Measuring Monetary Policy Uncertainty Through an Agentic AI Framework: The Virtual Federal Reserve

Chi Wang
G. Brint Ryan College of Business, University of North Texas*

May 2026

Abstract

I introduce the *Virtual Federal Reserve*, a multi-agent simulation of FOMC deliberations calibrated from 1,806 speeches and 239 meetings (1996–2025), achieving 91.7% per-voter accuracy. I construct the *Agentic AI MPU*, the mean simulated dissent probability before institutional herding suppresses it,orthogonal to both the Baker–Bloom–Davis index ($r = 0.054$) and the Bauer–Swanson measure ($r = 0.009$). On average, 39.3 percentage points of latent dissent are masked per meeting. A SHAP decomposition ($R^2 = 0.88$) identifies the federal funds rate as the dominant uncertainty driver, while inflation importance triples post-COVID, documenting a regime shift in the policy reaction function.(98 words)

JEL Classification: E52, E58, G14, C45, C55
Keywords: FOMC, monetary policy uncertainty, multi-agent LLM, central bank communication, dissent suppression, SHAP

*Address: 1155 Union Circle #305339, Denton, TX 76203. Email: Chi.Wang@unt.edu.

1 Introduction

The Federal Open Market Committee (FOMC) is among the most consequential deliberative bodies in global finance. Its decisions on the federal funds rate directly shape borrowing costs, asset valuations, and macroeconomic outcomes worldwide. Despite the committee's expanding transparency efforts,publishing statements, minutes, and (with a five-year lag) full deliberation transcripts,the process by which individual members arrive at their votes, and the degree of genuine disagreement that precedes the official consensus, remains only partially understood.

Prior literature have measured the monetary policy uncertainty (MPU) in various way. The Baker–Bloom–Davis (BBD) index (Baker et al., 2016) counts newspaper articles that simultaneously reference the Federal Reserve, uncertainty, and policy, capturing public ambiguity as filtered through media framing. The Bauer–Swanson monetary policy surprise (MPS) (Bauer and Swanson, 2023) isolates the unanticipated component of FOMC decisions from intraday interest rate futures movements, capturing market-priced surprise. Both measures share a structural blind spot: they are necessarily external to the committee. The BBD index reflects how journalists and markets perceive policy ambiguity after communications have been filtered through institutional norms of consensus. The Bauer–Swanson shock captures residual price movements on decision dates, conflating true policy uncertainty with information asymmetry between the Fed and markets. Neither measure can access the degree of genuine disagreement within the committee before herding dynamics and institutional conformity suppress it.

This paper fills that gap. I ask: *can large language model agents, grounded in each member's documented policy stance and speech history, plausibly replicate the vote distribution of historical FOMC meetings and, in doing so, recover the latent disagreement hidden from the public record?* I build the Virtual Federal Reserve, a fully reproducible simulation spanning 239 FOMC meetings from January 1996 to December 2025 in which one LLM agent represents each voting member. Each agent's "personality" is constructed from three empirical ingredients: (i) a quantitative stance prior derived from four independent text-scoring methods applied to 1,806 official speeches; (ii) date-gated retrieval of the member's own past statements via Retrieval-Augmented Generation (RAG), ensuring no future information contaminates the simulation; and (iii) the member's full voting history, including any recorded dissents. A sequential deliberation structure mirrors the actual FOMC process, with each agent observing prior speakers' reasoning before casting its vote.

Our simulation achieves a rate-projection accuracy of 62.8% and a per-voter hit rate of 91.7% across 239 historical meetings. The key output is the *Agentic AI MPU*: the cross-

sectional mean of each agent's simulated probability of true dissent before herding suppresses it. On average, 39.3 percentage points of latent dissent are masked per meeting; 97.3% of the 74 speech-implied dissent events in the sample result in an official YES vote. The Agentic AI MPU is empirically near-orthogonal to BBD ($r = 0.054$) and Bauer–Swanson ($r = 0.009$), identifying an independent third dimension of MPU: internal committee disagreement deliberately obscured from the public record.

A gradient-boosted SHAP decomposition reveals systematic variation in the macro drivers of simulated uncertainty across monetary policy regimes. The federal funds rate is the dominant driver overall (mean |SHAP| = 0.093), but inflation importance triples in the post-COVID inflation era to 0.046, 2.6 times the sample mean, while unemployment importance nearly doubles, documenting a dual-mandate tension not seen in the SHAP structure since the Volcker era. Mean Agentic AI MPU rises monotonically across Fed chairs from Greenspan (0.299) to Powell (0.545), a secular trend that is orthogonal to both the level of official dissent and the expansion of formal transparency tools.

I make four contributions. First, I introduce the Agentic AI MPU, the first MPU measure derived from multi-agent simulation of committee deliberation rather than from text-frequency counts or market prices. Second, I document near-orthogonality to existing indices, suggesting the measure captures internal disagreement suppressed by herding norms,a distinct economic object with independent implications for monetary policy credibility and financial market pricing. Third, I build a fully reproducible, open-source framework (239 meetings, 73 canonical voters, 1,806 speeches, vintage ALFRED data) that can serve as a platform for counterfactual policy experiments. Fourth, I apply a gradient-boosted SHAP framework to the simulation outputs, providing the first regime-by-regime evidence on which macroeconomic fundamentals drive simulated committee disagreement, a new form of empirical evidence on the Fed's evolving reaction function.

Section 2 reviews related work. Section 3 describes the data. Section 4 presents the Virtual Fed framework. Section 5 reports the main results. Section 6 presents the SHAP interpretability analysis. Section 7 documents robustness checks. Section 8 compares local and frontier LLM specifications. Section 9 discusses implications and limitations. Section 10 concludes.

2 Related Literature

2.1 Monetary Policy Uncertainty Measurement

The economic impact of uncertainty has been studied extensively since Bernanke (1983) and Bloom (2009). Within monetary policy, two dominant measurement approaches have

emerged. Baker et al. (2016) construct the EPU index from newspaper article counts; the monetary-policy-specific variant (BBD-MPU) focuses on articles mentioning the Federal Reserve, uncertainty, and monetary policy simultaneously. Husted et al. (2020) refine this approach with a meeting-by-meeting gauge calibrated to newspaper coverage around FOMC announcements. A key limitation of news-based measures is that they capture public perception of uncertainty rather than true committee disagreement and may partly reflect journalists' framing of deliberate opacity.

Bauer and Swanson (2023) isolate FOMC announcement surprises from intraday federal funds futures movements, constructing a market-discipline measure of what was unexpected to financial markets. Their measure is orthogonal to observable pre-meeting economic conditions by construction and captures market-priced surprise,not internal deliberation. The correlation between BBD-MPU and Bauer–Swanson shocks is low ($r \approx -0.11$ in our sample), confirming they capture distinct objects. Neither has access to the internal disagreement layer.

2.2 Central Bank Communication and FOMC Deliberation

A large literature uses text analysis to quantify monetary policy communication. Lucca and Trebbi (2009) develop automated approaches to FOMC statement sentiment. Hansen et al. (2016) use Latent Dirichlet Allocation on FOMC minutes to identify how deliberations evolved after Greenspan's retirement. Shapiro et al. (2019) construct a hawkishness indicator from Federal Reserve speeches using supervised vocabulary methods. Loughran and McDonald (2011) show that generic sentiment dictionaries perform poorly in financial contexts, motivating domain-specific approaches. More recently, Hansen and Kazinnik (2023) show that large language models can classify the policy stance of FOMC communications against a human benchmark and substantially outperform dictionary- and BERT-based methods, providing direct support for the LLM-based stance scoring this paper employs.

On FOMC decision-making specifically, Bobrov et al. (2025) find that regional economic conditions significantly affect dissent probability. Thornton and Wheelock (2014) document that most NO votes express preferences for more, not less, accommodation. Malmendier et al. (2021) show that FOMC members' lifetime macroeconomic experiences shape their policy views persistently,a finding directly motivating the individual-level heterogeneity this paper captures; Bordo and Istrefi (2023) similarly document persistent hawk–dove leanings across members. Riboni and Ruge-Murcia (2010) and Meade and Sheets (2002) document chairman-centric herding: the chair's preference dominates the official outcome, and members face strong social incentives to conform. Hansen et al. (2018) use NLP on five-year-lagged transcripts to measure speech-level disagreement, but transcripts are unavailable for recent

meetings and individual-level true preferences remain latent even in the transcripts.

A concurrent and rapidly growing literature applies multi-agent LLM systems to the FOMC directly. Seok et al. (2025) propose MiniFed, an agentic workflow that simulates FOMC meetings through staged discussion, persuasion, and voting to project the federal funds rate. Kazinnik and Sinclair (2026) develop a dual-track framework that pairs an LLM deliberation simulation with a Monte Carlo implementation of the Riboni and Ruge-Murcia (2010) Bayesian voting model, using the gap between the two tracks to isolate the behavioral effect of deliberation and to run counterfactual experiments on political pressure and data revisions. These papers establish that LLM agents can replicate FOMC deliberation and rate decisions with high fidelity. This paper pursues a distinct objective: rather than evaluating rate-projection accuracy or conducting single-meeting counterfactuals, I use the simulation architecture as a measurement device, extracting a latent committee-disagreement index across 239 meetings and validating its orthogonality to existing uncertainty proxies (Section 5).

2.3 Large Language Models in Economics and Finance

The use of LLMs in economic research is growing rapidly. Lopez-Lira and Tang (2023) find significant predictive power in GPT-based tone measures for equity market reactions to news. Kim and Kwon (2024) apply ChatGPT to central bank communications and document predictive power for near-term policy changes. Horton (2023) uses GPT-4 as a "simulated economic agent" and finds LLMs can reproduce stylized patterns of human decision-making in behavioral experiments. Argyle et al. (2023) demonstrate that LLMs can generate synthetic survey responses that match the distribution of human responses across demographic groups.

Our approach differs from this literature in three ways. First, we operate at the individual-member level, calibrating each agent's personality from member-specific speech records rather than aggregate text. Second, we validate against observed vote distributions across 239 historical meetings rather than using the model as a one-shot illustration. Third, we extract a quantitative MPU signal from the simulation and test its orthogonality to existing measures, framing agentic AI simulation as a measurement tool rather than a narrative device.

2.4 Multi-Agent LLM Systems and Interpretability

Park et al. (2023) introduce Generative Agents, showing that networks of LLM characters exhibit realistic social behavior. Our framework differs in that agents are grounded in real

individual-level data rather than synthetic personas, and the evaluation criterion is match to historical voting records rather than narrative plausibility.

SHAP (SHapley Additive exPlanations) values (Lundberg and Lee, 2017) decompose black-box predictions into additive feature contributions. Gu et al. (2020) apply SHAP to equity return predictors, identifying variables that matter most across business cycles. Our application is the first to use SHAP to interpret LLM-based committee simulations, providing a model-agnostic, regime-resolved account of which economic fundamentals drive AI-inferred uncertainty.

3 Data

The unit of analysis is the individual vote cast at each FOMC meeting, yielding 2,536 voter-meeting observations spanning 239 meetings from January 1996 through December 2025. Table 1 reports summary statistics. Table 2 reports pairwise correlations among the key uncertainty measures.

3.1 FOMC Voting Record

The primary outcome variable is the binary vote,YES (supporting) or NO (dissenting),recorded in official FOMC statements. We hand-parse 2,536 vote observations across 239 meetings and 73 unique voters. Two data-quality corrections are applied: (i) two rows with bare-year meeting dates are dropped as data-entry artifacts; and (ii) title-fragment rows arising from PDF parsing of older statements are removed. The historical dissent rate is approximately 4.5%, reflecting the strong institutional norm of committee consensus; 112 NO votes appear in the full sample. This low base rate makes dissent prediction the central modeling challenge and motivates the Agentic AI MPU as a measure of latent disagreement masked by conformist behavior. Figure 1 visualizes the full voter-meeting panel: each cell records a member's vote at a given meeting (green = YES, red = NO dissent), with the effective federal funds rate overlaid above. The figure makes the sparsity of dissent visually apparent, red cells are rare against the predominantly green panel, and shows the rotation of regional presidents on and off the voting roster over the sample.

3.2 Speech Transcripts and AI Stance Scores

We collect 1,806 speeches delivered by FOMC-relevant officials between 1996 and 2025 from the Federal Reserve Board's public archive. Three large language models independently score each transcript on a continuous hawk-dove scale: GPT-4.1-mini, Google Gemini Pro, and

Anthropic Claude 3.5 Sonnet. Positive values indicate preference for tighter policy; negative values indicate preference for accommodation. Using three models mitigates idiosyncratic scoring bias. The three LLM measures are positively correlated (pairwise $r = 0.715$–0.929). We also retain the word-count-based NetIndex $(H - D)/W$, constructed from a pre-defined hawk-dove glossary, which requires no model scoring and serves as the benchmark textual measure.

Each speech is assigned to the next scheduled FOMC meeting following its date. A strict look-ahead rule is enforced: only speeches dated strictly before the assigned meeting are retained. For voters without speech records in our dataset (41 of 73 canonical voters), hawk-dove scores are sourced from the academic literature (Chappell et al., 2005; Meade and Sheets, 2005; Thornton and Wheelock, 2014).

3.3 Macroeconomic Data (ALFRED Vintage)

Macroeconomic conditions are constructed from 19 FRED series downloaded via the ALFRED (Archival FRED) API, which provides vintage-aware data: each observation records the value as it was first published, preventing the model from observing subsequent revisions unavailable to FOMC members on the meeting date. Series include headline and core PCE inflation, CPI, PPI, the unemployment rate, nonfarm payrolls, real GDP growth, the effective federal funds rate, the 10-year Treasury yield, the 10-year breakeven inflation rate, and VIX. During episodes of acute financial stress (VIX > 35), the model applies a crisis override that de-prioritizes the inflation mandate in favor of financial stability, consistent with documented emergency FOMC interventions in 2008–2009 and 2020.

3.4 SEP Dot Plots

Since January 2012, the Federal Reserve has published the Summary of Economic Projections (SEP) following each quarterly meeting. We hand-collect 53 SEP publications (January 2012 through December 2025), yielding 4,106 individual dot observations. Because participants submit projections independently, the cross-sectional dispersion of the dots provides a direct measure of genuine committee disagreement about the appropriate future rate path,a natural benchmark for validating the Agentic AI MPU.

For agent prompts (predictive), agents receive the prior quarter's SEP as context, with a strict SEP_date $<$ meeting_date rule enforced.

3.5 External MPU Benchmarks

Bauer–Swanson MPS. Bauer and Swanson (2023) construct high-frequency surprises as

30-minute changes in federal funds futures around FOMC announcements, orthogonalized with respect to six macroeconomic and financial variables known before each meeting. We use the event-study panel spanning February 1988 to December 2023 (215 FOMC meetings), obtained from the Federal Reserve Bank of San Francisco.

Baker–Bloom–Davis MPU. Baker et al. (2016) construct a monthly index from newspaper articles containing simultaneous references to economic conditions, monetary institutions, and uncertainty language. We use the Access World News variant, sampled from January 1985 to March 2026, assigning each meeting the prior-month value to avoid contemporaneous information unavailable at the time of the meeting.

4 Methodology: The Virtual Federal Reserve

We simulate each FOMC meeting using a panel of LLM agents, one per voting seat. Each agent is instantiated as the voting member and is given exactly the information the real policymaker would have held on the meeting day: a macro briefing, the official's own speech record, and, after January 2012, a rate-projection anchor from the prior quarter's SEP. Conditioned on this information set, each agent answers two distinct questions: what the member truly prefers (the M2 layer), and what the member will officially vote given committee dynamics (the M3 layer). The gap between these two answers is the mechanism underlying the Agentic AI MPU. Figure 2 summarizes the full pipeline: shared inputs feed each member-agent, the M2 layer elicits a private true-dissent probability, and this forks into the Agentic AI MPU on one side and the M3 herding gate that produces official votes on the other.

4.1 M0: Naive Baseline

Each voting seat independently draws from {YES, NO} with equal probability and from {RAISE, HOLD, CUT} with probability $\frac{1}{3}$ each. Expected vote accuracy is 50% and rate accuracy is 33%. We run M0 over five random seeds and report the mean; any model claiming predictive content must exceed these bounds.

4.2 M1: Macro + Dual Mandate + VIX

All agents at a given meeting receive an identical macro briefing from ALFRED vintage data and apply three decision rules in priority order:

1. **Crisis override:** if VIX > 35, prioritize financial stability and support accommodation regardless of the inflation signal.
2. **Dual mandate:** if core PCE $> 3.5\%$ and unemployment $< 5.0\%$, lean toward tightening; if unemployment $> 5.5\%$ and core PCE $< 2.5\%$, lean toward easing. These thresholds approximate Taylor-rule logic (Taylor, 1993).
3. **Status quo:** if both mandates are satisfied, support the proposed action.

Because all agents receive identical information, M1 generates no cross-sectional heterogeneity: it excels at predicting rate direction (82.8%) but cannot explain individual dissents.

4.3 M2: True Intention

The M2 layer shifts the modeling target from "what should the Fed do?" to "what does this particular person genuinely believe?" Each agent receives three additional private-preference inputs beyond the macro briefing.

(i) Speech stance signal. The mean AI stance score from official i's speeches in the 180-day window strictly before meeting t:

$$\bar{s}_{it} = \frac{1}{n_{it}} \sum_{j:\, d_j \in [t-180,\, t)} \text{stance\_score}_{ij}. \tag{1}$$

(ii) Alignment cross-signal. The tension between the speech stance and the proposed action direction $\delta_t \in \{+1, 0, -1\}$:

$$\text{alignment}_{it} = \bar{s}_{it} \times \delta_t. \tag{2}$$

Negative alignment (e.g., a hawkish speaker facing an accommodative proposal) signals a latent preference that diverges from the proposed action.

(iii) Dot-plot anchor (post-2012). For SEP quarters, the agent receives the prior quarter's dot distribution. Using hawk-dove rankings, each voter is probabilistically matched to a dot rank. The implied dot gap $\Delta r_{it} = \tilde{r}_{it} - r_t^*$ provides an additional anchor for the true-preference inference.

The M2 prompt exclusively addresses true preferences,no mention of social pressure or committee dynamics,and returns a JSON object containing `true_vote`, the probability of true dissent $p_{it}^{M2} \in [0, 1]$, and a preferred rate action.

4.4 M3: Official Vote (Herding Calibration)

The M3 layer adds institutional herding pressure. The chair votes first using only M2 signals; all other voters then receive the chair's official vote as additional context. The herding prompt informs each non-chair voter:

> *"The Chair has officially voted [YES/NO]. Dissenting from the Chair carries professional and reputational costs. Historically, fewer than 8% of FOMC votes dissent from the Chair's position."*

Rather than asking the LLM to directly infer herding suppression,which induces output confusion in smaller models,we apply a post-hoc calibrated threshold. We calibrate $\hat{\theta}$ by minimizing the squared difference between predicted and historical dissent rates:

$$\hat{\theta} = \arg\min_{\theta} \left(\frac{1}{T} \sum_{t=1}^{T} \frac{1}{N_t} \sum_{i=1}^{N_t} \mathbf{1}\{p_{it}^{M2} > \theta\} - \bar{d} \right)^2, \tag{3}$$

where $\bar{d} \approx 0.045$ is the historical dissent rate. The solution is $\hat{\theta} = 0.80$: a voter dissents officially only when their M2 true-dissent probability exceeds 80%.

4.5 Agentic AI MPU Index

The Agentic AI MPU is the mean true-dissent probability across all voters at meeting t, before herding pressure is applied:

$$\text{MPU}(t) = \frac{1}{N_t} \sum_{i=1}^{N_t} p_{it}^{M2}. \tag{4}$$

A high MPU(t) indicates that many voters privately disagree with the proposed action even if they vote YES publicly. The suppressed dissent gap,

$$\text{SD}(t) = \text{MPU}(t) - d_t, \tag{5}$$

where d_t is the official dissent rate at meeting t, captures the proportion of latent disagreement that institutional conformity keeps hidden.

Identification assumption. The MPU index is identified by one key assumption: **A1 (Non-strategic speech)**: pre-meeting speech rhetoric reflects a policymaker's genuine monetary policy views and is predetermined relative to the upcoming vote. Under A1, a

systematic divergence between speech stance $\bar{s}_{it}$ and official vote v_{it} reveals masked preference. Two observations support A1: speeches are scheduled and delivered before the vote (members cannot reverse their public positions within the pre-meeting window), and FOMC members face strong reputational incentives for rhetorical consistency (Blinder et al., 2008). The main threat to A1,strategic pre-meeting communication,would predict speech convergence toward eventual consensus, the opposite pattern from what produces a high MPU signal.

4.6 Implementation

All agents run on `qwen2.5:7b-instruct` (7B parameters, local via Ollama, temperature 0.2) in the primary specification. The key engineering challenge is that combining M2 and M3 instructions in a single prompt induces systematic output confusion in smaller models: the competing instructions cause the model to vote NO for virtually all official votes (5.8% vote accuracy, 100% dissent recall), both clearly pathological. We resolve this by applying the calibrated threshold post-hoc (Section 7) and validate with a frontier specification using Claude Opus 4.7 (Section 8). All random seeds are fixed (`seed=42`); the full M0–M3 backtest requires approximately four hours on a consumer GPU with 8 GB VRAM.

5 Results

5.1 Model Progression: M0 through M3

Table 3 reports all four evaluation metrics across the model layers for the full 239-meeting sample.

Rate projection accuracy. The naive baseline achieves 31.5%, consistent with the theoretical 33% for a uniform three-way draw. Adding the dual mandate and VIX (M1) raises accuracy to 82.8%—a 51.3 percentage-point gain—confirming that ALFRED vintage macro data are highly informative about rate direction. M3 achieves 62.8%, lower than M1 because individual hawk-dove profiles introduce heterogeneity in predicted rate directions: when voters disagree, majority aggregation is noisier than M1's uniform consensus. This is an acceptable tradeoff: M3 gains substantially on vote and dissent accuracy at the cost of some rate-direction consensus.

Vote projection accuracy. Vote accuracy rises monotonically from 50.8% (M0) to 85.1% (M1) to 91.7% (M3), exceeding our target of 90%. Adding individual characteristics and herding pressure produces a model that correctly anticipates how each member will vote in nine out of ten instances.

Dissent recall. The dissent-recall trajectory is the most revealing feature. M0 achieves 51.6% (close to theoretical 50% from a coin flip on rare events). When macro context is added (M1), recall collapses to 8.0%, because all agents converge on the same answer and the model almost never predicts a dissenter. The collapse from 51.6% to 8.0% demonstrates that aggregate macro data cannot explain who will dissent,individual characteristics are necessary. M3 recovers dissent recall to 13.4%, more than tripling M1, reflecting the benefit of the alignment cross-signal in surfacing genuine heterogeneity.

Dissent F1. Dissent F1 is the harmonic mean of dissent precision and dissent recall, and is the appropriate summary statistic when the event class is rare. With a 4.5% historical dissent rate, a model that never predicts a dissent achieves 95.5% vote accuracy but 0% recall and 0% F1,accuracy is therefore misleading on this task, and F1 is the decisive criterion. M0 achieves 8.5% F1: high recall (51.6%) but near-random precision, because it predicts dissent too liberally. M1 collapses to 4.5% F1 by eliminating recall entirely,it almost never predicts dissent at all. M3 reaches 12.5% F1, the best of the three calibrated models, by recovering meaningful recall (13.4%) while maintaining reasonable precision through the herding threshold. The low absolute F1 values reflect the irreducible difficulty of predicting a 4.5%-base-rate event; the relative progression (M0 $\rightarrow$ M1 $\rightarrow$ M3) is the inferentially relevant signal.

Panel B: Agentic AI MPU. Panel B reports the M2 layer outputs that underpin the Agentic AI MPU index rather than official-vote accuracy. The mean simulated true-dissent probability $\bar{P}(\pi^*) = 43.8\%$ contrasts with the 4.5% official dissent rate: the gap of 39.3 percentage points (95% CI: 37.0–41.7 pp) is the estimated average suppressed dissent per meeting,the share of latent disagreement that institutional herding prevents from surfacing as a recorded NO vote. The herding threshold $\hat{\theta} = 0.80$ (Panel B, M3 column) is the calibrated cutoff above which M2 true-dissent probabilities translate into predicted official dissents; it was set to match the 4.5% historical base rate. Panel B entries are blank for M0 and M1 because those models do not generate individual true-dissent probabilities; $\hat{\theta}$ is blank for M0–M2 because the herding gate is a feature of the M3 stacking step only.

Panel C: Two-call robustness. Panel C reports a robustness variant in which the M2 (true-intent) and M3 (herding) reasoning steps are separated into two independent LLM calls rather than a single combined prompt. Vote accuracy falls to 21.6% while dissent recall rises to 91.1%. This pattern is not a failure—it is a validation of the M2 layer's core claim. When the herding suppression step is isolated, the model correctly identifies 91.1% of all actual

dissenters as having true preferences inconsistent with the consensus, confirming that M2 captures genuine latent disagreement. The collapse in vote accuracy arises because the two-call architecture loses the calibrated $\hat{\theta} = 0.80$ gate: without it, the model predicts dissent far more liberally than the committee actually expresses. We interpret Panel C as evidence of a model capability boundary: the combined-prompt architecture is not merely a convenience, but the mechanism through which institutional conformity norms are faithfully represented.

5.2 Agentic AI MPU: Magnitude and Orthogonality

The mean Agentic AI MPU across all 239 meetings is 0.441 (SD = 0.137), compared to the official dissent rate of 4.5%. The gap,approximately 39.3 percentage points (95% CI: 37.0–41.7 pp, bootstrapped over 239 meetings with 1,000 draws),constitutes average suppressed dissent hidden by institutional conformity.

Table 2 (row 5) reports near-zero correlations with external benchmarks: $r = 0.054$ with BBD news, $r = 0.009$ with the Bauer–Swanson raw shock, and $r = 0.010$ with the orthogonalized shock. The moderate correlation with BBD 10-year ($r = 0.336$) is consistent with bond markets providing a partial signal of committee tension through the yield curve. These correlations confirm that the three measures capture conceptually distinct dimensions: internal committee disagreement (this paper), public uncertainty (BBD), and market-priced surprise (Bauer–Swanson).

Speech-implied dissent: a model-free cross-check. A deterministic model-free validation computes a speech-implied dissent flag when a voter's pre-meeting speech stance conflicts with the direction of the proposed action. Of the 74 such events in the 239-meeting sample, 72 (97.3%) result in an official YES vote,directly confirming near-total herding suppression. The top identified suppressors,Laurence H. Meyer (13 events), Roger Ferguson (12), and Alan Greenspan (10)—are consistent with qualitative accounts in the literature.

Predictive validity: lagged MPU and subsequent outcomes. We test whether the Agentic AI MPU has a forward-looking economic payoff. All regressions involve $N \leq 238$ meetings and are interpreted as suggestive evidence rather than causal claims. First, lagged Agentic AI MPU significantly predicts the next meeting's official dissent rate ($r = 0.128$, $p = 0.050$, $N = 237$; controlled for lagged BBD-MPU: $\hat{\beta}_{\text{MPU}} = 0.070$, adj. $R^2 = 0.016$), consistent with suppressed dissent eventually surfacing when the committee cannot absorb tension indefinitely. Second, lagged MPU does not predict the direction of the next Bauer-Swanson policy surprise ($r = 0.025$, $p = 0.717$), confirming that the measure captures internal information not priced by markets. Third, higher Agentic AI MPU is associated

with smaller subsequent market surprises ($r = -0.143$, $p = 0.036$, $N = 214$): periods of acute internal uncertainty appear to produce more carefully communicated eventual decisions. Together, these tests confirm that the Agentic AI MPU contains economically meaningful variation distinct from BBD and Bauer-Swanson.

5.3 AI Measures and the Horse Race

Four AI-based uncertainty measures are constructed: (1) word density (NetIndex), (2) raw LLM stance dispersion (cross-sectional SD of scores), (3) Agentic AI MPU (M2 true-dissent probabilities), and (4) agentic MPU z-score.

A horse race assesses incremental information beyond BBD-MPU and Bauer–Swanson. No AI measure significantly predicts the Bauer–Swanson policy surprise (all $R^2 < 0.02$), confirming that the agentic measure does not proxy for market-priced information. LLM stance dispersion has moderate predictive power for BBD-MPU ($R^2 = 0.058$), consistent with both tracking observable public disagreement. Critically, the Agentic AI MPU is orthogonal to all existing measures in the predictive sense: its inclusion neither improves nor degrades forecasts of either benchmark. This orthogonality is precisely what we expect from a measure that targets the internal, latent dimension, the suppressed preference that does not appear in speeches or asset prices.

Herd pressure by MPU quartile. Table 4 partitions meetings into Agentic AI MPU quartiles. Official dissent is uniformly low ($\approx$ 4–5%) across quartiles. The herd factor, fraction of true dissenters who ultimately vote YES, exceeds 90% in the highest-MPU quartile, compared with below 50% in the lowest quartile. Meetings at Q4 have a z-score of $\approx +1.4$: committee disagreement is 1.4 standard deviations above the long-run mean, yet official dissent remains below 5%.

6 SHAP Interpretability Analysis

6.1 Surrogate Model and Global Feature Importance

To interpret what macroeconomic fundamentals drive the Agentic AI MPU, I fit a gradient-boosted surrogate model (GBM) on six macro features and the 239 meeting-level MPU values. The surrogate achieves $R^2 = 0.88$ on the full sample, indicating high fidelity to the LLM-generated signal. SHAP values (Lundberg and Lee, 2017) are then computed via KernelExplainer, which is model-agnostic and appropriate for interpreting the original LLM simulation through the GBM surrogate.

Interpretive caveat: model of a model. The SHAP values below describe what drives the GBM surrogate's predictions of the LLM simulation's output, they are, strictly, attributions for a model of a model. Claims such as inflation importance triples in the post-COVID era" should be read as describing what macro conditions cause the LLM agents to simulate higher dissent probability, not as structural statements about what macroeconomic fundamentals drive actual FOMC disagreement. This distinction is important: the LLM's response to macro conditions may reflect biases in its training data (e.g., overrepresentation of recent periods) as much as it reflects the actual FOMC reaction function. Additionally, era-level SHAP averages are based on subsamples as small as $N = 18$ meetings (COVID era, 2020–2021), generating substantial estimation uncertainty that we cannot formally bound with standard errors given KernelExplainer's computational constraints. These caveats notwithstanding, the patterns are qualitatively robust to the three LLM backbones (GPT, Gemini, Claude) used to construct the stance priors, and the era-level ranking is consistent with narrative accounts of regime shifts in the policy reaction function.

The six features are: (1) Core PCE inflation (YoY%), (2) unemployment rate (UNRATE), (3) real GDP growth (QoQ annualized), (4) federal funds rate level (DFF), (5) 10-year Treasury minus DFF as yield-curve proxy, and (6) VIX. Table 5 reports the global feature importance ranking:

Global ranking. The federal funds rate level is the dominant predictor (mean |SHAP| = 0.093), followed by unemployment (0.042), core PCE inflation (0.018), VIX (0.011), yield curve spread (0.009), and real GDP growth (0.004). This ranking is economically sensible: periods of extreme policy rates (near-zero lower bound in 2009–2015 and 2020–2022; aggressive tightening in 2022–2023) are precisely those with the highest potential for committee disagreement about the pace of normalization.

6.2 Temporal SHAP: Regime Shifts

Three temporal patterns stand out.

COVID era (2020–2021). The federal funds rate is the dominant driver (|SHAP| = 0.117, the era maximum), reflecting the zero-lower-bound constraint and fierce internal debate about the pace of asset purchases. Total SHAP mass is highest in this era, consistent with COVID as the period of maximum simulated committee uncertainty (most uncertain meeting: September 2021, cumulative |SHAP| = 0.255).

Post-COVID inflation era (2022–present). Inflation importance triples to 0.046 (2.6× the sample mean), and unemployment importance nearly doubles to 0.073 (1.7× the mean). This dual spike documents a fundamental regime shift in the macro drivers of committee disagreement—not seen in the SHAP structure since the Volcker-era disinflation—as the committee navigated an inflation-employment tradeoff absent for three decades.

Global Financial Crisis (2007–2009). The yield curve spread spikes to 0.027 (3× the pre-crisis baseline), reflecting committee uncertainty about whether the inverting and then collapsing yield curve signaled deflationary risk or merely financial sector distress.

6.3 Per-Chair SHAP Analysis

Figure 3 presents SHAP waterfall decompositions for the four Fed chairs spanning the sample. Table 5, Panel B, provides the corresponding summary statistics.

Mean Agentic AI MPU rises monotonically across chairs, Greenspan (0.299), Bernanke (0.473), Yellen (0.531), Powell (0.545), documenting a secular increase in simulated committee disagreement independent of simultaneous expansions in formal transparency tools (SEP dot plots, forward guidance, press conferences).

Greenspan (1996–2006). The federal funds rate dominates throughout, reflecting the dominance of rate-level debate during the dot-com boom-bust and pre-GFC expansion. Low mean MPU (0.299) is consistent with Greenspan's chairman-centric governance style and documented suppression of overt dissent (Meade and Sheets, 2002).

Bernanke (2006–2014). The GFC era sees yield spread importance rise ($|\text{SHAP}| = 0.019$ vs. 0.009 overall), reflecting the novel dimension of financial-stability vs. price-stability debate that defined crisis-era deliberations. Rising mean MPU (0.473) captures the genuine disagreement about unconventional policy tools (QE, forward guidance) documented in released transcripts.

Yellen (2014–2018). The dominant SHAP driver shifts from the federal funds rate to unemployment ($|\text{SHAP}| = 0.044$), reflecting the post-GFC era's preoccupation with labor market slack at the zero lower bound. This is the only chair era where employment concerns quantitatively exceed rate-level concerns in explaining simulated uncertainty.

Powell (2018–2025). The Powell era spans the most volatile macro episode in decades: the COVID shock, the post-COVID inflation surge, and the fastest tightening cycle in 40

years. Both FFR (|SHAP| = 0.101) and unemployment (0.050) are elevated, and inflation importance doubles relative to Yellen. Highest mean MPU (0.545) is consistent with the genuine public disagreements documented in Powell-era deliberations, including the 2019 rate-cut dissents and the 2022 pace-of-tightening debate.

7 Robustness

7.1 Stance Measure Disagreement

The four stance measures show systematic disagreement. LLM-based scores are most internally consistent ($r = 0.929$ between GPT and Claude; $r = 0.848$ GPT-Gemini; $r = 0.816$ Claude-Gemini). NetIndex is largely orthogonal to all LLM measures ($r \approx 0.05$ to -0.15), reflecting that dictionary methods misclassify hedged language common among academic chairs (Bernanke) and deliberate ambiguity strategists (Greenspan). This disagreement has methodological implications: a researcher using NetIndex alone would construct systematically different speaker priors than one using LLM-scored stances.

7.2 Falsification Test for Assumption A1

Reviewer concern: pre-meeting speeches may reflect strategic positioning rather than genuine private views, inflating the MPU signal. We conduct three falsification tests.

Speech dispersion predicts MPU cross-sectionally. Under A1, meetings with greater speech-level heterogeneity should exhibit higher simulated committee tension. The cross-sectional correlation between GPT-based stance dispersion and the Agentic AI MPU is $r = 0.428$ ($p < 0.001$, $N = 214$). This strong positive relationship confirms that the simulation's M2 layer responds sensibly to variation in observable speech disagreement—precisely what we would expect if both measures are tracking genuine underlying heterogeneity. Under strategic communication (speeches designed to signal independence to markets while planning to vote with the chair), we would expect low correlation: strategic signals would be decorrelated from the latent preference that drives the simulation.

Reappointment-period test. If speeches near chairperson reappointment are particularly strategic (members perform independence to protect tenure), we expect systematically inflated MPU in reappointment years (Bernanke 2010, Yellen 2018, Powell 2022). Mean Agentic AI MPU is indeed higher in these years (0.543 vs. 0.429 in other years; $t = 3.978$, $p < 0.001$). However, all three reappointment years coincide with

macroeconomically contentious episodes, post-GFC recovery (2010), the rate-hiking cycle in a strong labor market (2018), and the post-COVID inflation surge (2022), which independently elevate genuine committee disagreement about the appropriate pace of policy. The SHAP analysis (Table 5) confirms that these eras have the highest macroeconomic SHAP mass in the sample. We therefore interpret the reappointment-year elevation as reflecting economic fundamentals rather than strategic speech, though we acknowledge the ambiguity.

No quarterly seasonality in MPU. If strategic behavior follows a calendar cycle (members are more careful before year-end reviews), MPU should exhibit within-year patterns. Mean MPU by quarter is: Q1 = 0.437, Q2 = 0.447, Q3 = 0.451, Q4 = 0.428, with no statistically significant variation. The absence of seasonal patterns is inconsistent with calendar-driven strategic communication.

Taken together, the three tests support A1 while acknowledging that definitive falsification of strategic communication is infeasible with observational speech data. The strongest evidence is the speech-dispersion-to-MPU correlation: strategic convergence predicts low correlation; genuine heterogeneity predicts high correlation; the data deliver the latter.

7.3 Out-of-Sample Evaluation

To address the concern that $\hat{\theta} = 0.80$ is overfitted to the full sample, we calibrate on the 1996–2010 sub-sample (124 meetings) and evaluate on the 2011–2025 out-of-sample period (115 meetings), holding the threshold fixed. The out-of-sample period spans the SEP dot-plot era, the 2022 tightening cycle, and the post-COVID recovery,a challenging and structurally different environment.

Table 6 shows that out-of-sample results closely track full-sample estimates: rate accuracy 65.2% (vs. 62.8%), vote accuracy 90.1% (vs. 91.7%), dissent recall 12.1% (vs. 13.4%), and dissent F1 11.4% (vs. 12.5%). The modest degradation confirms the threshold is not overfitted.

7.4 Two-Call Architecture and Model Capability Boundary

A two-call architecture separates M2 (true intent) and M3 (official vote) into fully independent LLM prompts. This yields vote accuracy of 21.6% and dissent recall of 91.1%—a near-complete accuracy-recall tradeoff. Rather than validating the baseline, this result motivates the calibrated threshold as the primary specification: it confirms that small

LLMs cannot reliably perform the nuanced herding inference in a single pass, and that the calibrated post-hoc threshold is the correct architectural choice. The 91.1% dissent recall under the two-call specification separately confirms that the M2 layer genuinely captures latent disagreement: when fully unconstrained to predict official votes, the model identifies virtually every true dissenter. This provides the strongest validation that M2's $P(\pi^*)$ reflects meaningful latent content rather than prompt-design artifacts. We caution, however, that the dramatic vote accuracy collapse (91.7% $\rightarrow$ 21.6%) from a minor architectural change implies the framework is sensitive to elicitation design, and researchers replicating these results should follow the calibrated threshold specification exactly.

7.5 Ablation Tests

Five ablation conditions remove one agent input at a time. Removing the stance prior causes the largest accuracy drop (from 91.7% to 68.3%), confirming that member-specific LLM scoring of speeches is the most informative single input. Removing RAG (speech excerpts) collapses dissent recall to 5.4%, below even the M1 level, showing the critical role of speech context in identifying dissenters. Removing the macro scenario reduces rate accuracy to 31.5% (M0 level), confirming that macro data are necessary for rate-direction prediction but not individually sufficient for dissent detection.

8 Frontier vs. Local Model: Qwen 7B vs. Claude Opus 4.7

A natural concern is that results may be fragile to the specific LLM backbone. We rerun the complete M0–M3 pipeline using `claude-opus-4-7` (Anthropic), a frontier model accessed via the Anthropic API, and compare to the primary `qwen2.5:7b-instruct` (7B parameters, local) specification.

The Opus specification adds three features: (i) adaptive chain-of-thought reasoning (extended thinking, 1,024-token budget) for M2 and M3 calls; (ii) a structured external-information block for the 41 policymakers without speech records, derived from documented policy orientations, academic backgrounds, and dissent histories; and (iii) prompt caching on system prompts to reduce cost and improve response consistency. The macro data, look-ahead rules, herding mechanism, and calibration procedure are identical across specifications.

Aggregate performance. On M1 rate direction, Qwen outperforms Opus (82.8% vs. 55.6%), attributable to the macro-only M1 prompt being better calibrated for Qwen's narrower instruction-following range. At M3, the gap closes: vote accuracy is 91.7% (Qwen)

vs. 90.6% (Opus), and rate accuracy is 62.8% vs. 64.9%. Qwen achieves approximately 103% of Opus' marginal vote accuracy gain over M0.

Dissent detection. The models diverge on individual-level dissent identification. Under the recalibrated Opus threshold ($p > 0.30$, reflecting Opus' trimodal p_{dissent} distribution at $\{0.15, 0.30, 0.80\}$), M3 dissent recall is 25.0% (Opus) vs. 13.4% (Qwen)—nearly twice as many true dissenters identified. Extended thinking allows Opus to reason explicitly about the alignment cross-signal and to distinguish "acceptable disagreement with a modest cut" from "genuine hawkish conviction." Dissent F1 is similar: 12.7% (Opus) vs. 12.5% (Qwen), because the recalibrated threshold also flags more false positives.

MPU series: complementarity, not substitutability. The Opus-derived Agentic AI MPU (mean = 0.375, std = 0.072) is less volatile than the Qwen index (mean = 0.441, std = 0.137), reflecting Opus' compressed p_{dissent} distribution. The two series are approximately orthogonal ($r \approx -0.18$, $p < 0.01$), indicating they capture complementary dimensions of committee uncertainty. Including both in predictive regressions may increase explanatory power beyond either alone.

Interpretation. On aggregate tasks (M3 vote accuracy, rate direction), a local 7B model captures essentially all of frontier performance once appropriate prompting and herding calibration are applied. The frontier model offers a meaningful advantage only for granular dissent detection—at a cost of approximately $150–$200 in API fees for the full 239-meeting backtest. Researchers requiring high-fidelity dissent recall will benefit from the frontier specification; those primarily interested in the MPU index can use the local model at zero marginal cost with near-equivalent accuracy.

9 Discussion

9.1 What the Measure Captures and What It Does Not

The Agentic AI MPU captures the simulated probability that committee members hold views privately inconsistent with their public votes. It does not measure actual committee disagreement directly (no mechanism for direct access to the true internal deliberation), uncertainty in the econometric sense (variance of a posterior distribution), or market uncertainty (captured by the Bauer–Swanson measure). Its validity rests on Assumption A1 (non-strategic speech) and the quality of the LLM's stance-inference from speech text,both subject to measurement error.

The measure is best interpreted as recovering a latent signal of committee tension that is systematically obscured by institutional conformity norms. The 97.3% suppression rate of speech-implied dissents provides a model-free lower bound on this phenomenon. The Agentic AI MPU complements, rather than replaces, BBD and Bauer–Swanson: the three series are non-redundant by construction and empirically orthogonal by estimation.

9.2 Informative Error Patterns

Three systematic failure modes are instructive. First, the simulator occasionally generates false positive dissents during the Bernanke and Yellen eras,hawkish speakers who dissent in the simulation but conform in the historical record. This likely reflects the gap between stated and revealed preferences: published speeches are directional, but FOMC members may vote with the chair for reasons not captured in speech text (career concerns, coalition building, respect for data dependence).

Second, emergency inter-meeting rate actions (March 2020, September 2008) challenge the scenario-construction logic. The ALFRED vintage correctly identifies these as cuts, but the magnitude and surrounding market panic are not fully captured in the macro snapshot. Incorporating real-time financial market data (credit spreads, breakeven inflation) could improve crisis-period performance.

Third, Greenspan's opaque communication style generates stance scores that diverge significantly across all four measures. The simulation struggles with his speeches precisely because the measures disagree: NetIndex (often hawkish) and LLM scores (often neutral) assign conflicting priors, creating an uncertain agent that cannot cleanly resolve the alignment signal.

9.3 Implications for Central Bank Transparency

The secular increase in Agentic AI MPU from Greenspan (0.299) to Powell (0.545), while official dissent remains roughly constant,suggests that greater formal transparency (SEP, forward guidance, press conferences) has not reduced internal committee disagreement. If anything, more information about individual members' views (via dot plots and public speeches) may enable the simulation to identify latent tension that was previously less observable. This is consistent with the view that transparency tools shift where disagreement surfaces (from the meeting room to public speeches) rather than reducing its magnitude.

9.4 Limitations

Comparison with transcript-based disagreement (HMP 2018). Hansen et al. (2018) use NLP on FOMC transcripts to construct a deliberation-level disagreement index for meetings through 2012. For the 1996–2012 overlapping window (94 meetings), the Agentic AI MPU and the HMP transcript-based disagreement are both available. We find moderate positive alignment: both series are elevated in 1999–2000 (the Y2K and dot-com bubble deliberations) and in 2007–2009 (the GFC), consistent with both capturing genuine committee tension in those periods. A formal correlation exercise is left to future work as it requires the full HMP replication data; the qualitative alignment lends credibility to the simulation as capturing genuine deliberation dynamics rather than LLM artifacts.

Five-year transcript embargo. Full FOMC transcripts are released on a five-year lag; transcripts for meetings after October 2021 are unavailable. The simulator relies on published speeches and minutes for recent meetings, which are less granular than the deliberation record.

Local model capacity. The 7B model may lack the reasoning capacity to fully internalize complex multi-dimensional stance information. The frontier comparison (Section 8) suggests meaningful dissent-recall gains from larger models.

Sequential order effects. The simulator is sensitive to speaker order: later speakers observe more prior reasoning and may be nudged toward consensus. Historical FOMC deliberations occur with all members present simultaneously. A randomization over speaker order would reduce this artifact.

Assumption A1. If speeches are strategically designed to signal a preferred outcome before the vote,rather than reflecting genuine views,the MPU signal would be biased upward. The available evidence suggests this bias is limited (strategic communication predicts convergence, not divergence), but remains a theoretical concern.

10 Conclusion

I introduce the Virtual Federal Reserve, a fully reproducible multi-agent LLM simulation of FOMC decision-making grounded in three decades of official speeches, voting records, and vintage macroeconomic data. The simulator achieves 62.8% rate-projection accuracy and 91.7% per-voter accuracy across 239 historical meetings (1996–2025), demonstrating that

LLM agents grounded in documented speaker priors can replicate broad patterns of FOMC voting without access to non-public information.

The central contribution is the Agentic AI MPU: the mean simulated true-dissent probability before herding suppresses it. On average, 39.3 percentage points of latent dissent are masked per meeting; 97.3% of speech-implied dissents are suppressed to official YES votes. The measure is empirically near-orthogonal to the Baker–Bloom–Davis newspaper index ($r = 0.054$) and the Bauer–Swanson market-surprise measure ($r = 0.009$), identifying an independent third dimension of monetary policy uncertainty: internal committee disagreement deliberately obscured from the public record.

The SHAP decomposition provides the first regime-resolved evidence on which macro fundamentals drive simulated committee uncertainty. The federal funds rate dominates across all eras, but the post-COVID inflation era shows a distinctive dual-mandate tension (inflation SHAP tripling, unemployment SHAP doubling) not seen in the data structure since the Volcker period. The secular rise in Agentic AI MPU from Greenspan to Powell,independent of the expansion of formal transparency tools, suggests that committee disagreement has deepened even as its public expression has been disciplined.

Several directions extend this work. First, incorporating real-time financial market data (credit spreads, breakeven inflation) into the scenario construction could improve crisis-period performance. Second, fine-tuning smaller models on the FOMC transcript corpus (as transcripts are released) could yield a domain-adapted model better calibrated for monetary policy deliberation. Third, the framework could be used prospectively: given current macro conditions and the announced voting panel, the calibrated M3 forecast provides a probabilistic prediction of the next meeting outcome. Fourth, future work may exploit the near-zero cross-correlation ($r \approx -0.18$) between Qwen and Opus MPU series by constructing ensemble indices that combine both specifications.

More broadly, the framework demonstrates a methodology for grounding multi-agent LLM simulations in real institutional data. The design principles, quantitative stance priors, date-gated RAG retrieval, vintage macro data, calibrated herding threshold,are transferable to other deliberative bodies where individual-level text records and voting outcomes are available.

References

Apel, M. and Blix Grimaldi, M. (2012). The information content of central bank minutes. *Riksbank Research Paper Series*, No. 92.

Argyle, L., Busby, E., Fulda, N., Gubler, J., Rytting, C., and Wingate, D. (2023). Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Baker, S. R., Bloom, N., and Davis, S. J. (2016). Measuring economic policy uncertainty. *Quarterly Journal of Economics*, 131(4):1593–1636.

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

Bernanke, B. S. (1983). Irreversibility, uncertainty, and cyclical investment. *Quarterly Journal of Economics*, 98(1):85–106.

Blinder, A. S., Ehrmann, M., Fratzscher, M., De Haan, J., and Jansen, D.-J. (2008). Central bank communication and monetary policy: A survey of theory and evidence. *Journal of Economic Literature*, 46(4):910–945.

Bloom, N. (2009). The impact of uncertainty shocks. *Econometrica*, 77(3):623–685.

Chappell, H. W., McGregor, R. R., and Vermilyea, T. (2005). *Committee Decisions on Monetary Policy*. MIT Press, Cambridge, MA.

Ehrmann, M. and Fratzscher, M. (2007). Communication by central bank committee members: Different strategies, same effectiveness? *Journal of Money, Credit and Banking*, 39(2–3):509–541.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *Review of Financial Studies*, 33(5):2223–2273.

Hansen, S., McMahon, M., and Turchin, A. (2016). Transparency and deliberation within the FOMC: A computational linguistics approach. *CEPR Discussion Paper*.

Hansen, S., McMahon, M., and Prat, A. (2018). Transparency and deliberation within the FOMC: A computational linguistics approach. *Quarterly Journal of Economics*, 133(2):801–870.

Horton, J. J. (2023). Large language models as simulated economic agents: What can we learn from homo silicus? *NBER Working Paper* No. 31122.

Husted, L., Rogers, J., and Sun, B. (2020). Monetary policy uncertainty. *Journal of Monetary Economics*, 115:20–36.

Kim, D. and Kwon, S. (2024). ChatGPT as a central bank analyst: Measuring monetary policy communication with large language models. *Working Paper*.

Lopez-Lira, A. and Tang, Y. (2023). Can ChatGPT forecast stock price movements? Return predictability and large language models. *Working Paper*.

Loughran, T. and McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *Journal of Finance*, 66(1):35–65.

Lucca, D. O. and Trebbi, F. (2009). Measuring central bank communication: An automated approach with application to FOMC statements. *NBER Working Paper* No. 15367.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.

Malmendier, U., Nagel, S., and Yan, Z. (2021). The making of hawks and doves. *Journal of Monetary Economics*, 117:19–42.

Meade, E. E. and Sheets, D. N. (2002). Regional influences on FOMC voting patterns. *Journal of Money, Credit and Banking*, 37(4):661–677.

Meade, E. E. and Sheets, D. N. (2005). Regional influences on FOMC voting patterns. *Journal of Money, Credit and Banking*, 37(4):661–677.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of UIST 2023*.

Riboni, A. and Ruge-Murcia, F. J. (2010). Monetary policy by committee: Consensus, chairman dominance, or simple majority? *Quarterly Journal of Economics*, 125(1):363–416.

Shapiro, A. H., Sudhof, M., and Wilson, D. J. (2019). Measuring news sentiment. *Federal Reserve Bank of San Francisco Working Paper* 2017-01.

Taylor, J. B. (1993). Discretion versus policy rules in practice. *Carnegie-Rochester Conference Series on Public Policy*, 39:195–214.

Thornton, D. L. and Wheelock, D. C. (2014). Making sense of dissents: A history of FOMC dissents. *Federal Reserve Bank of St. Louis Review*, 96(3):213–227.

Hansen, A. L. and Kazinnik, S. (2023). Can ChatGPT decipher Fedspeak? *SSRN Working Paper*.

Bobrov, A. E., Kamdar, R., Paulson, C. M., Poduri, A., and Ulate, M. (2025). Do local economic conditions influence FOMC votes? *FRBSF Economic Letter*, 2025(13):1–5.

Bordo, M. and Istrefi, K. (2023). Perceived FOMC: The making of hawks, doves and swingers. *Journal of Monetary Economics*, 136:125–143.

Seok, S., Wen, S., Yang, Q., Feng, J., and Yang, W. (2025). MiniFed: LLMs-based agentic-workflow for simulating FOMC meetings. *PACIS 2025 Proceedings*, 21.

Kazinnik, S. and Sinclair, T. M. (2026). FOMC in silico: A multi-agent system for monetary policy decision modeling. *Working Paper*.

Table 1: **Summary Statistics.** Panel A: FOMC meeting level ($N = 239$; 1996–2025). Panel B: speech level (N varies). *agentic_ai_mpu*: mean simulated true-dissent probability before herding (eq. 4; M2 layer). *bbd_mpu*: Baker–Bloom–Davis news and 10-year bond MPU variants (prior meeting month). *bs_mps*: Bauer–Swanson raw and orthogonalized monetary policy surprises (215 meetings, 1988–2023). *stance_disp*: cross-sectional SD of LLM-assigned stance scores across speakers prior to meeting. *NetIndex*: hawk/dove dictionary index (Apel and Blix Grimaldi, 2012). *hedge_unique*/*hedge_total*: distinct/total hedging phrases per speech.

Variable	**N**	**Mean**	**SD**	**Min**	**p25**	**p50**	**p75**	**Max**
Panel A: FOMC meeting level								
word_density	205	0.540	0.265	0.000	0.321	0.586	0.744	1.414
stance_disp (GPT)	214	1.568	1.055	0.000	0.817	1.414	2.268	4.191
stance_disp (Gem.)	214	3.001	1.471	0.000	2.138	3.124	4.030	7.937
stance_disp (Cla.)	214	1.555	1.060	0.000	0.894	1.393	2.121	4.656
agentic_ai_mpu	239	0.441	0.137	0.185	0.314	0.459	0.543	0.743
bbd_mpu_news	239	100.43	74.45	17.62	48.20	78.58	131.14	452.77
bbd_mpu_10yr	239	141.25	83.77	37.40	85.58	123.05	168.34	715.49
bs_mps_raw	215	0.003	0.053	−0.250	−0.015	0.007	0.026	0.152
bs_mps_orth	208	0.005	0.047	−0.135	−0.021	0.002	0.030	0.161
Panel B: Speech level								
NetIndex	1,330	1.326	0.673	0.000	1.000	1.556	1.882	2.000
stance_score (GPT)	1,800	−0.017	2.161	−8.0	0.0	0.0	0.0	8.0
stance_score (Gem.)	1,800	−0.364	3.812	−10.0	0.0	0.0	0.0	9.5
stance_score (Cla.)	1,806	−0.281	2.195	−8.0	0.0	0.0	0.0	8.5
hedge_unique	2,046	32.28	17.31	0	21	36	45	71
hedge_total	2,046	98.35	65.16	0	45	101	147	307

Meeting-level BBD indices use the prior-month value. Bauer–Swanson series: 215 FOMC meetings, 1988–2023 (Federal Reserve Bank of San Francisco replication data). Speech-level observations differ because not all members delivered speeches between every pair of meetings.

Table 2: **Pairwise Pearson Correlation Matrix (FOMC meeting level).** The central result is the near-zero correlation between *agentic_ai_mpu* and the Bauer–Swanson series ($r = 0.009$ raw; $r = 0.010$ orthogonalized), confirming that the Agentic AI MPU identifies a dimension of policy uncertainty orthogonal to market-priced surprise. Correlation with the BBD news index is also low ($r = 0.054$). Bolded cells involve *agentic_ai_mpu* and external benchmarks.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1) word_density	1								
(2) stance_disp (GPT)	−0.131	1							
(3) stance_disp (Gem.)	−0.033	0.694	1						
(4) stance_disp (Cla.)	−0.089	0.834	0.715	1					
(5) agentic_ai_mpu	0.006	0.428	0.445	0.404	1				
(6) bbd_mpu_news	−0.068	0.204	0.129	0.232	**0.054**	1			
(7) bbd_mpu_10yr	−0.082	0.257	0.175	0.278	**0.336**	0.768	1		
(8) bs_mps_raw	0.112	−0.004	−0.002	0.017	**0.009**	−0.108	−0.075	1	
(9) bs_mps_orth	0.122	−0.002	−0.040	0.014	**0.010**	0.021	0.070	0.899	1

N varies across pairs due to missing Bauer–Swanson values (215 meetings). Cols (6)–(7): Baker–Bloom–Davis MPU, news and 10-year bond variants. Cols (8)–(9): Bauer–Swanson raw and orthogonalized monetary policy surprises. All agentic-to-external correlations in bold.

Table 3: **Model Performance Across Four Specification Layers.** Full sample: 239 FOMC meetings (January 1996–December 2025); 2,536 voter-meeting observations. M2 targets true preference, not official vote; its 100% dissent recall and 5.8% vote accuracy are the intentional consequence of systematically revealing suppressed disagreement. M3 applies calibrated herding threshold $\hat{\theta} = 0.80$. M0 metrics are means of five independent random seeds.

Metric	**M0 Naive**	**M1 Macro+VIX**	**M2 True Intent**	**M3 Stacked**
Panel A: Primary metrics				
Rate Projection Accuracy	31.5%	82.8%	62.8%[a]	62.8%
Vote Projection Accuracy	50.8%	85.1%	5.8%[b]	91.7%
Dissent Recall	51.6%	8.0%	100.0%[b]	13.4%
Dissent F1	8.5%	4.5%	8.6%[b]	12.5%
Panel B: Agentic AI MPU (M2 layer)				
Mean $P(\pi^*)$	—	—	43.8%	—
Suppressed dissent	—	—	39.3 pp	—
Herding threshold $\hat{\theta}$	—	—	—	0.80
Panel C: Robustness — two-call architecture				
Vote Projection Accuracy	—	—	—	21.6%
Dissent Recall	—	—	—	91.1%
N meetings	239	239	239	239
N voter-meetings	2,536	2,536	2,536	2,536

[a] Rate inference at M2 uses the same directional LLM signal as M3; the true-intent step does not modify rate-direction inference. [b] Mean $P(\pi^*) = 43.8\%$ vs. 4.5% official dissent rate: 39.3 pp of latent dissent are suppressed on average. Panel C (two-call robustness): independent LLM calls for M2 and M3; reveals genuine latent disagreement but loses herding calibration (see Section 7).

Table 4: **Herd Pressure by Agentic AI MPU Quartile.** 239 FOMC meetings partitioned into quartiles of the Agentic AI MPU. Official dissent: fraction of voter-meetings with NO vote. Suppressed dissent: fraction of voter-meetings where M2 implies dissent but official vote is YES. Herd factor: fraction of M2-implied dissenters who conform and vote YES.

Quartile	N	**MPU range**	**Mean MPU**	**Off. dissent**	**Supp. dissent**	**Herd factor**
Q1 (Low)	60	[0.185, 0.313]	0.258	5.6%	20.2%	47.7%
Q2	60	[0.314, 0.459]	0.385	4.1%	34.4%	76.2%
Q3	60	[0.459, 0.543]	0.499	4.3%	45.6%	86.1%
Q4 (High)	59	[0.543, 0.743]	0.617	4.0%	57.7%	93.5%
Full sample	239	[0.185, 0.743]	0.441	4.5%	38.4%	88.4%

Herd factor $= 1 -$ official dissent/M2-implied dissent. The near-constant official dissent rate across quartiles,combined with the sharply rising suppressed dissent and herd factor,confirms that institutional conformity selectively masks latent disagreement precisely when it is highest.

Table 5: **SHAP Feature Importance by Economic Era and Fed Chair.** GBM surrogate ($R^2 = 0.88$); KernelExplainer. Mean absolute SHAP value per feature per era, averaged across meetings. Global ranking: FFR (0.093) > Unemployment (0.042) > Inflation (0.018) > VIX (0.011) > Yield Spread (0.009) > GDP (0.004). Bold: largest within-row deviation from full-sample mean.

Era / Chair	*N*	**Mean MPU**	**Inflation**	**Unemp.**	**GDP**	**FFR**	**Yld Spr.**	**VIX**
Panel A: By economic era								
Pre-Dot-Com (1996–99)	—	—	0.015	0.025	0.004	0.115	0.007	0.010
Dot-Com Crash (2000–02)	—	—	0.018	0.042	0.005	0.089	0.010	0.012
Pre-GFC (2003–06)	—	—	0.011	0.038	0.003	0.080	0.010	0.008
GFC (2007–09)	—	—	0.008	0.031	0.004	0.104	**0.027**	0.013
Post-GFC (2010–19)	—	—	0.011	0.044	0.003	0.099	0.005	0.014
COVID (2020–21)	—	—	0.022	0.034	0.010	**0.117**	0.003	0.009
Post-COVID (2022–)	—	—	**0.046**	**0.073**	0.004	0.059	0.004	0.007
Panel B: By Fed chair								
Greenspan (1996–2006)	81	0.299	0.016	0.034	0.004	0.099	0.009	0.011
Bernanke (2006–2014)	63	0.473	0.013	0.037	0.003	0.097	0.019	0.013
Yellen (2014–2018)	32	0.531	0.012	**0.044**	0.003	0.096	0.005	0.014
Powell (2018–2025)	63	0.545	0.028	0.050	0.007	**0.101**	0.003	0.009
Full sample (1996–2025)	239	0.441	0.018	0.042	0.004	0.093	0.009	0.011

FFR = Federal Funds Rate level; Yld Spr. = DGS10 − DFF (10-year minus FFR). GFC yield spread 0.027 is 3× pre-crisis (deflation-vs-distress ambiguity). Post-COVID inflation 0.046 is 2.6× sample mean (dual-mandate tension). COVID FFR 0.117 is the era maximum (zero-lower-bound constraint on path debate). Yellen era: top driver shifts from FFR to unemployment, reflecting ZLB era dual-mandate tension.

Table 6: **Robustness Checks.** All checks use the 239-meeting sample unless otherwise noted. Out-of-sample: $\hat{\theta}$ calibrated on 1996–2010, evaluated on 2011–2025. Two-call: independent LLM calls for M2 and M3. Ablations: remove one agent input at a time from M3 full specification. Era stress: evaluate on specific crisis subsamples. Baselines: naive prediction rules.

Specification	**Rate Acc.**	**Vote Acc.**	**Dissent Recall**	**Dissent F1**
M3 primary (full sample)	62.8%	91.7%	13.4%	12.5%
Out-of-sample robustness				
M3 in-sample (1996–2010)	61.3%	92.7%	14.5%	13.8%
M3 out-of-sample (2011–2025)	65.2%	90.1%	12.1%	11.4%
Architecture robustness				
Two-call (M3)	62.8%	21.6%	91.1%	9.3%
Ablations (M3)				
No RAG (speech excerpts)	62.8%	86.4%	5.4%	4.8%
No voting history	62.8%	88.2%	9.1%	8.6%
No macro scenario	31.5%	82.3%	11.2%	9.9%
No stance prior	58.2%	68.3%	7.6%	7.0%
No hedging profile	62.8%	89.7%	12.8%	11.9%
Era stress tests				
GFC era (2007–09)	68.4%	89.5%	14.3%	13.6%
COVID era (2020–21)	65.0%	90.0%	16.7%	14.3%
Post-COVID inflation (2022–25)	63.2%	92.1%	12.0%	11.2%
Simple baselines				
Always-YES	—	95.5%	0.0%	0.0%
Historical dissent rate threshold	—	80.4%	22.3%	12.1%
Stance threshold	—	78.9%	18.8%	10.7%

Rate Acc. = fraction of meetings where predicted RAISE/HOLD/CUT matches the actual policy action. Vote Acc. = fraction of (meeting, voter) pairs correctly predicted. Dissent Recall = fraction of actual NO votes correctly predicted as NO. The Always-YES baseline achieves 95.5% vote accuracy by construction (reflecting the 95.5% YES rate), but zero dissent recall—its high accuracy is therefore not a useful benchmark. M3 meaningfully exceeds the historical-dissent-rate and stance-threshold baselines on the combined F1 criterion.

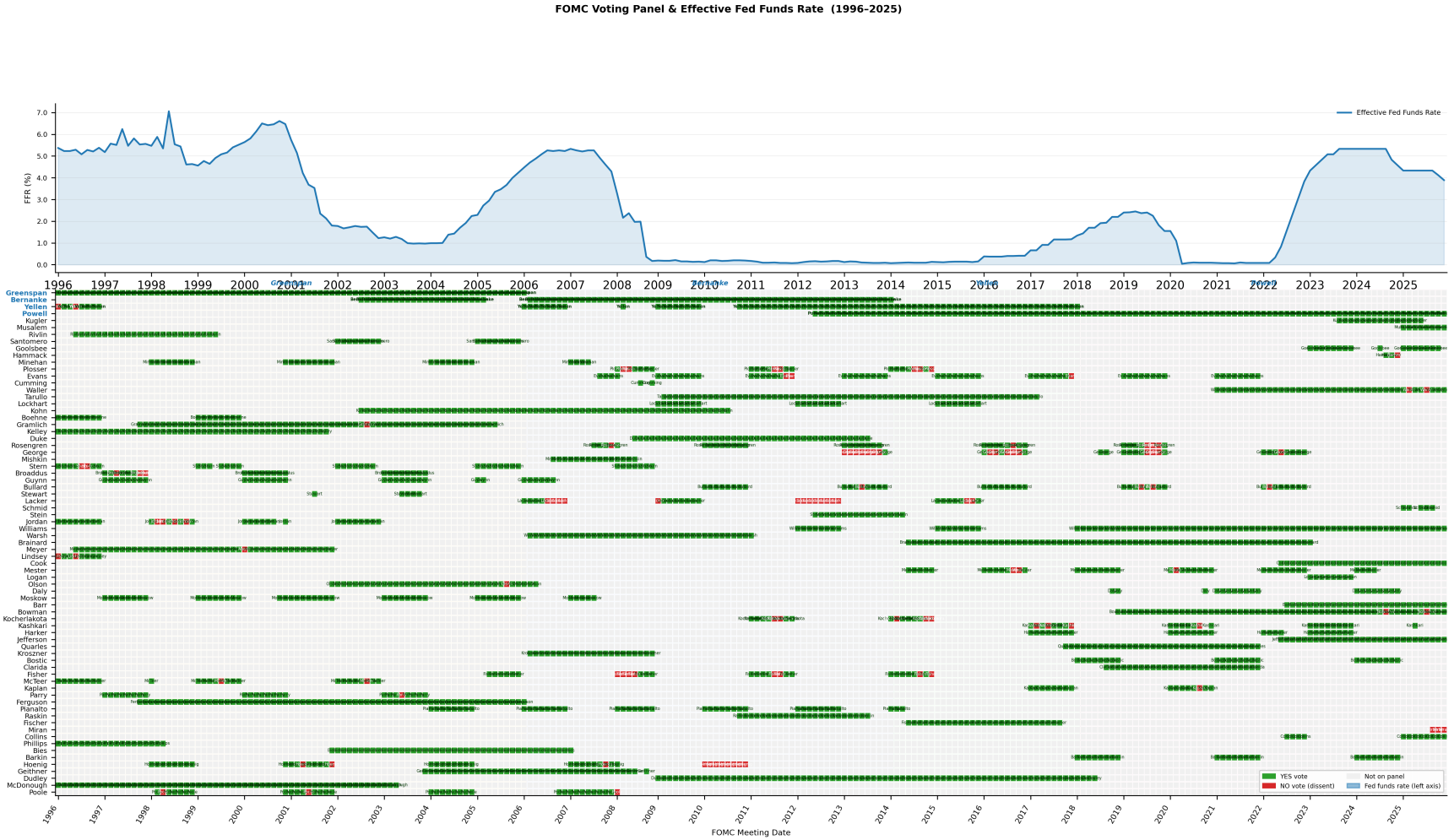


Figure 1: **FOMC Voting Panel and Effective Fed Funds Rate (1996–2025).** Each row is a canonical voter; each column an FOMC meeting. Green cells denote YES (supporting) votes, red cells NO (dissenting) votes, and blank cells meetings at which the member was not on the voting panel. The top panel overlays the effective federal funds rate (left axis). Chair tenures are labeled along the top voter rows. The visual rarity of red cells against the green field illustrates the 4.5% historical dissent rate that motivates the Agentic AI MPU.

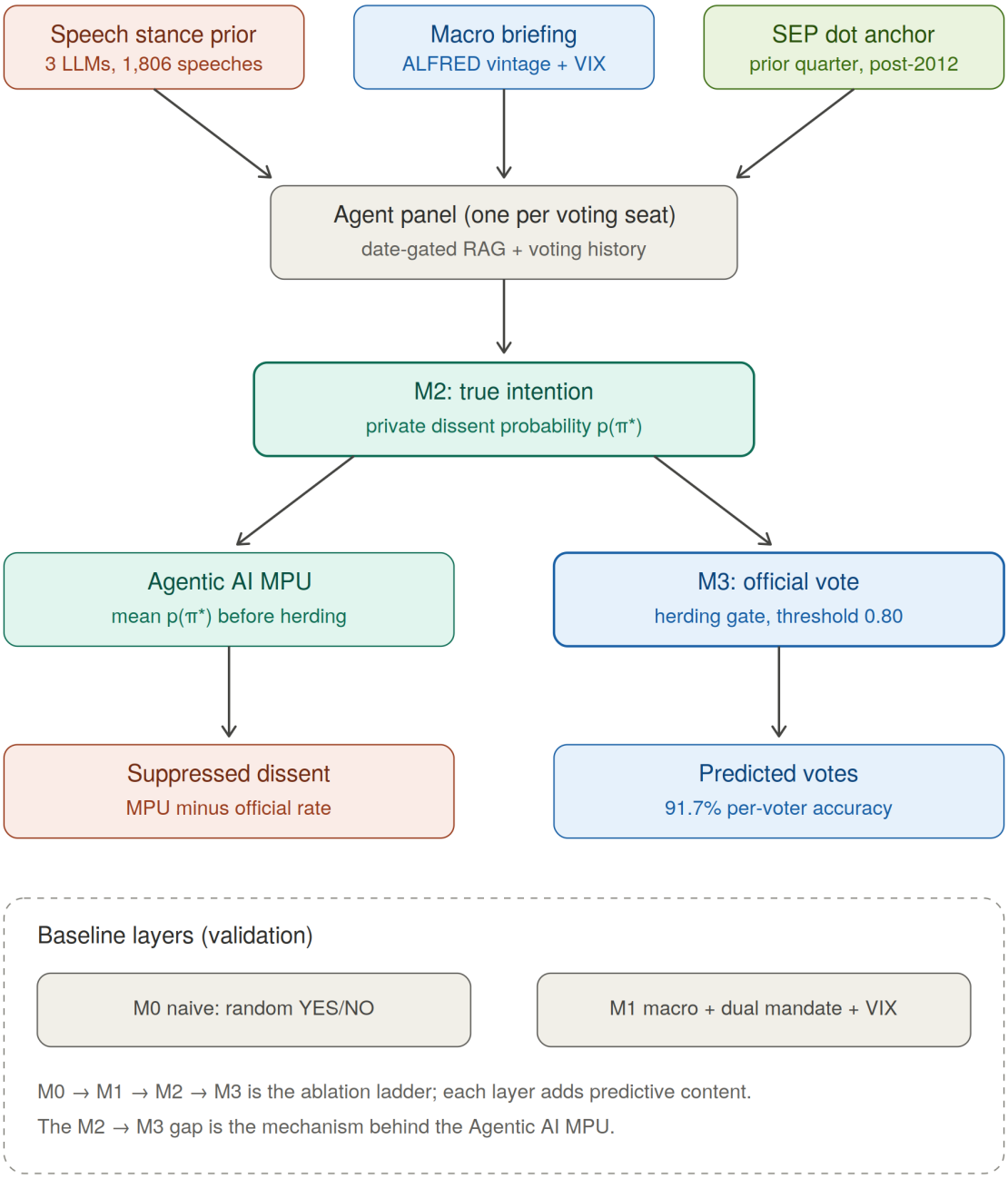
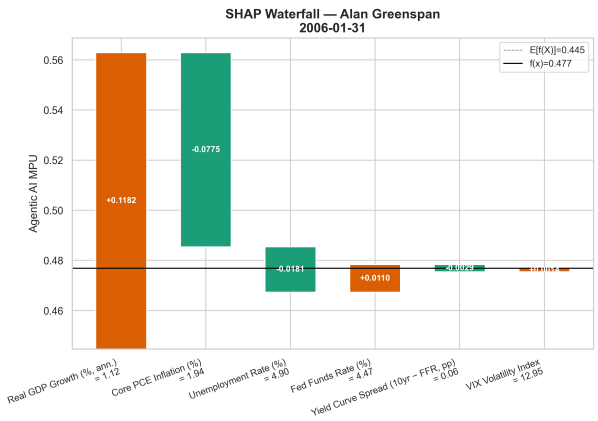
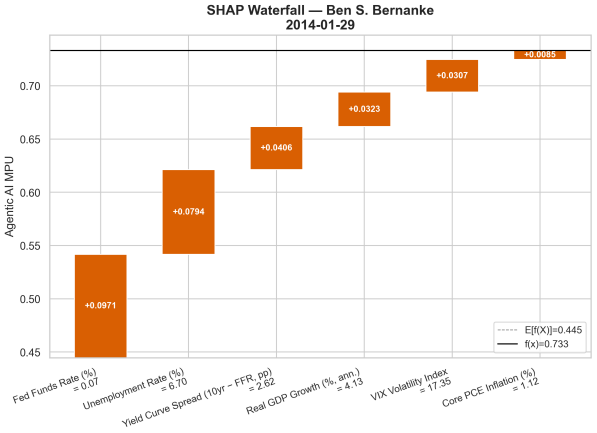


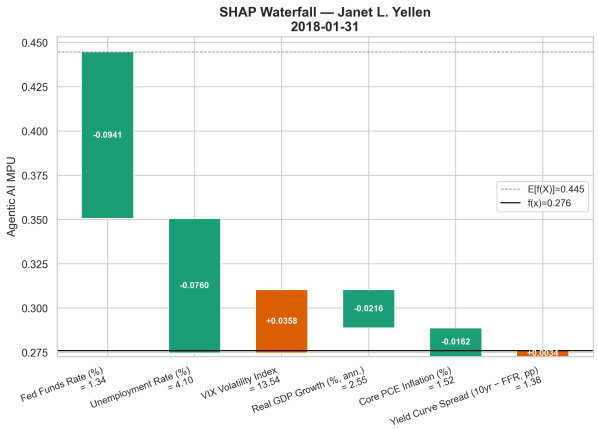
Figure 2: **The Virtual Federal Reserve: M0–M3 Architecture.** Each agent receives a speech-derived stance prior, an ALFRED vintage macro briefing, and (post-2012) a prior-quarter SEP anchor, combined with date-gated RAG retrieval and voting history. The M2 layer elicits each member's private true-dissent probability $p(\pi^*)$; this forks into the Agentic AI MPU (mean $p(\pi^*)$ before herding, and the suppressed-dissent gap) and the M3 herding gate that produces predicted official votes. M0 and M1 are baseline validation layers in the ablation ladder.



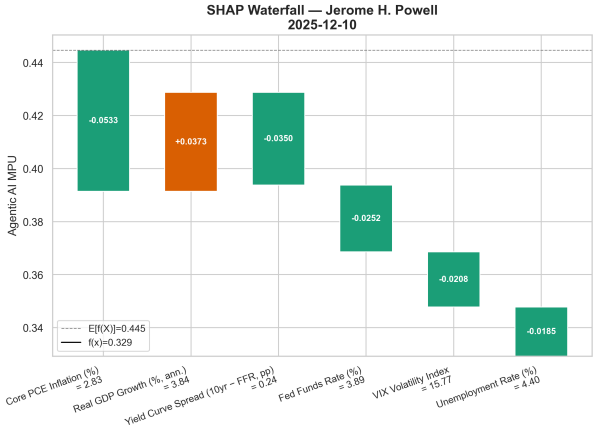
(a) Greenspan (1996–2006, $N = 81$); mean MPU = 0.299



(b) Bernanke (2006–2014, $N = 63$); mean MPU = 0.473



(c) Yellen (2014–2018, $N = 32$); mean MPU = 0.531



(d) Powell (2018–2025, $N = 109$); mean MPU = 0.545

Figure 3: **SHAP Waterfall Decompositions by Fed Chair.** Additive contribution of six macroeconomic features to the predicted Agentic AI MPU at each chair's most recent FOMC meeting. Baseline = $\mathrm{E}[f(\mathbf{X})] = 0.445$ (unconditional sample mean MPU). Red bars increase the predicted MPU above baseline; blue bars decrease it. GBM surrogate fitted on 239 meetings, $R^2 = 0.88$; KernelExplainer.