The Conservative AI: Diagnosing Hold Bias and Reliability Limits in Persona-Based Monetary Policy Simulation

Giyong Kim and Sojung Kim*
Bank of Korea
{kgy0617, sojung.kim}@bok.or.kr

Abstract

We examine whether large language models (LLMs) can reliably simulate historical FOMC policy decisions and whether persona-based agentic deliberation improves performance. Using strictly time-consistent vintage economic information, we evaluate multiple state-of-the-art LLMs on a three-way Hike/Hold/Cut classification task in both single-agent and multi-agent settings. Single-LLM baselines achieve nontrivial accuracy and track broad policy regime shifts, establishing a simple but strong benchmark. However, we identify a systematic behavioral asymmetry that we term *Hold bias*: models disproportionately favor Hold decisions and remain reluctant to predict Cut outcomes even during easing cycles. This conservatism is especially costly around regime turning points, where reliable adaptation matters most. We further find that standard agentic workflows, including debate and consensus-style aggregation, do not mitigate this problem and often amplify caution rather than improve accuracy. Overall, our results show that plausible deliberation is not sufficient for trustworthy decision support. Progress will require agentic systems explicitly designed to diagnose and correct structural bias, rather than merely reproducing surface-level committee interaction.

1 Introduction

Large language models (LLMs) and agentic workflows are increasingly used for decision support in high-stakes domains. In such settings, trustworthiness depends not only on average accuracy, but also on systematic bias, adaptability to regime shifts, and robustness beyond simple prompting. Monetary policy provides a natural stress test for these issues. Federal Open Market Committee (FOMC) decisions are discrete—hike, hold, or cut—but are formed by combining heterogeneous numerical and textual evidence under uncertainty.

Traditional models, such as the Taylor rule, VARs, DSGEs, and Random Forests, focus on predicting outcomes rather than modeling how decisions are formed. LLMs offer a different perspective. They can process numerical inputs and text jointly, while generating coherent rationales. This makes them a useful testbed for studying both decision quality and decision behavior. Recent work has explored LLM-based simulations of FOMC decisions, often using multi-agent systems to mimic committee interaction (Seok et al., 2024; Hou et al., 2025; Kazinnik and Sinclair, 2025; Takano et al., 2025).

Despite this progress, three gaps remain. First, prior studies focus on multi-agent architectures without strong single-LLM baselines. This makes it difficult to isolate the effect of agent interaction. Second, most analyses rely on a single foundation model (e.g., GPT-type), raising concerns about model-specific artifacts. Third, existing work provides limited evidence on trustworthiness: Are errors asymmetric? Do models underreact at turning points? Does deliberation improve reliability or only produce more plausible explanations?

Accordingly, we frame this study as a trustworthy-NLP diagnostic of LLM-based decision systems in a real, high-stakes domain. We address three questions: (1) How accurately can a single LLM reproduce historical FOMC decisions given vintage economic information? (2) Do models exhibit systematic behavioral biases, especially around regime shifts? (3) Does explicit deliberation through agentic workflows meaningfully improve reliability or interpretability?

To answer these questions, we evaluate multiple state-of-the-art LLMs in both single-agent and agentic settings under a controlled design with strictly time-consistent inputs. This setup lets us separate model effects from architectural effects while keeping the information environment fixed.

Our findings yield several insights. First,

* The views expressed in this paper are those of the authors and do not represent the official views of the Bank of Korea.

Proceedings of the 6th Workshop on Trustworthy NLP (TrustNLP 2026), pages 663–677

single-LLM baselines—particularly models such as Qwen3-Max and Gemini 3 Pro—can reproduce policy regime shifts with high accuracy using only snapshot indicators. This suggests that contemporary LLMs can effectively track policy regimes by recognizing salient economic signals, establishing a simple yet strong baseline.

Second, we identify a pervasive structural tendency toward conservative decisions, which we term *Hold bias*. Most models exhibit a strong preference for inaction (Hold), even when the realized decision involves a rate change. Notably, models with weaker Hold bias perform better, due to their willingness to switch policies at transition points.

Finally, we find limited evidence that standard agentic workflows improve decision quality. In many cases, direct synthesis and multi-round deliberation fail to outperform the single-agent baseline and can even degrade performance. Agentic interaction may improve interpretability, but interpretability and reliability do not necessarily move together.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the task and experimental setup. Section 4 presents the results, and Section 5 concludes.

2 Related works

Quantitative and Agent-Based Modeling. Monetary policy has traditionally been modeled using econometric frameworks and Taylor-type rules (Sarno et al., 2005; Brayton et al., 2014), recently extended by machine learning classifiers to predict discrete policy outcomes (Yoon and Fan, 2024). A parallel stream of research employs Reinforcement Learning (RL) and Agent-Based Modeling (ABM) to frame policy-making as a dynamic optimization problem. For instance, Hinterlang and Tänzer (2021) and Chen et al. (2025) demonstrate that RL agents can learn policy rules in stylized environments, while systems like ABIDES-Economist (Dwarakanath et al., 2025) simulate learning within complex markets. However, these frameworks typically rely on simplified information structures and lack the capacity for natural language deliberation.

NLP for Central Bank Communication. Another strand of research focuses on interpreting central bank texts. Studies such as FinBERT-FOMC (Gössi et al., 2023), central bank-specific model analysis (Kim et al., 2024), and interpretations of "Fedspeak" (Yao et al., 2025) examine whether language models can comprehend complex economic communications.

LLM Agents for Policy Simulation. Most relevant to our work is the nascent body of literature utilizing LLMs to simulate FOMC deliberation. This line of research is still in its early stages, primarily focusing on architectural exploration. Seok et al. (2024) introduced a role-based framework (MiniFed) where agents mimic FOMC members through analysis, discussion, and voting. Hou et al. (2025) extended this with explicit debate mechanisms and a coordinating agent. Other approaches include combining LLM simulations with Bayesian voting models to isolate behavioral dynamics (Kazinnik and Sinclair, 2025), and using latent belief variables (e.g., hawkish vs. dovish) to mediate social influence during deliberation (Takano et al., 2025).

Prior work on LLM-based FOMC simulation has focused mainly on agentic frameworks built on a single model, with limited comparison to strong single-LLM baselines. This leaves open whether observed behaviors are general or model-specific, and whether multi-agent interaction mitigates or amplifies decision bias.

3 Task Definition and Experimental Design

Our primary task is to predict the policy action taken at each FOMC meeting using only contemporaneous economic information. The prediction target is a three-class categorical variable—Hike, Hold, or Cut of the federal funds target rate. This formulation deliberately abstracts from the magnitude of rate changes and focuses on the directional policy stance, enabling consistent evaluation across heterogeneous economic regimes. We adopt this direction-only setting as a first-step diagnostic: Hold bias manifests first as failure to switch direction at regime transition points.

Data and Information Sets. Model inputs consist of vintage macroeconomic indicators observable prior to each meeting, covering key dimensions such as inflation, labor market conditions, real activity, housing, and financial conditions. To examine the impact of data representation, we organize these inputs into six predefined information sets (E1–E6). These sets vary along two dimensions: (1) the temporal granularity of numerical data (provided as either snapshot values or short-

Table 1: Summary of information sets (E1–E6) used in the experiments. All inputs are constructed from vintage data available prior to each FOMC meeting.

Info Set	Numeric Indicators	Trend Horizon	Beige Book
E1	Yes	Snapshot	No
E2	Yes	Snapshot	Yes
E3	Yes	3-month trend	No
E4	Yes	3-month trend	Yes
E5	Yes	6-month trend	No
E6	Yes	6-month trend	Yes

Table 2: Distribution of interest-rate decisions in the test set (2022–2025). Each year contains 8 FOMC meetings.

Decision	2022	2023	2024	2025	Total
Hold	1	4	5	5	15
Hike	7	4	0	0	11
Cut	0	0	3	3	6

horizon trends over three to six months), and (2) the inclusion of qualitative context. For the latter, selected sets incorporate textual summaries from the Federal Reserve's Beige Book, released shortly before each FOMC meeting. This allows us to evaluate whether narrative assessments can compensate for publication lags in hard data. A summary of these sets is provided in Table 1, and the full list of indicators is detailed in Appendix A.

In addition to meeting-specific vintage information (E1–E6), selected agentic configurations are permitted to retrieve historical policy precedents via retrieval-augmented generation (RAG). The retrieval corpus consists of all scheduled FOMC meetings from January 2000 through December 2025, including realized policy decisions and associated macroeconomic conditions.

Evaluation Period and Regimes. We evaluate all models on a test set of 32 scheduled FOMC meetings from 2022 to 2025. This period is intentionally selected to span distinct monetary policy regimes: a hike-intensive tightening phase (2022–2023) and a subsequent easing phase marked by rate cuts (2024–2025). As shown in Table 2, the distribution of policy decisions varies substantially across years. This diversity is crucial for assessing whether models can adapt their policy stance under markedly different macroeconomic environments.

Modeling Frameworks. We compare two distinct modeling paradigms. First, we evaluate Single-LLM Baselines, where a model directly maps information sets to a decision. These serve as a reference point to determine if sophisticated

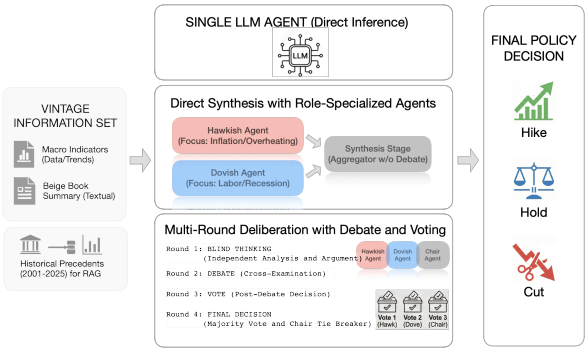

Figure 1: Overview of single-LLM and agentic decision frameworks used to predict FOMC policy actions from vintage information and historical precedents

architectures are strictly necessary. Second, we examine Agentic Workflows that decompose the decision process into interacting roles (e.g., Hawkish vs. Dovish perspectives). Within this class, we test both a direct synthesis setup and a multi-round deliberation framework involving debate, voting, and a final decision by a designated Chair. All approaches are evaluated on the identical task and test set to ensure a controlled comparison.

Evaluation Metrics and Inference Settings. Accuracy is our primary metric. To diagnose systematic bias, we additionally report Cut→Hold and Hike→Hold confusion rates. Because the number of realized Cut meetings is small, pooled counts across information sets can overstate certainty if treated as independent samples. We therefore supplement the pooled counts with meeting-level analyses: meeting-cluster bootstrap confidence intervals and leave-one-meeting-out (LOMO) sensitivity analysis. To quantify temporal responsiveness, we also report Regime-Shift Delay (RSD): for each change point where the ground-truth action switches, we measure the number of subsequent meetings until the model first predicts the new ground-truth label; cases never recovered within the evaluation horizon are treated as right-censored. Unless otherwise stated, single-LLM baselines use deterministic decoding (temperature 0.0, top-p 1.0) to minimize stochastic variation. The agentic frameworks use modest role-specific temperatures to encourage diverse viewpoints.

Figure 1 provides an overview of the experimental frameworks considered in this study.

4 Results

Single LLMs perform competitively under minimal information.

Table 3: Accuracy of FOMC Decision Prediction across Input Configurations (Single LLM - Base Prompt)

Information	**E1**	**E2**	**E3**	**E4**	**E5**	**E6**
Model	(Snapshot)	(Snapshot+BB)	(Trend 3M)	(Trend 3M+BB)	(Trend 6M)	(Trend 6M+BB)
gpt-4o	25/32 (78.1%)	24/32 (75.0%)	24/32 (75.0%)	24/32 (75.0%)	23/32 (71.9%)	23/32 (71.9%)
gpt-5.2	26/32 (81.2%)	27/32 (84.4%)	**27/32 (84.4%)**	**27/32 (84.4%)**	27/32 (84.4%)	27/32 (84.4%)
gemini-2.0-flash	25/32 (78.1%)	25/32 (78.1%)	25/32 (78.1%)	25/32 (78.1%)	25/32 (78.1%)	25/32 (78.1%)
gemini-3-pro-preview	29/32 (90.6%)	29/32 (90.6%)	26/32 (81.2%)	26/32 (81.2%)	**28/32 (87.5%)**	**29/32 (90.6%)**
Qwen3-Max	**30/32 (93.8%)**	**30/32 (93.8%)**	26/32 (81.2%)	**27/32 (84.4%)**	27/32 (84.4%)	28/32 (87.5%)
DeepSeek V3.2	24/32 (75.0%)	22/32 (68.8%)	21/32 (65.6%)	23/32 (71.9%)	22/32 (68.8%)	26/32 (81.2%)

4.1 Single LLM Baselines

This section evaluates the performance of single LLMs on the FOMC policy decision prediction task. The base prompt provides a concise description of the task and the available information. We consider six input configurations (E1–E6); an example of the full prompt for E6 is provided in Appendix A.2.1. Each configuration is evaluated across multiple state-of-the-art LLMs: gpt-4o, gpt-5.2, gemini-2.0-flash, gemini-3-pro-preview, Qwen3-Max, and DeepSeek-V3.2.

As shown in Table 3, several models already achieve high accuracy under the most restrictive setting (E1). In particular, Qwen3-Max and gemini-3-pro-preview perform strongly using snapshot-level macroeconomic indicators alone, suggesting that some LLMs can extract policy-relevant signals without explicit temporal or textual context. At the same time, performance varies substantially across models, indicating strong model dependence.

We next examine the marginal utility of additional information. Averaged across models, accuracy is highest for E1 (82.81%) and E6 (82.29%), followed by E2 (81.77%), E5 (79.17%), E4 (78.65%), and E3 (77.08%). Thus, richer inputs do not systematically improve performance. Models that already perform well in E1, such as Qwen3-Max and gemini-3-pro-preview, often experience performance degradation when partial additional information is introduced, with recovery only under the most comprehensive configuration (E6). In contrast, gpt-5.2 exhibits a more monotonic pattern, benefiting from trend information and showing stable or improved performance when Beige Book summaries are included.

Figure 2 compares, for each FOMC meeting, the ground-truth policy decision with model predictions under the most restrictive E1 configuration. The figure also includes a Taylor-rule benchmark. Ground truth corresponds to realized FOMC target rate decisions (Hike, Hold, or Cut) at each scheduled meeting. From the temporal patterns, Qwen3-Max and gemini-3-pro-preview stand out as closely tracking regime shifts across tightening and easing cycles. In particular, both models correctly identify four out of six Cut decisions, reacting more promptly to easing transitions than other LLMs and thereby achieving higher overall accuracy.

For a lightweight classical reference under the same evaluation period, the Taylor-rule benchmark achieves 16/32 (50.0%) accuracy. We do not treat this as a fully matched econometric benchmark suite—direct VAR/DSGE or ML comparisons would require additional design choices under our vintage-data, three-way classification setup—but it provides a useful lower-complexity reference point for the single-LLM results.

Single LLM decisions exhibit a strong Hold bias. A clear and consistent pattern emerges across models: a strong preference for the Hold decision. Even when the ground truth corresponds to an active policy change, many models continue to predict Hold. We refer to this phenomenon as *Hold bias*, a structural property of single-LLM decision making.

Hold bias manifests asymmetrically across policy regimes. During easing cycles, models frequently predict Hold even when the true decision is Cut, indicating reluctance to commit to policy reversals despite sufficiently strong easing signals. During tightening episodes, some models similarly predict Hold despite clear Hike decisions. When faced with uncertainty or mixed evidence, single LLMs tend to adopt a risk-averse strategy, favoring policy inaction over explicit directional moves.

Model-specific conservatism drives performance differences. Table 4 quantifies this behavior by reporting, for each model, the frequency with which true Cut or Hike decisions are misclassified as Hold. Substantial heterogeneity is observed across models. Qwen3-Max and gemini-3-pro-preview exhibit relatively lower Cut→Hold rates, while gpt-4o, gemini-2.0-flash, and DeepSeek-V3.2 show pronounced reluctance to predict easing. These results suggest that overall accuracy is driven less by input richness than by model-specific deci-

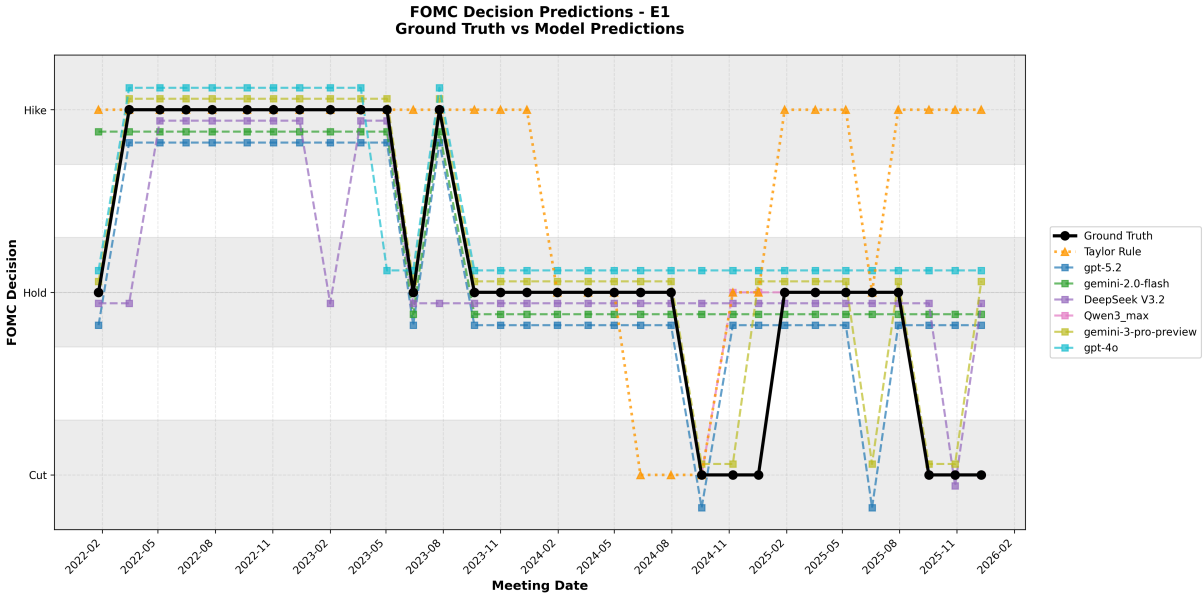

Figure 2: Ground truth FOMC policy decisions versus model predictions under E1

sion conservatism.

Table 4: Hold Bias in Single-LLM Policy Predictions

Model	Cut→Hold	(%)	Hike→Hold	(%)
gpt-4o	36/36	100.0	10/66	15.2
gpt-5.2	28/36	77.8	0/66	0.0
gemini-2.0-flash	35/36	97.2	0/66	0.0
gemini-3-pro-preview	17/36	47.2	0/66	0.0
Qwen3-Max	17/36	47.2	0/66	0.0
DeepSeek V3.2	30/36	83.3	20/66	30.3

The pooled Cut→Hold counts in Table 4 aggregate six input configurations for the same six Cut meetings and therefore should not be interpreted as 36 independent observations. To make the uncertainty explicit, Table 5 reports meeting-level bootstrap estimates that treat meetings—not prompt variants—as the resampling unit.

Table 5: Meeting-level uncertainty for Cut→Hold rates. Meetings, not prompt variants, are treated as the resampling unit.

Model	Mean Cut→Hold	95% CI
gpt-4o	1.00	[1.00, 1.00]
gemini-2.0-flash	0.97	[0.92, 1.00]
DeepSeek V3.2	0.83	[0.72, 0.94]
gpt-5.2	0.78	[0.44, 0.97]
gemini-3-pro-preview	0.47	[0.17, 0.83]
Qwen3-Max	0.47	[0.25, 0.72]

The qualitative ranking remains stable under leave-one-meeting-out (LOMO) analysis. In particular, Qwen3-Max and gemini-3-pro-preview remain in the 0.37–0.57 range across exclusions, well below DeepSeek-V3.2 (0.80–0.87) and gemini-2.0-flash (0.97–1.00). Thus, the lower-bias pattern of Qwen3-Max and gemini-3-pro-preview is not driven by one or two idiosyncratic meetings.

Table 6: Regime-Shift Delay (RSD) across single-LLM baselines. Lower is better.

Model	RSD
gemini-3-pro-preview	0.10
Qwen3-Max	0.14
gpt-5.2	0.24
gemini-2.0-flash	1.71
DeepSeek V3.2	3.38
gpt-4o	3.52

Lower Hold bias corresponds to faster regime adaptation. To quantify temporal responsiveness, we compute Regime-Shift Delay (RSD). As shown in Table 6, models with lower Cut→Hold bias adapt much faster to policy turning points. gemini-3-pro-preview and Qwen3-Max have near-zero delay, gpt-5.2 exhibits modest delay, and gpt-4o and DeepSeek-V3.2 often fail to switch promptly. Across models, we observe 46 right-censored transition cases in which the model never predicts the new label within the evaluation horizon. These failures are concentrated on Hold→Cut and Hold→Hike transitions, whereas transitions into Hold are typically recognized immediately. This supports the interpretation of Hold bias as a robustness problem of delayed regime adaptation rather than merely a static class-imbalance effect.

Reasoning and retrieval do not mitigate Hold bias. To assess whether additional reasoning and retrieval signals can alter the hold-biased behavior

Table 7: Performance of `gpt-4o` under different input configurations (E1–E6) and inference strategies. Accuracy is reported as the number of correct predictions out of 32 FOMC meetings, with percentages in parentheses.

Method	**E1**(Snapshot)	**E2**(Snapshot+BB)	**E3**(Trend 3M)	**E4**(Trend 3M+BB)	**E5**(Trend 6M)	**E6**(Trend 6M+BB)
Baseline (zeroshot)	25/32 (78.1%)	24/32 (75.0%)	24/32 (75.0%)	24/32 (75.0%)	23/32 (71.9%)	23/32 (71.9%)
CoT (zeroshot)	19/32 (59.4%)	19/32 (59.4%)	21/32 (65.6%)	21/32 (65.6%)	23/32 (71.9%)	23/32 (71.9%)
Few-shot (fixed)	23/32 (71.9%)	23/32 (71.9%)	24/32 (75.0%)	23/32 (71.9%)	23/32 (71.9%)	24/32 (75.0%)
Few-shot (dynamic)	18/32 (56.2%)	18/32 (56.2%)	20/32 (62.5%)	20/32 (62.5%)	20/32 (62.5%)	21/32 (65.6%)
RAG (summary)	21/32 (65.6%)	21/32 (65.6%)	20/32 (62.5%)	21/32 (65.6%)	20/32 (62.5%)	20/32 (62.5%)
RAG (structured)	23/32 (71.9%)	23/32 (71.9%)	22/32 (68.8%)	22/32 (68.8%)	22/32 (68.8%)	23/32 (71.9%)

of `gpt-4o`, we conduct auxiliary experiments incorporating chain-of-thought prompting, few-shot learning, and retrieval-augmented generation. Table 7 summarizes the results. Overall, these interventions do not materially shift the model away from its conservative preference for *Hold*. Notably, chain-of-thought prompting does not mitigate the bias and, in some settings, appears to reinforce it. One plausible interpretation is that step-by-step reasoning makes competing macroeconomic signals more explicit, which heightens perceived uncertainty and leads the model to default to the status quo. This pattern suggests that Hold bias is not merely a byproduct of shallow inference, but may persist—and occasionally strengthen—even under structured reasoning. Full details are provided in Appendix A.3.

4.2 Agentic Decision-Making Frameworks

This section turns to agentic frameworks that explicitly decompose monetary policy judgment into multiple interacting roles. We study two configurations that differ in how policy perspectives are combined: a direct synthesis framework with role-specialized agents and a multi-round deliberation framework involving debate, voting, and a final chair decision. These setups should be interpreted as *surface-level deliberation proxies* rather than full institutional simulations of the FOMC.

4.2.1 Direct Synthesis with Role-Specialized Agents

We begin with a minimal agentic extension that introduces explicit role specialization. The system consists of three agents: two role-specialized policy agents and a separate synthesizer agent. This design allows us to examine whether exposing the model to polarized policy arguments alone—without inter-agent debate—can influence the final decision outcome.

Agent configuration. The framework comprises two parallel policy agents and one aggregation agent. The policy agents represent contrasting monetary policy orientations: a hawkish agent that prioritizes price stability and a dovish agent that emphasizes employment conditions. Both policy agents are provided with the same meeting-specific information set, including the current federal funds rate and vintage macroeconomic indicators. Each agent independently analyzes the information and argues for a preferred policy action (Hike, Hold, or Cut) from its assigned perspective, ending its response with an explicit recommendation.

Aggregation and interaction structure. The outputs of the hawkish and dovish agents are passed to a third agent acting as a synthesizer. This synthesizer aggregates the two viewpoints into a final decision and produces a discrete choice, a confidence level, and a brief justification that explicitly references both perspectives. Crucially, the synthesizer does not participate in the initial analysis and does not engage in debate, voting, or iterative interaction with the policy agents. The aggregation is performed in a single step, reflecting a one-shot synthesis rather than a deliberative process.

By separating policy reasoning from aggregation, this setup isolates the effect of role specialization under minimal interaction. It allows us to test whether the presence of explicitly polarized policy arguments can mitigate conservative default behavior, or whether the final decision continues to converge toward Hold even when opposing perspectives are explicitly represented. The exact prompt templates used for all three agents are provided in Appendix A.2.2.

Direct synthesis does not systematically improve accuracy. Table 8 reports the performance of the direct synthesis agentic framework across the six information sets. For ease of comparison with single-LLM baselines, we report both the raw accuracy (number of correct predictions out of 32 meetings) and, in parentheses, the change in the number of correct predictions relative to the corresponding single-LLM result. Positive values indicate improvements over the single-LLM baseline, while

Table 8: Accuracy of direct synthesis agentic framework across information sets.

Model	E1	E2	E3	E4	E5	E6
GPT-4o	24/32	24/32	23/32	24/32	23/32	24/32
	(-1)	(0)	(-1)	(0)	(0)	(+1)
GPT-5.2	26/32	26/32	26/32	26/32	26/32	26/32
	(0)	(-1)	(-1)	(-1)	(-1)	(-1)
gemini-2.0-flash	25/32	25/32	25/32	25/32	25/32	25/32
	(0)	(0)	(0)	(0)	(0)	(0)
gemini-3-pro-preview	28/32	27/32	26/32	26/32	27/32	25/32
	(-1)	(-2)	(0)	(0)	(-1)	(-4)
Qwen3-Max	24/32	24/32	23/32	24/32	24/32	23/32
	(-6)	(-6)	(-3)	(-3)	(-3)	(-5)
DeepSeek-V3.2	22/32	22/32	20/32	24/32	22/32	21/32
	(-2)	(0)	(+1)	(-1)	(0)	(-5)

negative values indicate performance deterioration.

Overall, the direct synthesis framework fails to deliver systematic accuracy gains over single-LLM baselines. Across models and information sets, performance differences are generally small and frequently negative. This indicates that explicit role specialization alone—without deliberation or interaction—is insufficient to reliably improve policy decision prediction. In many cases, the aggregation step appears to inherit the dominant tendencies of the underlying model rather than resolving conflicting policy signals through synthesis.

Performance degradation in strong single-LLM models. The negative effects of direct synthesis are most pronounced for strong single-LLM models. Qwen3-Max and gemini-3-pro-preview show consistent performance deterioration across information sets, suggesting that centralized synthesis can disrupt otherwise effective single-model decision behavior.

Systematic failure to predict rate cuts. For GPT-5.2, the direct synthesis framework yields the same accuracy (26/32) across all information sets. Notably, this corresponds exactly to failing to predict all six Cut decisions, while Hike and Hold outcomes are otherwise predicted correctly.

4.2.2 Multi-Round Deliberation with Debate and Voting

The limitations of direct synthesis motivate a more structured agentic design that explicitly incorporates deliberation and aggregation. We implement a multi-round deliberation framework—a mini council—that captures core elements of committee-based monetary policy decision making.

Agent configuration. The council consists of three agents: a hawkish agent, a dovish agent, and a chair agent. The hawk and dove represent opposing policy orientations, while the chair acts as a neutral moderator rather than a synthesizer.

The chair operates under explicit constraints. First, it cannot recommend a policy stance more hawkish than the hawk or more dovish than the dove (median-voter feasibility). Second, during disinflationary phases, it applies asymmetric risk considerations that favor Cut over Hold. Third, it seeks majority coalitions while maintaining coherence with forward guidance.

In addition to meeting-specific vintage information (E1–E6), agents are allowed to retrieve historical precedents via retrieval-augmented generation (RAG). Retrieval is restricted to past FOMC decisions that predate the meeting under consideration, preventing time leakage. Up to three historical episodes are retrieved based on similarity in inflation and unemployment conditions.

All agents are required to ground their reasoning by citing at least one historical precedent in the following format: `[Year/period; inflation x%; unemployment y%; policy rate z%] → [Fed action] → [Outcome]`. Agents must state two similarities and one key difference relative to the current context. This anchors reasoning in empirical policy history and mitigates hallucination.

Deliberative interaction and aggregation. The council follows a fixed four-round deliberation protocol.

Round 1: Independent analysis. Each agent independently analyzes the economic data and retrieved precedents without observing the positions of other agents. Each produces an initial vote (Hike, Hold, or Cut) with a magnitude specification (0, $\pm$25bp, $\pm$50bp, or $\pm$75bp).

Round 2: Debate. All initial positions are revealed simultaneously. Agents critique the arguments of the other agents and defend their own positions, identifying weaknesses in opposing views.

Round 3: Final vote. After considering the debate, each agent casts a final vote, which may differ from its initial position.

Round 4: Majority decision. The final policy decision is determined by majority vote. Unanimous (3:0) and split (2:1) outcomes are decided directly. In the rare case of a three-way split (1:1:1), the chair's vote serves as the tiebreaker.

We employ role-specific temperature settings to reflect policy stances: 0.2 for the hawk, 0.4 for the dove, and 0.1 for the chair. Full prompt templates are provided in Appendix A.2.3.

Table 9: Accuracy of multi-round deliberation framework across information sets.

Model	E1	E2	E3	E4	E5	E6
GPT-5.2	26/32	25/32	24/32	23/32	23/32	24/32
	(0)	(-2)	(-3)	(-4)	(-4)	(-3)
gemini-3-pro-preview	27/32	27/32	26/32	25/32	26/32	26/32
	(-2)	(-2)	(0)	(-1)	(-2)	(-3)
Qwen3-Max	23/32	27/32	21/32	21/32	23/32	19/32
	(-7)	(-3)	(-5)	(-6)	(-4)	(-9)

Limited gains from multi-round deliberation. Table 9 reports the performance of the multi-round deliberation framework across information sets for three flagship models. As before, values in parentheses denote changes relative to the corresponding single-LLM baseline.

Overall, multi-round deliberation does not yield systematic improvements in predictive accuracy. GPT-5.2 shows no evidence that deliberation mitigates its conservative decision pattern. More strikingly, Qwen3-Max—one of the strongest performers in the single-LLM setting—exhibits substantial performance degradation under deliberation.

One plausible interpretation is that Qwen3-Max's strong single-model accuracy partly reflects a lower threshold for predicting cuts during easing regimes. However, we cannot rule out alternative explanations, including sensitivity to longer agent prompts, cue ordering, or other context-format effects introduced by structured interaction. What is clear is that the single-model advantage does not survive deliberative aggregation.

4.3 Interpretable Deliberation: A Qualitative Illustration

As a qualitative illustration, we present one representative case that shows the reasoning structures that can emerge from agent-based deliberation.

In the October 29, 2025 decision, both agent frameworks arrive at a Hold decision through ideation and debate among agents with opposing policy views. Although this prediction differs from the realized outcome and from the single-LLM baseline, the resulting analysis reveals economically meaningful reasoning patterns that are largely absent from direct single-model predictions.

The deliberation exhibits policy coherence: inflation, labor market conditions, and demand indicators are interpreted jointly rather than in isolation, for example through a "jobless growth" narrative in which spending remains resilient despite a sharp slowdown in hiring. It also makes the dual-mandate trade-off explicit, framing the decision as a tension between "entrenched inflation expectations" and a "stalling labor market" rather than as a one-sided signal for immediate action. In addition, the agentic synthesis shows rejection completeness. Both alternative actions are explicitly considered and dismissed with distinct justifications—rejecting a rate hike as "reckless in a bending labor market," while rejecting a rate cut as a threat to "price stability and policy credibility." The Hold decision thus emerges as a residual choice under mandate conflict and uncertainty, rather than as a default outcome.

More broadly, the structure and tone of the reasoning resemble FOMC-style deliberation, emphasizing uncertainty, lagged policy transmission, and credibility concerns. These features arise endogenously from interaction among agents, without prompts designed to enforce economic theory or interpretability.

This example suggests that agentic workflows can provide interpretable reasoning pathways that complement accuracy-based evaluation. In our setting, such traces do not consistently improve decision quality and appear to retain some of the Hold bias observed in the single-LLM baseline. Yet this does not diminish their value. Instead, it clarifies where that value lies: agentic systems make policy trade-offs explicit, expose model failure modes, and provide an auditable basis for diagnosing and mitigating systematic biases. Preserving these interpretability benefits while reducing inherited Hold bias is therefore an important direction for future work.

5 Conclusion

We find that single-LLM baselines achieve nontrivial accuracy in the Hike/Hold/Cut classification task, effectively tracking broad regime changes. However, performance is hindered by a structural *Hold bias*, where models remain reluctant to predict "Cut" decisions even in easing cycles. Furthermore, we observe that naively "agentifying" the process—via debate or consensus—fails to improve performance. Instead, deliberative aggregation tends to amplify caution, suggesting that standard agentic architectures are insufficient to mitigate the bias. An important direction for future work is therefore to identify the sources of this bias and develop agentic systems explicitly designed to diagnose and mitigate it, rather than merely simulating surface-level committee deliberation.

Limitations

1. Information asymmetry and institutional realism. Real-world policy decisions rely on private internal materials unavailable to public-source-based models. Our setups also abstract away organizational hierarchy, reputational concerns, and strategic interaction, so they should be interpreted as controlled approximations rather than full institutional simulations.

2. Potential temporal leakage and API drift. Because most evaluated systems are closed commercial APIs, we cannot fully rule out pre-training contamination or other forms of temporal leakage on historical data. Although we use deterministic decoding to reduce run-to-run variance, future model updates may still affect reproducibility.

3. Hold bias attribution and agentic scope. Our results are diagnostic rather than causal: we cannot yet determine how much Hold bias arises from the economic data, alignment, or other training artifacts. Relatedly, our minimalist agent designs are useful for controlled comparison, but they may understate the value of stronger mechanisms such as tool use, uncertainty tracking, abstention, or adaptive aggregation.

4. Evaluation scope. We study Hike/Hold/Cut direction rather than basis-point magnitude, policy paths, or probabilistic signaling. More comprehensive evaluation should also examine calibration, cost-sensitive errors, prompt stability, and responsiveness at regime shifts.

Broader Impact

This work does not advocate automated monetary-policy decision making. Instead, it highlights a practical risk: models that appear prudent may encode asymmetric underreaction to regime change, which could be harmful if used without strong human oversight. At the same time, the benchmark can support safer decision-support research by exposing bias, interpretability gaps, and robustness failures before deployment.

References

Flint Brayton, Thomas Laubach, and David Reifschneider. 2014. The frb/us model: A tool for macroeconomic policy analysis. *FEDS Notes*. Board of Governors of the Federal Reserve System.

Mingli Chen, Rama Cont, Andreas Joseph, Michael Kumhof, Xinlei Pan, Wei Xiong, and Xuan Zhou. 2025. Deep reinforcement learning in a monetary model. Technical Report Staff Working Paper No. 1,142, Bank of England.

Kshama Dwarakanath, Tucker Balch, and Svitlana Vyetrenko. 2025. Abides-economist: Agent-based simulator of economic systems with learning agents. *Computing Research Repository*, arXiv:2402.09563.

Sandro Gössi, Ziwei Chen, Wonseong Kim, Bernhard Bermeitinger, and Siegfried Handschuh. 2023. Finbert-fomc: Fine-tuned finbert model with sentiment focus method for enhancing sentiment analysis of fomc minutes. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, ICAIF '23, pages 357–364. Association for Computing Machinery.

Natascha Hinterlang and Alina Tänzer. 2021. Optimal monetary policy using reinforcement learning. Technical Report Discussion Paper No. 51/2021, Deutsche Bundesbank.

Yuhan Hou, Tianji Rao, Jeremy Tan, Adler Viton, Xiyue Zhang, David Ye, Abhishek Kodi, Sanjana Dulam, Aditya Paul, and Yikai Feng. 2025. Fedsight ai: Multi-agent system architecture for federal funds target rate prediction. *Computing Research Repository*, arXiv:2512.15728. NeurIPS 2025 Generative AI in Finance Workshop.

Sophia Kazinnik and Tara M. Sinclair. 2025. Fomc in silico: A multi-agent system for monetary policy decision modeling. Technical Report Working Paper No. 2025-005, The George Washington University, Department of Economics. H. O. Stekler Research Program on Forecasting.

Wonseong Kim, Jan Spörer, Choong Lyol Lee, and Siegfried Handschuh. 2024. Is small really beautiful for central bank communication? evaluating language models for finance: Llama-3-70b, gpt-4, finbert-fomc, finbert, and vader. In *Proceedings of the 5th ACM International Conference on AI in Finance*, ICAIF '24, pages 626–633. Association for Computing Machinery.

Lucio Sarno, Daniel L. Thornton, and Giorgio Valente. 2005. Federal funds rate prediction. *Journal of Money, Credit and Banking*, 37(3):449–471.

Sungil Seok, Shuide Wen, Qiyuan Yang, Juan Feng, and Wenming Yang. 2024. Minifed: Integrating llm-based agentic-workflow for simulating fomc meeting. *Computing Research Repository*, arXiv:2410.18012.

K. Takano, M. Hirano, and K. Nakagawa. 2025. Modeling hawkish–dovish latent beliefs in multi-agent debate-based LLMs for monetary policy decision classification. In *PRIMA 2025: Principles and Practice of Multi-Agent Systems*, volume 16366 of *Lecture Notes in Computer Science*. Springer, cHAM.

Rui Yao, Qi Chai, Jinhai Yao, Siyuan Li, Junhao Chen, Qi Zhang, and Hao Wang. 2025. Interpreting fedspeak with confidence: A LLM-based uncertainty-aware framework guided by monetary policy transmission paths. *Computing Research Repository*, arXiv:2508.08001.

Jungyeon Yoon and Juanjuan Fan. 2024. Forecasting the direction of the Fed's monetary policy decisions using random forest. *Journal of Forecasting*, 43(7):2848–2859.

A Appendix

A.1 A List of Macroeconomic Indicators

Table 10 summarizes the macroeconomic indicators used to construct the information sets provided to the LLMs and agents, all of which are assembled using strictly vintage data available prior to each FOMC meeting.

A.2 Prompt Templates

A.2.1 An Example of Baseline Single-LLM Prompt (E6)

```
You are an expert Federal Reserve analyst.

INSTRUCTIONS:
1. Analyze the SPECIFIC economic data
     provided below (Inflation, Employment,
     Growth, etc.).
2. Do NOT state that data is missing. The
     provided report contains the necessary
     vintage data.
3. Based on the "Current Federal Funds
     Rate" and economic conditions, deciding
     whether to Hike, Cut, or Hold.

============================================
FOMC Meeting Date: 2022-01-26
Current Federal Funds Rate: 0.25%
============================================

============================================
Economic Indicators Report
(Context for 2022-01-26 with 6M Trend)
============================================

INFLATION & PRICES
------------------
- Core PCE (YoY): 4.68% (Prev: 4.68%,
     4.12%, 3.62%, 3.62%, 3.54%, 3.39%)
- Headline PCE (YoY): 5.73% (Prev: 5.73%,
     5.05%, 4.26%, 4.17%, 3.99%, 3.91%)
- Core CPI (YoY): 5.49% (Prev: 4.96%,
     4.58%, 4.04%, 3.98%, 4.23%, 4.45%)
- Headline CPI (YoY): 7.12% (Prev: 6.88%,
     6.24%, 5.38%, 5.20%, 5.28%, 5.32%)
- PPI Final Demand (YoY): 9.75% (Prev:
     9.74%, 8.63%, 8.59%, 8.28%, 7.70%,
     7.14%)
- Sticky Price CPI (YoY): 3.50% (Prev:
     3.23%, 3.05%, 2.66%, 2.37%, 2.35%,
     2.60%)
- 1-Year Expected Inflation: 4.80% (Prev:
     4.80%, 4.90%, 4.60%, 4.60%, 4.70%,
     4.20%)
- 10Y Breakeven Inflation: 2.41% (Prev:
     nan%, nan%, 2.66%, 2.34%, 2.35%, 2.35%)

LABOR MARKET
------------
- Unemployment Rate: 3.90% (Prev: 4.20%,
     4.60%, 4.80%, 5.20%, 5.40%, 5.90%)
- Non-farm Payrolls (MoM Change): 199.00
     (Prev: 210.00, 531.00, 194.00, 235.00,
     943.00, 850.00)k
- Avg Hourly Earnings (YoY): 4.68% (Prev:
     4.80%, 4.88%, 4.58%, 4.28%, 3.98%,
     3.58%)
- JOLTS Job Openings: 10562.00 (Prev:
     11033.00, 10438.00, 10439.00, 10934.00,
     10073.00, 9209.00)k
- Participation Rate: 61.90% (Prev: 61.80%,
     61.60%, 61.60%, 61.70%, 61.70%, 61.60%)
- Initial Jobless Claims: 286.00 (Prev:
     205.00, 199.00, 290.00, 351.00, 353.00,
     419.00)k

GROWTH & OUTPUT
---------------
- Real GDP (YoY): 4.95% (Prev: 4.95%,
     4.90%, 12.23%, 12.18%, 12.18%, 0.40%)
- Retail Sales (MoM): -1.91% (Prev: 0.26%,
     1.70%, 0.74%, 0.71%, -1.12%, 0.55%)
- Industrial Production: 101.89 (Prev:
     102.29, 101.61, 100.02, 101.59, 101.11,
     100.10)
- Personal Income (MoM): 0.44% (Prev:
     0.44%, 0.45%, 0.17%, 1.11%, 0.13%,
     -1.95%)
- Personal Spending (MoM): 0.64% (Prev:
     0.64%, 1.33%, 0.83%, 0.27%, 1.00%,
     0.02%)
- Consumer Sentiment (UM): 70.60 (Prev:
     70.60, 67.40, 72.80, 70.30, 81.20,
     85.50)

HOUSING
-------
- Housing Starts: 1702.00 (Prev: 1679.00,
     1520.00, 1555.00, 1615.00, 1534.00,
     1643.00)k
- New Home Sales: 811.00 (Prev: 744.00,
     745.00, 800.00, 740.00, 708.00, 676.00)k

FINANCIAL CONDITIONS
--------------------
- 2Y Treasury Yield: 1.02% (Prev: nan%,
     nan%, 0.47%, 0.29%, 0.23%, 0.22%)
- 10Y Treasury Yield: 1.78% (Prev: nan%,
     nan%, 1.64%, 1.47%, 1.35%, 1.30%)
- 10Y-2Y Spread: 0.76% (Prev: nan%, nan%,
     1.17%, 1.18%, 1.12%, 1.08%)
- High Yield Spread: 3.39% (Prev: nan%,
     3.32%, 3.11%, 3.05%, 3.22%, 3.22%)
- Dollar Index (YoY): 2.95% (Prev: nan%,
     nan%, -1.45%, -3.22%, -2.73%, -3.92%)
- WTI Crude Oil (YoY): 64.10% (Prev: nan%,
     nan%, 120.47%, 84.94%, 58.77%, 76.24%)
```

Table 10: Macroeconomic indicators used in the information sets. All variables are constructed using vintage data available prior to each FOMC meeting.

Category	Indicator	Description
Inflation	Core PCE	Core personal consumption expenditures inflation (YoY)
	Headline PCE	Headline PCE inflation including food and energy (YoY)
	Core CPI	Core consumer price inflation (YoY)
	Headline CPI	Headline consumer price inflation (YoY)
	PPI Final Demand	Producer price inflation for final demand (YoY)
	Sticky Price CPI	Inflation of relatively sticky-price components (YoY)
	1Y Exp. Inflation	One-year-ahead inflation expectations (survey-based)
	10Y Breakeven	Market-implied 10-year inflation expectations
Labor Market	Non-farm Payrolls	Monthly change in total non-farm employment
	Unemployment Rate	Civilian unemployment rate
	Avg. Hourly Earnings	Average hourly earnings growth (YoY)
	JOLTS Openings	Job openings from the JOLTS survey
	Participation Rate	Labor force participation rate
	Jobless Claims	Initial unemployment insurance claims
Growth & Output	Real GDP	Real gross domestic product growth (YoY)
	Retail Sales	Monthly growth in retail and food services sales
	Industrial Production	Index of industrial production
	Personal Income	Monthly growth in personal income
	Personal Spending	Monthly growth in personal consumption expenditures
	Consumer Sentiment	University of Michigan consumer sentiment index
Housing	Housing Starts	New privately-owned housing starts
	New Home Sales	Sales of new single-family homes
Financial Conditions	2Y Treasury Yield	Nominal yield on 2-year U.S. Treasury securities
	10Y Treasury Yield	Nominal yield on 10-year U.S. Treasury securities
	10Y–2Y Spread	Treasury yield curve slope (10Y minus 2Y)
	High Yield Spread	Corporate high-yield bond spread
	Dollar Index	Broad U.S. dollar exchange rate index
	WTI Crude Oil	West Texas Intermediate crude oil price (YoY)
	VIX	CBOE volatility index
	S&P 500	Equity market performance (YoY)

```
- VIX: 31.16 (Prev: nan, nan, 15.24, 17.75,
    16.79, 17.20)
- S&P 500 (YoY): 13.00% (Prev: nan%, nan%,
    34.27%, 37.24%, 30.57%, 36.35%)

==========================================
BEIGE BOOK SUMMARY (Released: 2022-01-12)
==========================================

NATIONAL SUMMARY:

Overall Economic Activity
Economic activity across the United States
    expanded at a modest pace in the final
    weeks of 2021. Contacts from many
    Districts indicated growth continued to
    be constrained by ongoing supply chain
    disruptions and labor shortages.
    Despite the modest pace of growth,
    demand for materials and inputs, and
    demand for workers, remained elevated
    among businesses. Lending activity
    picked up slightly toward the end of
    the year, led by commercial real estate
    borrowers. Consumer spending continued
    to grow at a steady pace ahead of the
    rapid spread of the Omicron COVID-19
    variant. Most Districts noted a sudden
    pull back in leisure travel, hotel
    occupancy and patronage at restaurants
    as the number of new cases rose in
    recent weeks. Although optimism
    remained high generally, several
    Districts cited reports from businesses
    that expectations for growth over the
    next several months cooled somewhat
    during the last few weeks. The
    manufacturing sector continued to
    expand nationally, with some regional
    differences in the pace of growth.
    Overall activity in the transportation
    sector expanded at a moderate pace.
    While farm incomes were elevated
    throughout 2021, agricultural
```

```
    conditions were marred by drought
    conditions across several Districts.

Employment and Wages
Employment grew modestly in recent weeks,
    but contacts from most Districts
    reported that demand for additional
    workers remains strong. Job openings
    were up but overall payroll growth was
    constrained by persistent labor
    shortages. Tightness in labor markets
    drove robust wage growth nationwide,
    with some Districts highlighting
    additional growth in labor costs
    associated with non-wage benefits.
    While many contacts noted that wage
    gains among low-skill workers were
    particularly strong, compensation
    growth remained well above historical
    averages across industries, across
    worker demographics, and across
    geographies. Besides wage gains, many
    contacts indicated adjustments to job
    demands - such as accommodating
    part-time work or adjusting
    qualification requirements - to attract
    more applicants and retain existing
    workforces.

Prices
Contacts from most Federal Reserve
    Districts reported solid growth in
    prices charged to customers, but some
    also noted that price increases had
    decelerated a bit from the robust pace
    experienced in recent months. Wholesale
    and materials prices contributed to
    pricing pressures across a wide range
    of industries, spanning service
    providers and goods producers. Many
    contacts attributed the high cost of
    inputs to ongoing supply chain
    disruptions. Some Districts reported
    that transportation bottlenecks had
    stabilized in recent weeks, though
    procurement costs remained elevated.
    Ongoing labor shortages and associated
    wage growth also added cost pressures
    to businesses.

==========================================

Based on this information, what decision
    should the FOMC make?

Provide your answer in the following format:
Decision: [Hike/Hold/Cut]
Confidence: [High/Medium/Low]
Reasoning: [One sentence explanation citing
    specific numbers from the report]

Your prediction:
```

A.2.2 Prompt Structure of Direct Synthesis with Role-Specialized Agents

```
# Hawk Agent Prompt
You are a HAWKISH Fed policy maker who
    prioritizes price stability.

Meeting date: {meeting_date}
Current Rate: {current_rate}%

{economic_context}

Argue for your preferred policy decision
    from a hawkish perspective.
End with: My recommendation is
    [Hike/Hold/Cut].

--------------------------------------------

# Dove Agent Prompt
You are a DOVISH Fed policy maker who
    prioritizes employment.

Meeting date: {meeting_date}
Current Rate: {current_rate}%

{economic_context}

Argue for your preferred policy decision
    from a dovish perspective.
End with: My recommendation is
    [Hike/Hold/Cut].

--------------------------------------------

# Synthesis Prompt
Two FOMC members have debated the policy
    for {meeting_date}:

HAWK VIEW:
{hawk_opinion}

DOVE VIEW:
{dove_opinion}

As the synthesizer, make a final decision.
Respond in EXACTLY this format:
Decision: [Hike/Hold/Cut]
Confidence: [High/Medium/Low]
Reasoning: [Brief synthesis of both views]
```

A.2.3 Structured Multi-Agent Prompt Design of Multi-Round Deliberation with Debate and Voting

```
# Structured Multi-Agent Prompt
    Specification

## Agent 1: Hawk (Christopher)
Priority:
- Price stability, data over ideology

Beliefs:
- Inflation is sticky; premature easing
    risks repeating 1970s mistakes
```

```
Default Priors:
- If Core CPI > 2.5% and trend is
    flat-to-rising -> lean HOLD or HIKE

Bias-Reversal Triggers:
- If Real Rate > r* + 0.5pp AND Core CPI
    falling for 3 consecutive months
  -> MAY support CUT

Historical Citation Requirement:
- Format: [Year/period; inflation x%;
    unemployment y%; policy rate z%]
  -> [Fed action] -> [Outcome]
- State 2 similarities and 1 key difference
    vs. current context

Guiding Principle:
"Persona is a starting point, NOT a
    constraint. Follow the evidence."

-----------------------------------------------

## Agent 2: Dove (Austan)
Priority:
- Maximum employment, data over ideology

Beliefs:
- Monetary policy works with lags; act
    preemptively against downturn risk

Default Priors:
- If unemployment rising or labor broadly
    softening -> lean CUT

Bias-Reversal Triggers:
- If inflation > 3% with acceleration AND
    labor remains tight
  -> MAY support HIKE

Historical Citation Requirement:
- Same format as Hawk
- State 2 similarities and 1 key difference

Guiding Principle:
"Persona is a starting point, NOT a
    constraint. Follow the evidence."

-----------------------------------------------

## Agent 3: Chair (Jerome)
Role:
- Neutral moderator, median voter,
    consensus builder

Data Discipline:
- Use only data available as of meeting
    month T (vintage-aware)
- Metrics: Core CPI YoY, U-3 unemployment,
    wages, FCI, breakevens
- Real Rate = Policy Rate - 1y expected
    inflation
- Neutral rate r*: 0.5 - 1.0%

Decision Algorithm:
1. Ideological Bound (Median Voter
    Feasibility)
2. Asymmetric Risk Protocol
3. Consensus Momentum
```

```
# Multi-Round Interaction Protocol

Round 1: Blind Thinking
Input:
- ECONOMIC DATA: {economic_context}
- HISTORICAL PRECEDENTS: {rag_context}

Output Format (EXACT):
Reasoning: ...
Vote: [Hike/Hold/Cut]
Magnitude: [+25bp / +50bp / +75bp / -25bp /
    -50bp / 0bp]

-----------------------------------------------

Round 2: Cross-Examination
Inputs:
- Your initial stance: "{my_opinion}"
- Opponent's argument:
    "{all_other_opinions}"

Task:
- Critically analyze and refute opponent
    arguments
- Defend your position

-----------------------------------------------

Round 3: Final Vote
Inputs:
- YOUR INITIAL POSITION
- YOUR REBUTTAL
- OPPONENT'S REBUTTAL

Output Format (EXACT):
Reflection: ...
Vote: [Hike/Hold/Cut]
Magnitude: [+25bp / +50bp / +75bp / -25bp /
    -50bp / 0bp]
```

A.3 Additional Experimental Results

This appendix reports additional experiments designed to examine whether augmenting the base input with advanced prompting and retrieval strategies can improve the policy decision accuracy of `gpt-4o`, particularly by mitigating its tendency toward hold-dominant predictions.

We first explore prompt-level interventions that explicitly encourage structured reasoning or learning from precedent. These include chain-of-thought (CoT) prompting and few-shot learning with both fixed and dynamically selected examples.

The CoT prompting setup explicitly instructs the model to reason step by step over quantitative indicators and qualitative Beige Book information before producing a policy decision. The full system

prompt used for this experiment is shown below.

```
You are an expert FOMC policy analyst.
Your task is to predict the Federal
    Reserve's interest rate decision based
    on economic data.

CRITICAL INSTRUCTIONS:
- Base your analysis ONLY on the provided
    economic data.
- Do NOT invent statistics or cite events
    not mentioned in the data.
- Consider both the dual mandate: Price
    Stability (2% inflation target) and
    Maximum Employment.
- If BEIGE BOOK data is provided, you MUST
    incorporate insights from "Districts"
    and "Contacts" in your analysis.

THINK STEP BY STEP before making a decision.

Output format (IMPORTANT: Thinking comes
    BEFORE Decision):
Thinking: [Step-by-step analysis:
  1) Quantitative: What do
    inflation/employment numbers suggest?
  2) Qualitative: What do Beige Book
    reports from Districts/Contacts
    indicate? (if provided)
  3) Synthesis: Weighing both sources,
    which risk is greater?]
Decision: [Hike/Hold/Cut]
Confidence: [High/Medium/Low]
```

In addition, we consider few-shot learning setups in which historical FOMC-like examples are provided to guide the model's decision. We evaluate both fixed example sets and dynamically selected examples based on similarity to the current meeting context. The corresponding prompt template is provided below.

```
You are a member of the FOMC.
Your task is to predict the policy decision
    by learning from HISTORICAL PRECEDENTS.

STRATEGY:
1. Analyze the provided 'Historical
    Examples'. Observe the relationship
    between economic data and the decision.
2. Compare the 'Current Situation' with
    those examples quantitatively.
3. If BEIGE BOOK data is provided,
    incorporate qualitative insights from
    "Districts" and "Contacts".
4. If the current situation resembles a
    past 'Hike' scenario (similar
    inflation, unemployment), recommend a
    Hike.

CRITICAL INSTRUCTIONS:
- Base your analysis ONLY on the provided
    data and historical examples.
- Do NOT invent statistics or cite events
    not mentioned in the data.
- Explicitly compare current numbers with
    historical precedent numbers.

Output format (IMPORTANT: Thinking comes
    BEFORE Decision):
Thinking: [Compare current data with
    history. Include Beige Book insights if
    provided.]
Decision: [Hike/Hold/Cut]
Confidence: [High/Medium/Low]
```

Across input configurations E1–E6, neither CoT prompting nor few-shot learning consistently improves accuracy relative to the baseline zeroshot setting. In some cases, these techniques appear to reinforce conservative behavior rather than correcting it.

We further investigate whether providing access to historically similar policy episodes via retrieval-augmented generation can help the model learn policy consistency and improve decision accuracy.

Retrieval design. For each test meeting, we retrieve past FOMC meetings with similar macroeconomic conditions, focusing on inflation, unemployment, and the prevailing policy rate. To prevent look-ahead bias, meetings corresponding to the same date are explicitly excluded from retrieval.

RAG variants. We consider two RAG implementations:

- **RAG (summary).** Retrieved meetings are summarized by an LLM into short textual descriptions (e.g., "inflation was elevated"), which are then provided to the prediction model along with a prompt that strictly encourages adherence to past decisions. As observed in the baseline experiments, this abstraction leads to information loss: semantic summaries obscure precise numerical conditions, resulting in degraded retrieval quality and lower predictive performance.

- **RAG (structured).** Retrieved cases are provided in a structured numerical format (e.g., `Core PCE at X%, Unemployment at Y%`), and the prompt allows the model to reference past cases without enforcing strict imitation. Using raw numerical values improves similarity matching at the vector search stage, and the more flexible prompt enables the model

to override historical precedents when warranted.

ChromaDB serves as the vector database with persistent local storage. Source documents (PDF files of FOMC minutes and materials) are processed using pypdf for text extraction, then chunked into 800-word segments with 100-word overlap. Each chunk is annotated with metadata including source filename, year, month, and document type. Semantic similarity search is enabled through OpenAI's text-embedding-3-small embeddings. The retrieval process incorporates three enhancements: (1) Query Enrichment - generating qualitative summaries via gpt-4o-mini combined with quantitative indicators (inflation, unemployment, policy rate); (2) Time-Travel Fix - restricting retrieval to documents dated before the target meeting to prevent data leakage; (3) Diversity - deduplicating results by month to return up to 3 precedents from distinct time periods.

Illustrative Examples of Hold-Convergent Reasoning To provide qualitative intuition for the chain-of-thought results, we include two stylized examples in which the model is exposed to conflicting hawkish and dovish signals. In both cases, the Chair explicitly acknowledges the validity of both sides yet ultimately converges on *Hold*. These examples are illustrative rather than causal evidence, but they are consistent with the quantitative pattern in Table 7: structured reasoning can make the policy conflict more explicit without shifting the final action away from the status quo.

Case 1: Rising inflation vs. higher unemployment. **Hawk:** Near-term inflation expectations remain elevated, and signaling premature easing could undermine the central bank's anti-inflation credibility.
Dove: Household income and spending have begun to soften, and the unemployment rate has risen to 4.1%, indicating emerging labor-market slack.
Chair: Both concerns are credible: inflation expectations still warrant caution, but weakening demand and a softer labor market argue against further tightening. With longer-run expectations still relatively anchored, maintaining the current restrictive rate while monitoring incoming data is the most prudent course. **Decision: Hold.**

Case 2: Rising inflation vs. recession risk. **Hawk:** Near-term inflation expectations remain elevated, and easing too early could reignite price pressures and send an overly dovish signal to markets.
Dove: Real activity indicators are deteriorating: consumer spending is slowing, business investment is weakening, and recession risk is rising.
Chair: The inflation outlook argues against premature easing, while the deterioration in real activity cautions against additional tightening. Given these offsetting risks, leaving the policy rate unchanged preserves anti-inflation credibility without further increasing downside growth risks. **Decision: Hold.**