658 | 0.0250 | 0.0 |
| *sim*Goolsbee | 4.209 | 0.036 | 0.822 | -0.1663 | 0.0360 | 0.0 |
| *sim*Kugler | 4.210 | 0.019 | 0.818 | -0.1658 | 0.0190 | 0.0 |
| *sim*Jefferson | 4.328 | 0.060 | 0.848 | -0.0482 | 0.0600 | 0.0 |
| *sim*Powell | 4.403 | 0.080 | 0.848 | 0.0275 | 0.0800 | 0.0 |
| *sim*Schmid | 4.577 | 0.117 | 0.849 | 0.2009 | 0.1170 | 0.0 |
| *sim*Musalem | 4.617 | 0.123 | 0.846 | 0.2409 | 0.1230 | 0.0 |
| *sim*Collins | 4.511 | 0.036 | 0.850 | 0.1348 | 0.0360 | 0.0 |
| **Final FFR** | **4.376** | 0.058 | | | | |
| **Target Range** | | | 4.25% – 4.50% | | | |
| **Effective FFR** | | | 4.33% | | | |

*Notes:* In practice, the FOMC sets the target range for the federal funds rate in 25 basis point increments. The simulated *Final FFR* of 4.3758% lies at the exact midpoint of the prevailing 4.25%–4.50% range, implying no change in the target range but maintaining the effective rate near its prior level (4.33%). *Deliberation Rate:* The interest rate each member publicly argues for during the formal meeting. *Variation:* Standard deviation of the deliberation rate across 100 simulations. *Uncertainty:* Member's self-reported confidence/uncertainty. *Vote Bias:* Difference between the deliberation rate and the simulated final FFR; positive = more hawkish, negative = more dovish. Mean dissents per simulation: 0.88; at least one dissent occurs in approximately 88% of runs (driven almost entirely by Waller).

*sim*Collins (4.511%) advocate above 4.50%.

The simulated *Final FFR* averages 4.376%, which sits at the midpoint of the existing target range. In practical terms, this would leave the range unchanged but position the effective federal funds rate near its prior level, a posture consistent with maintaining the stance of policy rather than easing or tightening. This stands in contrast to the baseline scenario, where the simulated mean implied an upward drift of roughly 10 basis points within the band. Figure 3 provides a visualization of the member vote distributions.

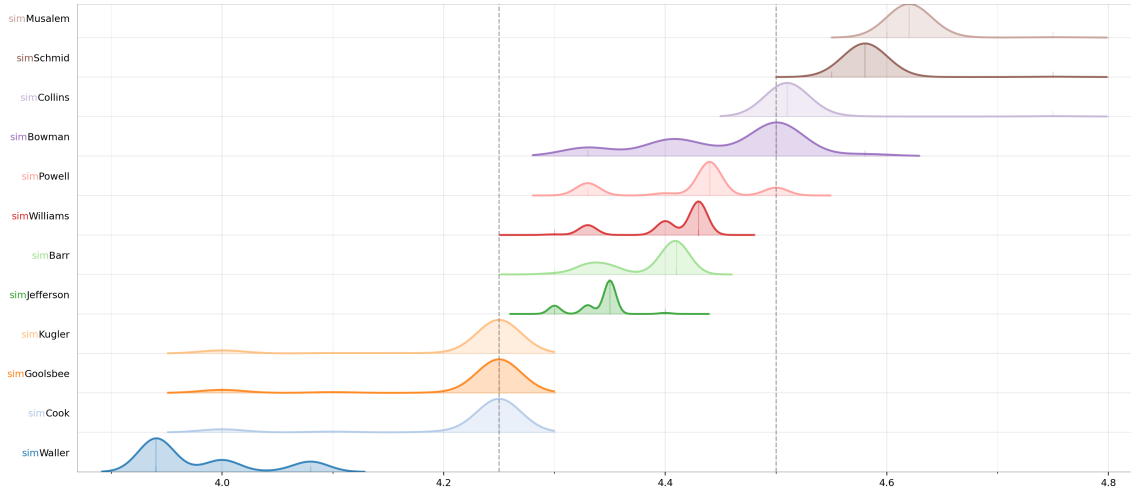



Figure 3: Political Pressure Scenario: Committee Vote Distribution

*Notes:* Each row plots a kernel density estimate of that member's votes (densities normalized within row). Dashed references at 4.25% and 4.50%.

While dissents are less frequent than the most polarized rhetoric might suggest, they are still notable: on average, 0.88 dissents per simulation. This pattern reflects the heightened influence of external pressure, which amplifies both ends of the policy spectrum while still allowing for a negotiated consensus on the midpoint as the institutional decision. This simulation shows that the Federal Reserve is only partially insulated from politics: credible threats to leadership succession can impose a measurable dovish drift without breaching the range defined by prior guidance. Under pressure, clear coalitions form, and policy positions closely reflect members' career incentives. The chair's influence also becomes conditional: when their authority is weakened, decisions tend to move toward the committee's median view rather than the chair's preference. This suggests that outside scrutiny can

shape internal decision-making, even in an institution guided by formal rules.

Next, we summarize the results from the Monte Carlo simulation under the political pressure scenario, incorporating 10,000 simulated policy paths. The Monte Carlo simulation of political pressure scenario confirm the presence of statistically significant political pressure effects. Across 10,000 simulated paths, the final federal funds rate averages 4.38%, roughly eight basis points below the baseline MC mean, and the standard deviation rises to 14 basis points, indicating a noticeably wider spread of outcomes. Although the modal decision remains the upper bound of the standing target range, dispersion around that anchor is much greater. Agendas concentrate at 4.50% (61.3% of cases) and 4.25% (32.6%), with smaller mass at 4.00% (4.6%) and 4.75% (1.5%), and essentially zero at 5.00%. Relative to the prior 4.25–4.50% target range (effective rate 4.33%), the distribution implies a modest upward tilt toward the top of the range, consistent with political pressure nudging decisions higher but rarely beyond a single 25 basis point step.

Political pressure also makes dissent a central feature. At least one dissent occurs in roughly 61 percent of meetings, for a total of 6,113 dissenting votes across all MC simulations, with dissent coming from six unique members: *sim*Waller, *sim*Goolsbee, *sim*Musalem, *sim*Williams, *sim*Collins, and *sim*Schmid. *sim*Waller dissents in 6,062 meetings, while the next most frequent dissenter, *sim*Musalem, appeared only 22 times.

Taken together, the MC results show that even without conversational dynamics or narrative persuasion, credible external pressure both shifts the committee's center of gravity downward and raises the likelihood of open disagreement. The policy path stays inside the established range, preserving formal continuity, yet the combination of a thicker lower tail and frequent dissents signals a materially weaker consensus and a heightened risk that future shocks could push rates further from their baseline trajectory.

## 4.4 Data Revision Scenario

On August 1, 2025, the Bureau of Labor Statistics (BLS) released the July employment report showing a gain of just 73,000 nonfarm payroll jobs, sharply below both

market expectations and the pace of previous months. The report also contained sizable downward revisions to prior months' data, subtracting a cumulative 258,000 jobs from the May and June totals, softening what had previously appeared to be a steady expansion. The release triggered an immediate political reaction. President Trump publicly accused BLS Commissioner Erika McEntarfer of "rigging" the numbers (Financial Times), and announced her dismissal within hours of publication.

From the perspective of monetary policy, the August 1 data release constituted a genuine shock. The disappointing payroll number and large downward revisions signaled that the labor market was weaker than previously thought. This suggested less economic momentum overall and raised the possibility that inflationary pressures in the future might be lower than expected. In our simulation model, we incorporate this scenario by adjusting the reported job growth figures. We set the July employment gain to 73,000 and revise the historical data to reflect the 258,000 job reduction for May and June. These changes directly reduce the recent employment growth rates that agents in the model observe in their macroeconomic data packets. This setup allows us to isolate the perceived economic effects of a weaker labor market.

We begin by re-running the initial belief formation process, given the changes in the underlying data. Table 6 reflects the committee's updated beliefs after a negative economic surprise and public controversy. The overall shift is slightly more dovish, with lower average rates, reduced confidence, and narrower internal spread. These adjustments help quantify how much the data revision reshaped committee thinking, even before formal policy meetings.

Table 7 summarizes the results. Under the data revision shock scenario, the Committee's internal dynamics shift toward a more cautious and data-sensitive posture. Significant downward revisions to payroll data weaken the labor market narrative, prompting many members to recalibrate their stances. The shock moderates previously hawkish voices and strengthens the hand of those favoring lower rates. Simulated Waller remains the most dovish, advocating for rates near 3.92% on average and dissenting in 73.7% of runs, often anchoring the lower end of the distribution. A dovish bloc, including *sim*Waller, *sim*Cook, *sim*Goolsbee, and *sim*Kugler, consistently recommends rates below the midpoint of the 4.25%–4.50% range.

Table 6: Committee Initial Rate Beliefs – August 7th, 2025 Vintage

| Member | Avg. Rate | Confidence | Key Reasoning |
|---|---|---|---|
| *sim*Powell | 4.31% (±0.077) | 79.9% | Balanced neutral stance; high core inflation, and recession risks |
| *sim*Cook | 4.24% (±0.034) | 77.0% | Dovish; supports employment and growth while remaining vigilant about inflation |
| *sim*Goolsbee | 4.25% (±0.000) | 76.5% | Slight dovish shift; supports growth despite high inflation and labor softness |
| *sim*Kugler | 4.24% (±0.034) | 76.3% | Dovish; cautious on weaker labor market data and potential recession risks |
| *sim*Williams | 4.32% (±0.072) | 80.2% | Neutral stance; balances inflation control with labor uncertainty |
| *sim*Jefferson | 4.29% (±0.040) | 80.3% | Neutral; steady rate reflects dual mandate balance amid mixed signals |
| *sim*Barr | 4.28% (±0.053) | 79.2% | Neutral; inflation concerns balanced with labor market considerations |
| *sim*Waller | 4.10% (±0.092) | 74.3% | Slightly dovish shift; accommodates due to weaker labor momentum and risks |
| *sim*Bowman | 4.41% (±0.160) | 83.5% | Typically hawkish; modest accommodation due to labor softness and risk outlook |
| *sim*Schmid | 4.50% (±0.000) | 84.8% | Strongly hawkish; upper-bound rate to combat inflation and preserve stability |
| *sim*Musalem | 4.51% (±0.038) | 84.6% | Hawkish; inflation control priority with stable economic backdrop |
| *sim*Collins | 4.50% (±0.008) | 84.8% | Hawkish; sustained inflation pressures warrant firm stance |

**Committee Initial Rate Beliefs (100 Iterations)**

| | |
|---|---|
| **Avg. Initial Recommendation:** | **4.333% (±0.038)** |
| **Range (Across Iterations):** | 4.10% – 4.51% |
| **Typical Spread:** | 0.410 percentage points (±0.058) |
| **Spread Range:** | 0.00pp – 0.75pp |
| **Members with opinions:** | 12 |
| **Hawkish (≥4.50%):** | 3 members (Schmid, Musalem, Collins) |
| **Moderate (4.41–4.44%):** | 2 members (Powell, Bowman) |
| **Leaning Dovish (4.24–4.35%):** | 5 members |
| **Dovish (<4.24%):** | 2 members (Barr, Waller) |
| **Most Consistent:** | Austan D. Goolsbee ($\sigma$=0.000) |
| **Least Consistent:** | Michelle W. Bowman ($\sigma$=0.160) |

Table 7: Data Revision Scenario: Member Deliberation and Voting

| Simulated Member | Deliberation Rate | Variation | Uncertainty | Vote Bias | | Dissent |
|---|---|---|---|---|---|---|
| | Mean (%) | Std | Mean | Mean | Std | Freq (%) |
| *sim*Waller | 3.919 | 0.381 | 74.35 | -0.3813 | 0.3813 | 73.7 |
| *sim*Bowman | 4.411 | 0.110 | 83.45 | 0.1100 | 0.1100 | 0.0 |
| *sim*Barr | 4.289 | 0.016 | 79.25 | -0.0113 | 0.0167 | 0.0 |
| *sim*Williams | 4.318 | 0.017 | 80.25 | 0.0171 | 0.0171 | 0.0 |
| *sim*Cook | 4.232 | 0.069 | 77.05 | -0.0690 | 0.0690 | 0.0 |
| *sim*Goolsbee | 4.241 | 0.060 | 76.45 | -0.0602 | 0.0604 | 0.0 |
| *sim*Kugler | 4.232 | 0.069 | 76.30 | -0.0689 | 0.0689 | 0.0 |
| *sim*Jefferson | 4.296 | 0.011 | 80.35 | -0.0044 | 0.0114 | 0.0 |
| *sim*Powell | 4.305 | 0.016 | 79.90 | 0.0038 | 0.0158 | 0.0 |
| *sim*Schmid | 4.500 | 0.199 | 84.85 | 0.1991 | 0.1991 | 0.0 |
| *sim*Musalem | 4.510 | 0.209 | 84.60 | 0.2090 | 0.2090 | 0.0 |
| *sim*Collins | 4.500 | 0.199 | 84.80 | 0.1991 | 0.1991 | 0.0 |
| **Final FFR** | **4.301** | 0.016 | | | | |
| **Target Range** | | 4.25% – 4.50% | | | | |
| **Effective FFR** | | 4.38% (midpoint of range) | | | | |

*Notes: Deliberation Rate:* Interest rate each member publicly argued for during formal meeting. *Variation (Std):* Standard deviation of member's true belief across runs. *Uncertainty:* Average confidence (in %) from member's stated reasoning. *Vote Bias:* Difference between member's true belief and final FFR; positive = hawkish, negative = dovish. *Dissent:* Frequency of dissent across runs. Across 99 successful simulations, mean dissents per run were 0.74; 73/99 (73.7%) had exactly one dissent (always simulated Waller). Final-rate distribution: min 4.24%, max 4.32%, mean 4.3009%, std 0.0160%.

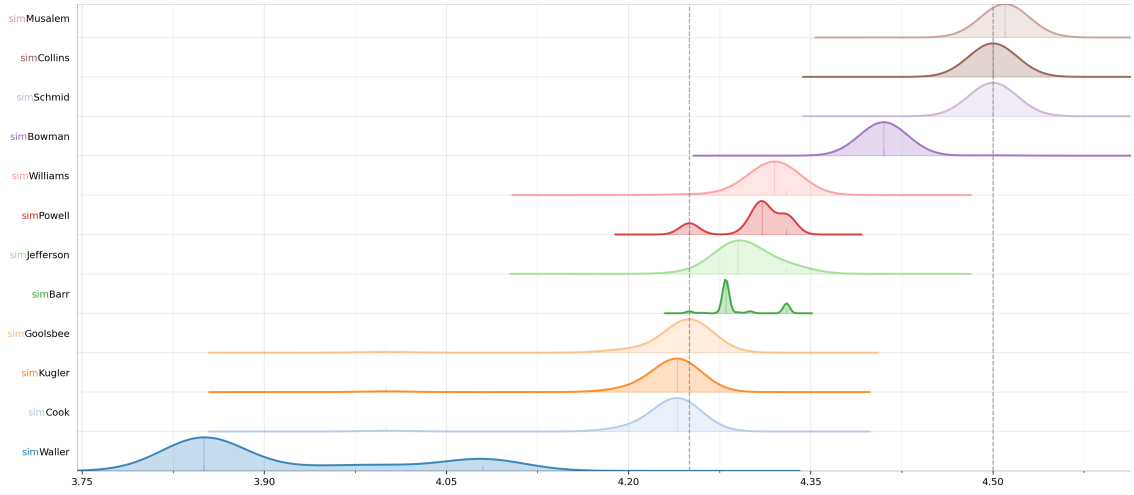



Figure 4: Data Revision Scenario:
Committee Vote Distribution

*Notes:* Each row plots a kernel density estimate of that member's votes (densities normalized within row). Dashed references at 4.25% and 4.50%.

On the hawkish side, *sim*Musalem, *sim*Schmid, and *sim*Collins remain firmly above 4.50%, though their margins narrow slightly in light of softer labor market data. Even traditionally hawkish members such as *sim*Bowman and *sim*Powell register modest downward adjustments in their preferred rates. The resulting distribution of preferences clusters more tightly than under political pressure, reflecting greater convergence toward the middle of the range. The simulated *Final FFR* averages 4.30%, sitting slightly above the lower bound of the prevailing target range, implying no change to the official band. In practice, this outcome signals a "hold" stance that maintains policy settings while acknowledging new information.

We next report results from the Monte Carlo simulation of the data revision shock scenario, incorporating 10,000 simulated policy paths under political pressure.[20] Across all runs, the final federal funds rate averages 4.315%, slightly below the baseline MC mean, with a standard deviation of 15 basis points, signaling a modest widening of the outcome distribution. The modal decision remains at the lower bound of the standing target range, with 42 percent of runs ending at 4.25%, followed closely by 4.41% (21 percent) and 4.50% (20 percent). Outcomes outside the 4.00–4.50% range are rare, and extreme hikes above 4.75% occur in less than one percent of simulations. The median outcome, 4.25%, reflects a distribution tilted slightly toward holding the lower bound.

Labor data revisions together with political pressure amplifies open disagreement within the Committee. One dissent occurs in 61.6 percent of meetings, with dissent drawn from six unique members: *sim*Waller, *sim*Musalem, *sim*Schmid, *sim*Goolsbee, *sim*Collins, and *sim*Williams. *sim*Waller accounts for the overwhelming majority of dissents (6,145 instances) while the next most frequent, *sim*Musalem, appears just 14 times. In sum, the combination of data shock and political pressure keeps the policy path anchored in the prevailing range but erodes consensus. Frequent dissents underscore a weaker center of gravity and a greater susceptibility to shifts if further shocks arise.

[20] Results that show pure data revision scenario without the political pressure add-on are available in the Online Appendix.

# 5 How Political Pressure Reshapes Deliberation

The aggregate results in Section 4 establish that political pressure shifts committee outcomes. In this section, we open the "black box" of deliberation further in order to identify the channels through which this effect operates. We do this through complementary analyses: (1) a narrative approach that classifies rhetorical stance using large language models, and (2) an influence map that quantifies persuasion dynamics within the committee.

## 5.1 Narrative Channel

We apply zero-shot classification following Hansen and Kazinnik (2023) to categorize each statement in the deliberation transcripts as *hawkish*, *mostly hawkish*, *neutral*, *mostly dovish*, or *dovish*. This approach captures contextual meaning beyond simple keyword matching, allowing us to detect subtle shifts in rhetorical framing.

For each statement by member $i$ in meeting $m$, we prompt GPT-4o-mini model to classify each FOMC deliberation statement as one of: dovish, mostly dovish, neutral, mostly hawkish, hawkish. We then convert labels to a numeric scale ($-2$ = dovish, $+2$ = hawkish) and aggregate across the full sample of 3,800 statements. Table 8 reports the aggregate classification results. Political pressure produces a statistically significant dovish shift of $-0.136$ points ($t = 2.95$, $p = 0.003$). The share of statements classified as "hawkish" falls from 35.0% to 30.6%, while "dovish" statements rise from 7.9% to 11.2%.

Table 8: LLM Classification of Deliberation Rhetoric

| **Category** | **Baseline** | **Political** | **Shift** |
|---|---|---|---|
| Hawkish | 35.0% | 30.6% | −4.4% |
| Mostly Hawkish | 10.7% | 12.6% | +1.9% |
| Neutral | 16.7% | 15.6% | −1.2% |
| Mostly Dovish | 29.7% | 30.0% | +0.3% |
| Dovish | 7.9% | 11.2% | +3.3% |
| **Mean Score** | **0.35** | **0.22** | −0.14** |

*Notes*: $N = 3{,}800$ statements (1,900 per scenario). Score scale: $-2$ = dovish, $+2$ = hawkish. Statistical test: $t = 2.95$, $p = 0.003$. $^{**}p < 0.01$.

All hawkish members show dovish shifts (Kugler: $-0.27$, Collins: $-0.22$, Gools-

bee: −0.22, Bowman: −0.21, Schmid: −0.20), while neutral members remain stable. Notably, the shift is asymmetric: members with hawkish baseline positions accommodate more than doves, consistent with direction-dependent career incentives under anticipated regime change.

Disaggregating by prior stance reveals that both hawks and doves shift dovish under pressure. The dovish shift is larger in absolute magnitude for doves, but hawks move a larger proportion of their baseline position. Neutral members, including the Chair, show no significant rhetorical change, suggesting the accommodation effect is concentrated among members with clear prior policy leanings.

## 5.2 Influence Channel: Persuasion Dynamics

Narrative analysis captures *what* members say, but not *how effectively* they shape others' positions. To measure persuasion, we construct an *influence map* that tracks belief updates during deliberation.

For each speaking turn by member $j$, we measure whether subsequent speakers shift their stated positions toward $j$'s view. Formally, influence power is:

$$\text{Influence}_j = \sum_{i \neq j} \sum_{t} \mathbf{1}\{\text{member } i \text{ moves toward } j \text{ at round } t\} \cdot |\Delta\mu_i^t|$$

where $\Delta\mu_i^t$ is member $i$'s stance shift following $j$'s statement.

Table 9 reports influence power by member role. The key finding is that influence reallocates toward likely successors under political pressure. Governor Bowman's influence power more than doubles (from 20.4 to 33.0, shift: +12.6), while Governor Collins loses substantial influence (−8.3). The Chair maintains influence (+5.9), possibly reflecting defensive positioning.

## 5.3 Two Channels, One Mechanism

The narrative and influence analyses reveal complementary aspects of a single mechanism: *anticipatory accommodation* under political pressure. Members soften their rhetoric (narrative channel) and reallocate their responsiveness toward likely

Table 9: Influence Power by Member Role

| Role | Member(s) | Baseline | Political | Shift |
|---|---|---|---|---|
| Chair | Powell | 28.0 | 33.9 | +5.9 |
| Likely Successor | Bowman | 20.4 | 33.0 | +12.6 |
| Hawks (other) | Schmid, Musalem, Collins | 79.8 | 73.6 | −6.2 |

*Notes*: Influence power measures cumulative ability to shift other members' positions toward one's own view. Higher values indicate greater persuasion effectiveness.

successors (influence channel). Both channels operate through forward-looking behavior as members anticipate regime change.

Two features of these results merit emphasis. First, hawks accommodate more than doves, consistent with direction-dependent career incentives. If dovish positions are cheaper under an expected dovish successor, hawkish members face stronger incentives to soften their tone. Second, influence power is not uniformly distributed. The Chair and likely successors command disproportionate ability to move others, and this asymmetry intensifies under political pressure.

# 6 Conditional Validation: Full Historical Sample

The single meeting counterfactual in Section 4 demonstrates the framework's mechanics under controlled conditions. A stronger test asks whether the same pipeline produces accurate rate decisions across the *entire* historical record. This section reports a comprehensive backtest covering 215 FOMC meetings from February 2000 through January 2026, every scheduled and emergency meeting in the sample.

Unlike the unconditional (chained) validation in Section A, the conditional backtest *resets to reality* before each meeting: the simulation receives the actual federal funds rate, the actual FOMC statement, and the true committee roster from the prior meeting. This isolates the framework's single meeting decision quality by eliminating error compounding.

For each of the 215 meetings, committee members receive biographical personas and form individual rate recommendations via GPT-4o, conditioned on the macro environment, the prior meeting's actual rate and statement, and the historical committee composition. Individual opinions are then aggregated through the RRM

agenda setting and voting mechanism described in Sections 3.2–3.3. Both the LLM based framework and the Monte Carlo baseline use 10,000 Monte Carlo draws per meeting for the RRM deliberation stage. As a baseline, we also report the pure RRM outcome, that is, the Monte Carlo median without LLM opinion formation, which isolates the contribution of the game theoretic voting mechanism from the LLM's behavioral priors.

Table 10 presents the aggregate performance of the two approaches.

Table 10: Conditional Backtest: One Step Ahead Accuracy (215 Meetings)

| **Condition** | **RMSE (bp)** | **MAE (bp)** | **Bias (bp)** | **Perfect (%)** | **≤25bp (%)** | **≤50bp (%)** | **Direction Acc. (%)** |
|---|---|---|---|---|---|---|---|
| LLM + RRM Framework | **15.9** | **9.1** | **−0.2** | **54.0** | **93.0** | **99.1** | **77.2** |
| MC baseline (RRM only) | 24.2 | 12.8 | — | — | 74.4 | 91.6 | 66.5 |

*Notes:* All metrics are computed over 215 meetings from February 2000 through January 2026. "Perfect" is the share with error $< 1$ bp. Direction accuracy measures correct classification of hike, cut, or hold. The MC baseline reports the median of 10,000 Monte Carlo RRM draws without LLM opinion formation. Bold marks the better performing approach.

Over half of all meetings (116 of 215) are predicted perfectly, meaning that the simulated rate matches the actual rate exactly. Ninety three percent of meetings fall within 25 basis points of the actual rate, and only two meetings (0.9%) exceed 50 bp of error. The aggregate bias is essentially zero (−0.2 bp), indicating no systematic tendency to over or under predict.

Comparing the LLM based framework to the pure MC baseline shows the marginal contribution of the LLM opinion formation stage. The LLM reduces MAE from 12.8 to 9.1 bp, a 29% improvement, increases the share of meetings within 25 bp from 74.4% to 93.0%, and improves direction accuracy from 66.5% to 77.2%. The LLM's behavioral priors, formed by processing each member's biographical persona, the macro environment, and the prior meeting's language, provide information that the mechanical RRM starting point does not capture.

Table 11 decomposes the framework's direction accuracy by the type of actual policy action.

The framework is most accurate when the Fed moves: it correctly predicts 90% of rate hikes and 88% of rate cuts. Hold accuracy is lower at 71% because holds

Table 11: Direction Accuracy by Action Type

| **Actual Action** | **Meetings** | **Correct** | **Accuracy (%)** |
|---|---|---|---|
| Hike | 40 | 36 | 90.0 |
| Cut | 34 | 30 | 88.2 |
| Hold | 141 | 100 | 70.9 |
| **Overall** | **215** | **166** | **77.2** |

*Notes:* Direction is classified as hike ($\Delta i > 0$), cut ($\Delta i < 0$), or hold ($\Delta i = 0$) relative to the previous meeting's actual rate. A predicted change of $\leq$12.5 bp in either direction is classified based on the sign of the simulated change.

constitute the modal action, 66% of all meetings, and minor predicted perturbations, typically 12.5 bp, are classified as directional errors even when the rate level error is small. These "near hold" misclassifications are concentrated in eras where the simulation correctly identifies rate stability but produces a modest level offset, for example the Pre Crisis period, where 9 consecutive holds are predicted with only 12.5 bp error but classified as directional misses.

Table 12 disaggregates the results by monetary policy era.

Table 12: Conditional Backtest by Monetary Policy Era

| **Era** | $n$ | **RMSE (bp)** | **MAE (bp)** | **Bias (bp)** | **$\leq$25bp (%)** | **Direction Acc. (%)** |
|---|---|---|---|---|---|---|
| Greenspan Late (2000–2006) | 55 | 17.1 | 14.8 | +1.1 | 92.7 | 60.0 |
| Pre-Crisis (2006–2007) | 9 | 12.5 | 12.5 | −6.9 | 100.0 | 0.0 |
| Financial Crisis (2007–2009) | 21 | 20.4 | 14.3 | −1.2 | 76.2 | 81.0 |
| ZIRP (2010–2015) | 47 | 0.0 | 0.0 | +0.0 | 100.0 | 100.0 |
| Normalization (2015–2019) | 30 | 12.1 | 5.8 | −0.8 | 100.0 | 76.7 |
| COVID & Response (2019–2021) | 20 | 18.5 | 6.2 | +6.2 | 95.0 | 95.0 |
| Inflation Hiking (2022–2023) | 16 | 28.0 | 18.8 | −15.6 | 75.0 | 87.5 |
| Easing Cycle (2024–2026) | 17 | 16.0 | 7.4 | +7.4 | 94.1 | 76.5 |
| **Full Sample** | **215** | **15.9** | **9.1** | **−0.2** | **93.0** | **77.2** |

*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 ZIRP era is predicted perfectly. All 47 meetings from January 2010 through October 2015 produce zero error and zero bias, and the simulation independently converges to rates at the effective lower bound when presented with the macro environment of that period. No constraint forces the rate to zero; the LLM agents,

reading the high unemployment and low inflation signals, consistently recommend rate stability.

The framework also handles regime transitions well in the conditional setting. The Financial Crisis era (2007–2009), which includes emergency inter meeting cuts and the rapid descent from 5.25% to near zero, achieves an MAE of only 14.3 bp and direction accuracy of 81%. Similarly, the COVID & Response era achieves 6.2 bp MAE with 95% direction accuracy. When anchored to the correct prior rate, the LLM agents recognize the severity of the economic shock and adjust accordingly.

The largest errors cluster around large, unexpected moves. The 15 meetings with error exceeding 25 bp (Table 13) share a common feature: the actual policy action involved a move of 50 bp or more, or the meeting represented the very first meeting in the sample (February 2000) where no prior synthetic context existed. The framework's modal error on these meetings is 37.5–50 bp, corresponding to under prediction of the magnitude of the rate change.

Table 13: Meetings with Conditional Error Exceeding 25 Basis Points

| **Date** | **Era** | **Actual Rate** | **Predicted Rate** | **Error (bp)** | **Actual Change** |
|---|---|---|---|---|---|
| 2000-02-02 | Greenspan Late | 5.750 | 6.375 | 62.5 | +25 bp |
| 2000-05-16 | Greenspan Late | 6.500 | 6.125 | 37.5 | +50 bp |
| 2002-01-30 | Greenspan Late | 1.750 | 1.375 | 37.5 | hold |
| 2002-11-06 | Greenspan Late | 1.250 | 1.625 | 37.5 | −50 bp |
| 2007-09-18 | Financial Crisis | 4.750 | 5.125 | 37.5 | −50 bp |
| 2008-01-22 | Financial Crisis | 3.500 | 3.875 | 37.5 | −75 bp |
| 2008-01-30 | Financial Crisis | 3.000 | 2.625 | 37.5 | −50 bp |
| 2008-06-25 | Financial Crisis | 2.000 | 1.625 | 37.5 | hold |
| 2008-10-29 | Financial Crisis | 1.000 | 0.625 | 37.5 | −50 bp |
| 2020-03-15 | COVID | 0.125 | 0.875 | 75.0 | −100 bp |
| 2022-06-15 | Inflation Hiking | 1.625 | 1.125 | 50.0 | +75 bp |
| 2022-07-27 | Inflation Hiking | 2.375 | 1.875 | 50.0 | +75 bp |
| 2022-09-21 | Inflation Hiking | 3.125 | 2.625 | 50.0 | +75 bp |
| 2022-11-02 | Inflation Hiking | 3.875 | 3.375 | 50.0 | +75 bp |
| 2024-09-18 | Easing Cycle | 4.875 | 5.375 | 50.0 | −50 bp |

*Notes:* These 15 meetings represent 7.0% of the 215 meeting sample. The maximum error is 75 bp, at the March 2020 emergency cut. No meeting exceeds 75 bp of error in the conditional backtest.

The largest single error, 75 bp on March 15, 2020, corresponds to the emergency inter meeting cut of 100 bp to the zero lower bound. The simulation predicted a

62.5 bp cut, correctly identifying both the direction and the urgency of the action but underestimating the magnitude by one 25 bp grid step. The four 2022 hiking meetings, each with 50 bp error, correspond to the Fed's four consecutive 75 bp hikes where the simulation consistently predicted 50 bp hikes instead. In both cases, the framework correctly reads the regime but underestimates the extremity of the Fed's response.

The conditional backtest establishes that the simulation framework achieves high accuracy when anchored to reality at each step. The 9.1 bp MAE and 93% within 25 bp rate place the framework well within the "noise band" of the FOMC's own 25 bp adjustment grid: the median error, 0 bp, indicates that the modal prediction is exactly correct. The contrast between the conditional results in this section and the unconditional (chained) results in Section A is sharp. Table 14 summarizes the comparison.

Table 14: Conditional vs. Unconditional Validation (LLM Based Framework)

| **Protocol** | **RMSE (bp)** | **MAE (bp)** | **Bias (bp)** | **≤25bp (%)** | **≤50bp (%)** | **Direction Acc. (%)** |
|---|---|---|---|---|---|---|
| Conditional | 15.9 | 9.1 | −0.2 | 93.0 | 99.1 | 77.2 |
| Unconditional | 159.3 | 103.6 | −25.9 | 44.1 | 53.7 | 64.5 |
| Ratio (uncond./cond.) | 10.0× | 11.4× | — | — | — | — |

*Notes:* Conditional results are from this section, covering 215 meetings. Unconditional results are from pooled LLM chains ($K$=3) in Section A, with 214 meetings per chain. The RMSE ratio quantifies the cost of error compounding.

The unconditional RMSE is ten times the conditional RMSE. This order of magnitude gap is almost entirely attributable to error compounding: once the synthetic rate level diverges from reality, the offset persists through subsequent meetings, inflating absolute errors even when meeting level increments are well calibrated. The fact that unconditional direction accuracy, 64.5%, remains reasonably high despite the RMSE inflation confirms that the chain's *incremental* decisions are sensible; it is the *level* that drifts.

The conditional results validate the framework's core mechanism: given accurate inputs, the LLM persona plus RRM deliberation pipeline produces rate decisions that closely match historical reality across a quarter century of diverse monetary

policy regimes. The remaining errors are concentrated at extreme moves of $\geq$75 bp, a pattern consistent with the rarity of such episodes relative to the more common 25 bp adjustments in the historical record.

# 7 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 generates plausible policy decisions and deliberation patterns across a range of counterfactual settings. Political pressure increases dissent and dispersion, while negative labor market revisions shift outcomes in a more dovish direction. In both cases, the main effect is not necessarily a large change in the policy band itself, but a meaningful change in internal disagreement, coalition formation, and the balance of influence across members. Opening the deliberative "black box" shows that these effects operate through identifiable channels: under political pressure, rhetoric softens and influence reallocates toward likely successors, consistent with anticipatory accommodation and career based incentives. Finally, the framework performs well against the historical record. In a conditional backtest covering 215 FOMC meetings from 2000 to 2026, it predicts the policy rate within 25 basis points in 93% of meetings, with a mean absolute error of 9.1 basis points. In a longer horizon validation exercise, it also tracks the broad path of the federal funds rate more closely than Taylor rule and other rule based benchmarks.

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


Ben S Bernanke. The taylor rule: A benchmark for monetary policy? *Ben Bernanke's Blog*, 28(7):4–9, 2015.

Ben S. Bernanke. *21st Century Monetary Policy: The Federal Reserve from the Great Inflation to COVID-19*. W. W. Norton, 2022.

Francesco Bianchi, Roberto Gómez-Cram, Thilo Kind, and Howard Kung. Threats to Central Bank Independence: High-frequency identification with twitter. *Journal of Monetary Economics*, 135:37–54, 2023.

Carola Conces Binder. Political Pressure on Central Banks. *Journal of Money, Credit and Banking*, 53(4):715–744, 2021.

Anton E Bobrov, Rupal Kamdar, Caroline M Paulson, Aditi Poduri, and Mauricio Ulate. Do Local Economic Conditions Influence FOMC Votes? *FRBSF Economic Letter*, 2025(13):1–5, 2025.

Henry W. Chappell and Rob Roy McGregor. Regional Information and FOMC Voting. *Journal of Money, Credit and Banking*, 45(4):643–668, 2013.

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.

Richard Clarida, Jordi Gali, and Mark Gertler. Monetary policy rules and macroeconomic stability: evidence and some theory. *The Quarterly journal of economics*, 115(1):147–180, 2000.

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.

Financial Times. Donald Trump's Attack on Statistics Agency Echoes Strongmen Leaders, Economists Say. *Financial Times*.

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.

Nadine Hinterlang and Lukas Tänzer. Deep Reinforcement Learning for Monetary

Policy Design. *Computational Economics*, 58:123–149, 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.

Anton Korinek. Generative ai for economic eesearch: Llms learn to collaborate and reason. Technical report, National Bureau of Economic Research, 2024.

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.

Allan H. Meltzer. *A History of the Federal Reserve, Volume 2: Book 2, 1970–1986*. University of Chicago Press, 2009.

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.

John B Taylor. Discretion versus policy rules in practice. In *Carnegie-Rochester conference series on public policy*, volume 39, pages 195–214. Elsevier, 1993.

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.

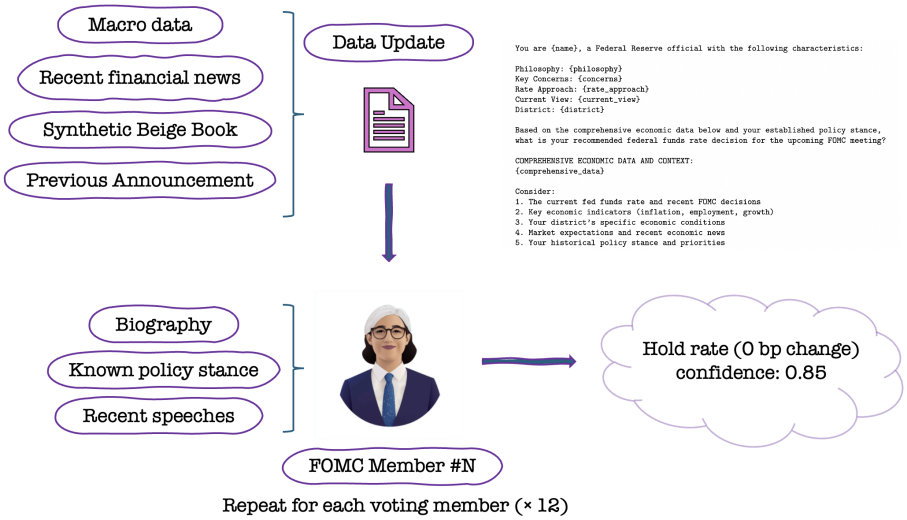
Chong Zhang, Xinyi Liu, Zhongmou Zhang, Mingyu Jin, Lingyao Li, Zhenting Wang, Wenyue Hua, Dong Shu, Suiyuan Zhu, Xiaobo Jin, et al. When AI Meets Finance (Stockagent): Large Language Model-based Stock Trading in Simulated Real-World Environments. *arXiv preprint arXiv:2407.18957*, 2024.

# Online Appendix for

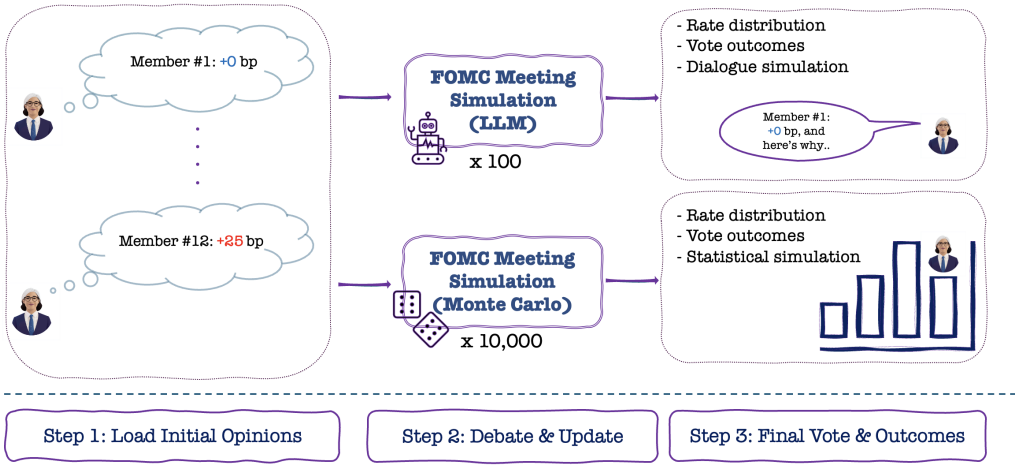
# "FOMC *In Silico*: A Multi-Agent System for Monetary Policy Decision Modeling"


## Abstract

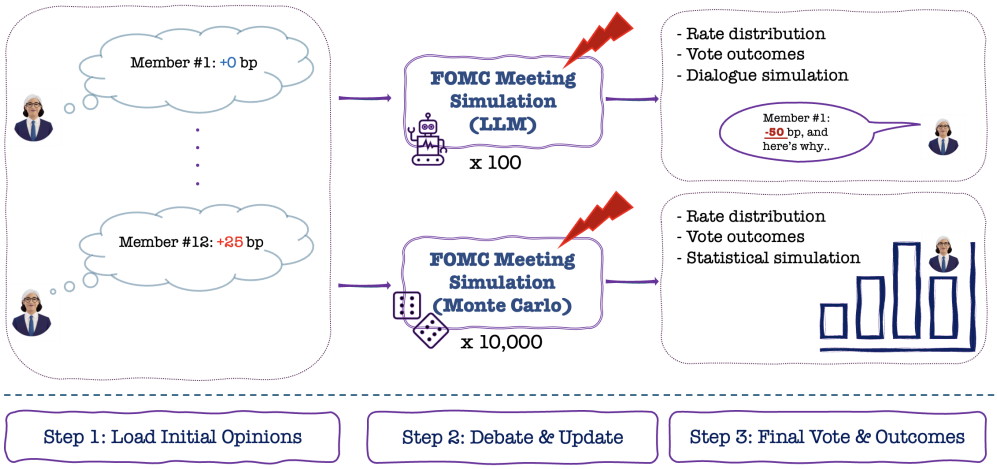
*This Internet Appendix provides additional material on the data, methodology, and some robustness exercises related to the main text.*




(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 data revisions) to this shared environment and recomputes outcomes in both tracks.

# Prompts

## Part 1: Individual Opinion Formation

### Main Opinion Generation Prompt

```
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
```

## Part 2: Meeting Deliberation

### System Prompt (Baseline)

```
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!
```

### System Prompt (Political Pressure)

```
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.

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!
```

### Chair Opening Prompt

```
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.
```

## .1 Additional Robustness Checks

Here, we present the results for data revision shock (without the political pressure treatment). In the LLM track, dovish members who prioritize employment become more uncertain when labor market conditions appear weaker. Deliberation transcripts show this as either caution in interpreting the revision or skepticism about its durability. The Chair's instructions include a nudge about the labor data revisions.[21]

Table .15: Data Revision Scenario: Member Deliberation and Voting (as of July 7th)

| Simulated Member | Deliberation Rate | Variation | Uncertainty | Vote Bias | | Dissent |
|---|---|---|---|---|---|---|
| | Mean (%) | Std | Mean | Mean | Std | Freq (%) |
| Austan D. Goolsbee | 4.250 | 0.036 | 0.822 | -0.1704 | 0.0360 | 0.0 |
| Lisa D. Cook | 4.240 | 0.025 | 0.820 | -0.1704 | 0.0250 | 0.0 |
| Adriana D. Kugler | 4.240 | 0.019 | 0.818 | -0.1704 | 0.0190 | 0.0 |
| Michael S. Barr | 4.280 | 0.088 | 0.848 | -0.0159 | 0.0880 | 0.0 |
| Philip N. Jefferson | 4.290 | 0.060 | 0.848 | -0.0710 | 0.0600 | 0.0 |
| John C. Williams | 4.320 | 0.086 | 0.850 | 0.0033 | 0.0860 | 0.0 |
| Jerome H. Powell | 4.310 | 0.080 | 0.848 | 0.0184 | 0.0800 | 0.0 |
| Christopher J. Waller | 4.100 | 0.088 | 0.770 | -0.2181 | 0.0880 | 0.0 |
| Michelle W. Bowman | 4.410 | 0.208 | 0.830 | 0.0449 | 0.2080 | 0.0 |
| Jeffrey R. Schmid | 4.500 | 0.117 | 0.849 | 0.1620 | 0.1170 | 0.0 |
| Alberto G. Musalem | 4.510 | 0.123 | 0.846 | 0.2017 | 0.1230 | 0.0 |
| Susan M. Collins | 4.500 | 0.036 | 0.850 | 0.0934 | 0.0360 | 0.0 |
| **Baseline FFR** | **4.3292** | 0.0229 | | | | |
| **Target Range** | | | 4.25% – 4.50% | | | |
| **Effective FFR** | | | 4.33% | | | |

*Notes:* In practice, the FOMC sets the target range for the federal funds rate in 25 basis point increments. The simulated *Final FFR* of 4.3292% falls within the 4.25%–4.50% range, corresponding to an effective rate near the midpoint (4.33%). *Deliberation Rate:* The interest rate each member publicly argues for during the formal meeting. *Variation (Std):* Standard deviation of the deliberation rate across 100 simulations. *Uncertainty:* Member's self-reported confidence/uncertainty (decimal form). *Vote Bias:* Difference between the deliberation rate and the simulated final FFR; positive = more hawkish, negative = more dovish.

Comparing these results to the baseline simulation, the labor data shock scenario produces a modest but measurable dovish drift in the committee's stance. The simulated final FFR declines from 4.420% in the baseline to 4.329%, an 9 basis point reduction, while remaining within the same 4.25%–4.50% target range. This shift reflects the Chair's slightly more dovish opening proposal and the increased weight given to the median committee signal during synthesis.

The largest effects are concentrated among employment-focused members. *sim*Goolsbee, *sim*Cook, and *sim*Kugler each maintain their baseline deliberation rates but show no hawkish shift, while their uncertainty remains elevated, making them marginally more inclined toward rate holds or cuts in close call scenarios. Notably, these members' vote biases stay dovish at 0.1704, but without the upward drift in rates observed in the baseline when labor momentum is perceived as stronger. Hawkish members, including *sim*Schmid and *sim*Musalem, retain higher deliberation rates but edge down slightly from their baseline positions (–8 and –11 basis points, respectively), narrowing the gap with the committee median. Variation across simulations remains essentially

[21] `NOTE: Labor data revision shows weaker payroll momentum; prior months overestimated. Treat labor as softer.`

unchanged, indicating that the dovish shift stems more from central tendency adjustments than from increased dispersion.

In sum, the scenario moves the committee toward a slightly lower and more consensus oriented outcome, softening the influence of more hawkish voices and reinforcing the weight of the median when the labor market is seen as weaker and data credibility is questioned.

In the Monte Carlo track, the labor market shock enters as a downward adjustment to the employment gap parameter in the initial belief formation stage, with a concurrent increase in posterior variance for employment-sensitive members. Inflation expectations and other inputs are held constant to isolate the labor data effect. Voting then proceeds under the usual Bayesian updating and institutional rules.

Comparing the data revision shock to the baseline MC simulation, the mean final rate declines from 4.376% to 4.321%, a reduction of roughly 5.5 basis points on a continuous scale. While both scenarios keep outcomes firmly within the existing 4.25%–4.50% target range (effective rate 4.33%), the shock tilts the distribution modestly downward, placing the mean just below the midpoint rather than slightly above it.

The agenda rate shows a larger adjustment, falling from 4.377% in the baseline to 4.297% under the shock, suggesting that the initial proposals entering deliberation are already shaped by weaker labor market assessments. This leaves less upward room for subsequent negotiation, and the Chair's moderation narrows the spread from proposal to final decision.

Dispersion in final rates remains virtually unchanged (0.074 vs. 0.075), and the absence of dissents in both cases underscores the strong central consensus despite the shock. The minimum and maximum simulated outcomes shift downward by similar magnitudes (about 13–15 basis points) reflecting a uniform leftward shift in the simulated rate distribution rather than a widening of tails. Overall, the Monte Carlo framework yields a slightly more dovish and consensus-driven outcome under the data revision shock, with lower agenda anchors feeding directly into marginally lower final rates.

Table .16: Baseline Scenario: MC Simulation Outcomes (Data Revision Shock)

| **Panel A: Summary Statistics** | | |
|---|---|---|
| Final rate (mean) | 4.321 | Average simulated terminal rate (continuous scale) |
| Final rate (s.d.) | 0.075 | Dispersion across simulations (continuous scale) |
| Final rate (min) | 4.058 | Minimum simulated terminal rate |
| Final rate (max) | 4.574 | Maximum simulated terminal rate |
| Agenda rate (mean) | 4.297 | Average initial proposal before deliberation |
| Agenda rate (s.d.) | 0.103 | Dispersion of agenda rates |
| Dissent rate | 0.0 | Share of simulations with at least one formal dissent |
| Average dissenters | 0.0 | Mean number of dissenters per simulation |
| Observations | 10,000 | Number of simulated scenarios |

**Panel B: Interpretation**

Final rates cluster within the 4.25–4.50% target range in place before the meeting (effective federal funds rate 4.33%). The mean of 4.321% lies one basis point from the midpoint, implying a modest downward tilt when viewed on a continuous scale. In actual policy terms, this corresponds to holding the rate steady, with no formal dissents recorded. The Chair's moderation narrows variability from agenda to final decision.

# A Historical Validation

The counterfactual analysis in Section 4 evaluates the framework at a single point in time, holding economic conditions fixed and varying institutional parameters. A natural follow up is to ask whether the framework can produce plausible *trajectories* of the federal funds rate over extended horizons (a more demanding test in which errors compound rather than reset at each meeting).

In Section 4 we anchor a simulated meeting to the actual rate and announcement from the prior meeting. This section performs *unconditional* validation, in which the synthetic output at meeting $t$ feeds forward as the input to meeting $t+1$. The design tests whether the framework, given only the FOMC roster and real time data, can reconstruct the broad developments of U.S. monetary policy over a quarter century.

## A.1 Methodology

We simulate a sequential chain of 214 FOMC meetings from February 2000 through December 2025. At each step $t$, the simulation receives:

1. the synthetic rate decision and FOMC statement from step $t-1$;
2. the historical FOMC roster for that meeting (committee composition and information);
3. annual macroeconomic indicators (CPI inflation, real GDP growth, unemployment rate, federal funds rate) for the relevant vintage year, sourced from FRED annual averages with a one year lag for January–June meetings to avoid look-ahead bias.

We compare three approaches, each representing a distinct point on the spectrum from mechanical rule to behavioral simulation. First, a purely mechanical rule based on Taylor (1993) with Clarida–Galí–Gertler inertial smoothing (Clarida et al., 2000):

$$i_t^* = r^* + \pi_t + 0.5\,(\pi_t - \pi^*) + 0.5\,(y_t - y^*), \tag{A.5}$$

$$\hat{i}_t = \rho\,\hat{i}_{t-1} + (1-\rho)\,i_t^*, \tag{A.6}$$

where $r^* = 0.5\%$ is the assumed real neutral rate, $\pi^* = 2.0\%$ the inflation target, $y^* = 2.5\%$ trend output growth, and $\rho = 0.85$ the interest-rate smoothing coefficient. Rates are rounded to the nearest 25 basis point increment to match the FOMC's discrete adjustment grid. This benchmark contains no institutional or behavioral parameters and serves as a lower bound on what any simulation framework should beat.

We also run five independent chains ($K = 5$) using the Riboni and Ruge-Murcia (2010) agenda setting model described in Sections 3.2–3.3. At each meeting, member beliefs are initialized from a Taylor-rule baseline with stance dependent offsets (Hawkish +25 bp, Neutral 0, Dovish −25 bp), then processed through Bayesian updating, the Chair's agenda proposal, and the committee vote. A template FOMC statement is generated from the macro data and rate decision, propagating forward into the next meeting's information set. Because the RRM chains do not use an LLM at any stage, they isolate the contribution of the game theoretic voting structure.

Finally, we run three independent chains ($K = 3$) using the full LLM opinion formation pipeline. Each committee member receives a biographical persona (constructed from historical records) and forms an individual rate recommendation via GPT-4o-mini, conditioned on the macro environment, the synthetic chain history, and a Taylor rule anchor injection to mitigate drift. A numeric voting mechanism aggregates opinions, and an LLM generated statement propagates forward. This is the same pipeline described in Sections 3–3.1, applied iteratively across 214 meetings.

## A.2 Taylor Rule Benchmark

Before turning to the simulation approaches, we establish the Taylor rule's performance as a mechanical baseline. Table A.17 reports the inertial Taylor rule's accuracy by monetary policy era.

The Taylor rule performs well during the Normalization era (2015–2019), when policy evolved gradually along a predictable tightening path (RMSE = 26 bp, 72% of meetings within 25 bp). It struggles most during the 2022–2023 inflation-hiking cycle (RMSE = 375 bp), where the inertial

Table A.17: Taylor Rule Benchmark: Error Metrics by Monetary Policy Era

| **Era** | $n$ | **RMSE (bp)** | **MAE (bp)** | **≤25bp (%)** | **≤50bp (%)** |
|---|---|---|---|---|---|
| Greenspan Late (2000–2006) | 55 | 134 | 120 | 16.4 | 21.8 |
| Pre-Crisis (2006–2007) | 8 | 75 | 75 | 0.0 | 0.0 |
| Financial Crisis (2007–2009) | 22 | 256 | 216 | 9.1 | 18.2 |
| ZIRP (2010–2015) | 47 | 211 | 188 | 10.6 | 12.8 |
| Normalization (2015–2019) | 29 | 26 | 21 | 72.4 | 93.1 |
| COVID & Response (2019–2021) | 21 | 168 | 133 | 14.3 | 23.8 |
| Inflation Hiking (2022–2023) | 16 | 375 | 366 | 0.0 | 0.0 |
| Easing Cycle (2024–2026) | 17 | 105 | 102 | 0.0 | 0.0 |
| **Full Sample** | **215** | **188** | **148** | **18.6** | **25.1** |

*Notes:* Inertial Taylor rule (Equation A.6) with $\rho = 0.85$. RMSE and MAE in basis points. "≤25bp" and "≤50bp" report the percentage of meetings where the TR prescribed rate falls within 25 or 50 basis points of the actual federal funds rate. Macro data sourced from FRED annual averages; meetings in the first half of the year use prior year data to avoid look ahead bias.

smoothing parameter prevents the rule from tracking the Fed's aggressive tightening, and during the ZIRP era (RMSE = 211 bp), where the rule's positive prescriptions conflict with the effective lower bound. The overall bias of +122 bp reflects the well-documented tendency of Taylor rules to prescribe higher rates than the Fed actually sets (Bernanke, 2015).

## A.3 Results

Table A.18 presents the head-to-head comparison across all three approaches. The RMSE in LLM simulation of 159 bp is 15% lower than the Taylor rule (188 bp) and 32% lower than the RRM model (233 bp). Nearly half (44%) of its meeting-level predictions fall within 25 basis points of the actual rate, compared to 19% for the Taylor rule and 12% for the RRM. The direction accuracy gap is equally pronounced: the LLM correctly identifies whether the Fed hiked, cut, or held in 64.5% of meetings, versus approximately 48% for both mechanical alternatives, only marginally better than a one-third random baseline.

Table A.18: Unconditional Validation: Taylor Rule vs. RRM vs. LLM

| **Approach** | **RMSE (bp)** | **MAE (bp)** | **Bias (bp)** | **≤25bp (%)** | **≤50bp (%)** | **Direction Acc. (%)** |
|---|---|---|---|---|---|---|
| Taylor (inertial) | 188 | 148 | +122 | 18.6 | 25.1 | 49.5 |
| RRM ($K$=5) | 233 | 158 | +135 | 12.3 | 26.0 | 46.5 |
| LLM ($K$=3) | **159** | **104** | **−26** | **44.1** | **53.7** | **64.5** |

*Notes:* All metrics computed over 214 FOMC meetings (February 2000–December 2025). RRM and LLM figures are pooled across all chains ($K \times 214$ observations each). Direction accuracy measures the share of meetings where the model correctly predicts hike, cut, or hold. Bias is the mean signed error; positive values indicate the model prescribes higher rates than actual. Bold marks the best-performing approach.

Both the Taylor Rule (+122 bp) and the RRM (+135 bp) systematically overshoot, prescribing higher rates than the Fed actually sets across the full 25-year sample. The LLM chains, by contrast, exhibit a small *downward* bias (−26 bp). This suggests that the LLM agents' behavioral reasoning partially captures the Fed's revealed preference for rates below Taylor-rule prescriptions, a "dovish tilt" rooted in the language of FOMC statements and member speeches that purely mechanical approaches discard.

Table A.19 disaggregates performance by monetary policy era, revealing where each approach has comparative advantage.

Table A.19: RMSE by Monetary Policy Era (basis points)

| **Era** | $n$ | **Taylor** | **RRM** | **LLM** |
|---|---|---|---|---|
| Greenspan Late (2000–2006) | 55 | 134 | **117** | 161 |
| Pre-Crisis (2006–2007) | 8 | **75** | 71 | 356 |
| Financial Crisis (2007–2009) | 22 | 256 | 212 | **138** |
| ZIRP (2010–2015) | 47 | 211 | 192 | **13** |
| Normalization (2015–2019) | 29 | 26 | 66 | **25** |
| COVID & Response (2019–2021) | 21 | **168** | 375 | 283 |
| Inflation Hiking (2022–2023) | 16 | 375 | 560 | **217** |
| Easing Cycle (2024–2026) | 17 | 105 | **55** | 79 |

*Notes:* RMSE in basis points for the inertial Taylor rule, pooled RRM chains ($K$=5), and pooled LLM chains ($K$=3). Bold indicates the best-performing approach in each era. Sample sizes ($n$) count meetings; RRM and LLM contribute $K \times n$ observations per era.

We show that during the ZIRP era (2010–2015), the LLM achieves a RMSE of just 13 bp, an order of magnitude lower than either alternative (Taylor: 211 bp; RRM: 192 bp). All three LLM chains independently converge to rates near zero and remain there for 47 consecutive meetings. The LLM agents appear to recognize that rates should stay at the effective lower bound when unemployment is elevated and inflation is subdued, without any explicit zero-lower-bound constraint in the model. Similarly, during the Normalization era (2015–2019), the LLM matches the Taylor rule (25 bp vs. 26 bp) and substantially outperforms the RRM (66 bp), accurately tracking the gradual tightening path.

In the Greenspan Late era (2000–2006), where rates traced a relatively smooth tightening/easing, the RRM achieves the lowest RMSE (117 bp), modestly beating both the Taylor rule (134 bp) and the LLM (161 bp). The same holds for the Easing Cycle (2024–2026), where the RRM's 55 bp RMSE reflects good calibration to gradual, predictable rate adjustments. This pattern is expected: when the actual policy path closely resembles the game theoretic equilibrium, the RRM's Taylor-type belief initialization gives it a structural advantage.

The COVID emergency (2019–2021) and the subsequent inflation hiking cycle (2022–2023) prove challenging across the board, with RMSE ranging from 168 to 560 bp. The unprecedented rate movements (emergency cuts to zero in March 2020, followed by 525 bp of cumulative tightening in 2022–2023) are inherently difficult to replicate in a forward chained simulation where momentum from the prior regime must be overcome. Nonetheless, the LLM outperforms during the 2022–2023 hiking cycle (217 bp vs. 375 bp for Taylor, 560 bp for RRM), suggesting that its conditioning on the evolving FOMC narrative helps it recognize the shift in the policy regime.

The LLM's elevated Pre-Crisis RMSE (356 bp) does not indicate that the model misidentifies the 2006–2007 environment. Rather, it reflects accumulated rate level divergence from earlier eras: the three LLM chains entered the pre-crisis period with a persistent level offset inherited from the early 2000s easing cycle, inflating absolute errors even when the *direction* of policy was correct. This is a fundamental property of unconditional validation (errors compound) and one reason the conditional results in Section 4 should be read alongside these findings.

Figure A.6 displays the fan chart of synthetic rate paths alongside the actual federal funds rate and the Taylor rule benchmark.

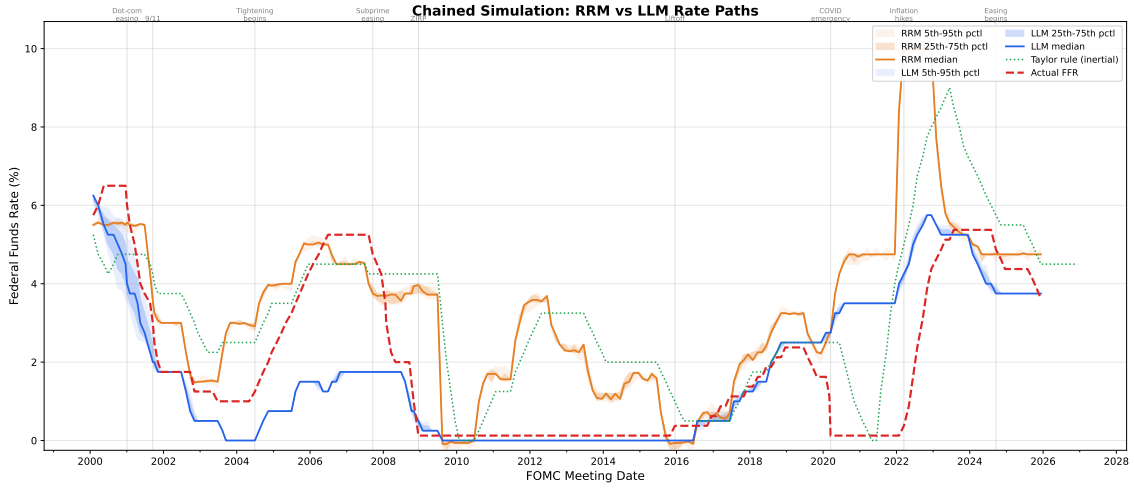



Figure A.6: Chained simulation fan chart: synthetic federal funds rate paths from RRM chains (orange, $K$=5) and LLM chains (blue, $K$=3) compared to the actual rate (red dashed) and the inertial Taylor rule (green dotted). Shaded bands represent the 25th–75th and 5th–95th percentile ranges. Vertical lines mark major policy events.

The LLM bands (blue) track the actual rate path far more closely than either the RRM bands (orange) or the Taylor rule (green), particularly during the 2008–2009 crisis, the ZIRP era, and the 2022–2023 hiking cycle. The RRM bands are wider and share the Taylor rule's persistent positive level shift. All three approaches converge during the 2015–2019 Normalization era, consistent with the low RMSE values in Table A.19.

Figure A.7 traces the rolling 8-meeting mean absolute error over time, showing how accuracy evolves as economic conditions shift. The LLM maintains the lowest rolling error through most of the sample, with clear convergence during the stable 2015–2019 period and error spikes corresponding to regime transitions.

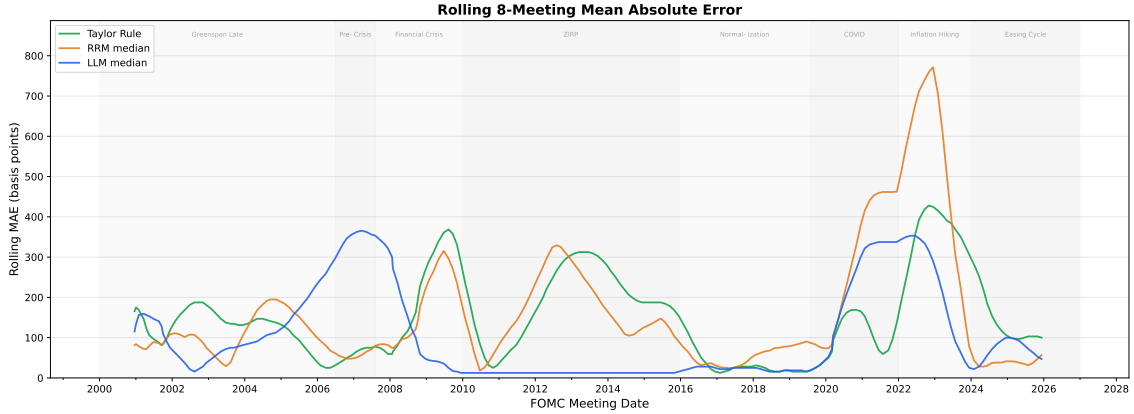



Figure A.7: Rolling 8-meeting mean absolute error. The LLM (blue) maintains the lowest rolling error through most eras. Error spikes correspond to the 2007–2008 crisis, the 2020 COVID shock, and the 2022 hiking cycle.

Figure A.8 plots cumulative absolute error over the full sample. By the end of 2025, the LLM

median has accumulated roughly 224 percentage points of total absolute error, compared to 317 for the Taylor rule and 339 for the RRM median. The LLM's advantage is not concentrated in a single era; rather, its flatter slope during the 2010–2019 decade (ZIRP through Normalization) reflects near-zero incremental error during this 76-meeting stretch.

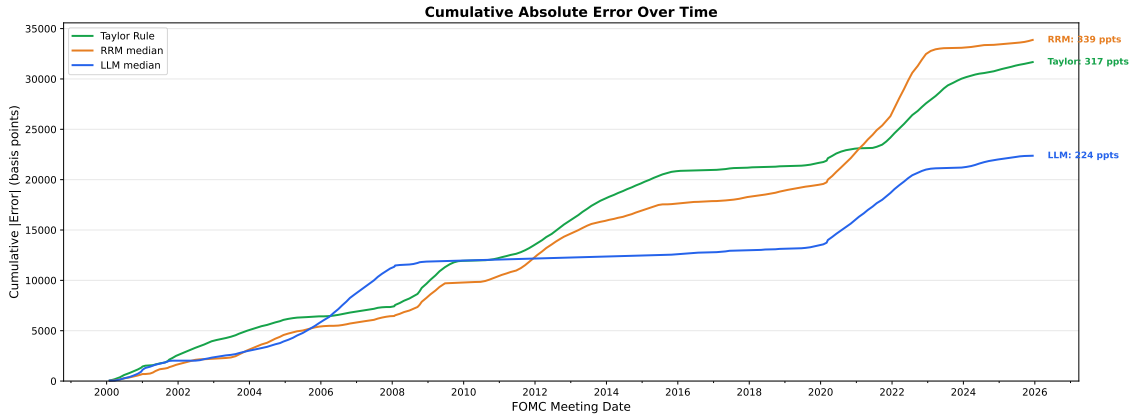



Figure A.8: Cumulative absolute error. The LLM accumulates errors at a slower rate throughout the sample, with the gap widening during the ZIRP and Normalization eras.

Figure A.9 decomposes direction accuracy by action type. The LLM's overall advantage (64.5% vs. ∼48%) is driven almost entirely by its ability to correctly predict hold decisions (80% vs. 54% for the RRM and 54% for Taylor). All three approaches perform similarly on hikes and cuts (roughly 30–35%), suggesting that the LLM's edge lies in recognizing when the Fed will *not* move rather than predicting the timing of rate changes.

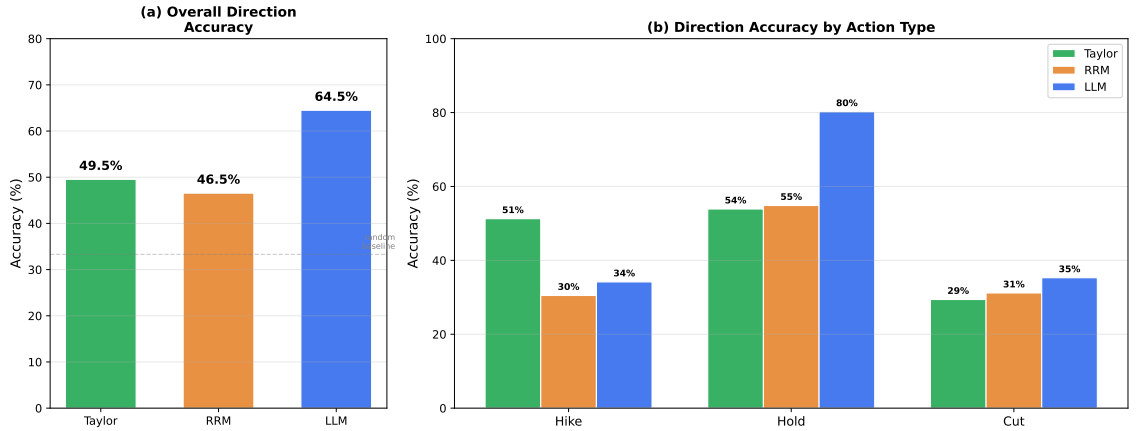



Figure A.9: Direction accuracy. Panel (a): overall accuracy. Panel (b): by action type. The LLM's advantage is concentrated in hold decisions, where it achieves 80% accuracy. The dashed line in panel (a) marks the 33% random baseline.

Finally, Figure A.10 presents the distribution of signed prediction errors. The Taylor rule and RRM distributions are centered well above zero (median errors of +112 and +93 bp), confirming their systematic upward bias. The LLM distribution is centered near zero (median −12 bp) with a substantially tighter interquartile range, consistent with its lower RMSE and near zero aggregate bias.

The chained validation establishes that the LLM framework, originally designed and validated for conditional (single meeting) analysis in Section 4, also generates plausible 25-year trajectories of

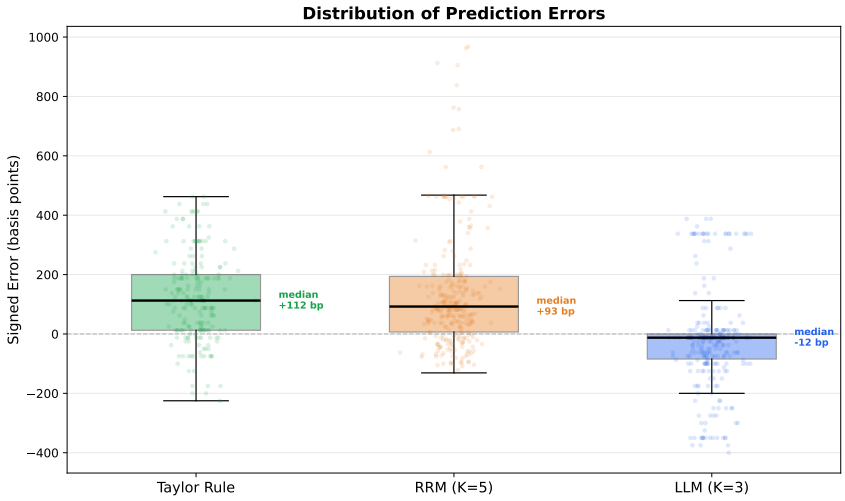



Figure A.10: Distribution of signed prediction errors (basis points). The Taylor rule and RRM exhibit positive median errors; the LLM distribution is centered near zero with less dispersion.

the federal funds rate.

The LLM chains' performance stems from two reinforcing mechanisms. First, the LLM's conditioning on the synthetic FOMC statement allows it to internalize contextual signals that a mechanical rule discards. When the statement warns of "significant downside risks to growth" or notes "inflation expectations remain well-anchored," the LLM agents adjust their rate recommendations in ways that a Taylor rule, which sees only annual macro aggregates, cannot. Second, the Taylor rule anchor injection prevents the model from fixating on the prior meeting's rate (a common LLM failure mode) while still allowing the model's behavioral "judgment" to deviate from the mechanical prescription. The result is a system that benefits from both the discipline of a rule-based anchor and the flexibility of language-conditioned reasoning.

The sign of the aggregate bias tells a complementary story. The Taylor rule's +122 bp bias and the RRM's +135 bp bias confirm that the Fed has historically set rates below what simple rules prescribe, the well-documented "Taylor rule gap" (Bernanke, 2015). The LLM's near-zero bias (−26 bp) suggests that behavioral reasoning, informed by the qualitative texture of FOMC communications, implicitly captures the institutional preference for gradualism and risk management that drives this gap.

The era-level decomposition reveals that no single approach dominates uniformly. The RRM performs best when policy follows a textbook trajectory (Greenspan Late, Easing Cycle), the Taylor rule provides the most stable benchmark during COVID, and the LLM excels in regimes characterized by the effective lower bound or rapid policy adjustment. This heterogeneity suggests that an ensemble approach could potentially improve on either track alone, a direction we leave to future research.

This exercise uses annual macro data, which limits the model's ability to respond to within-year shifts in economic conditions. Higher-frequency data (e.g., monthly CPI, weekly jobless claims) or real time financial market inputs (yield curves, equity volatility) would likely improve trajectory accuracy, particularly at regime transitions. The LLM chains also rely on a single model (GPT-4o-mini); robustness across model families remains untested. Finally, three LLM chains may be too few to fully characterize the distribution of behavioral outcomes; we are constrained by the computational cost of running 214 sequential LLM-based meetings per chain.