Generative AI for Central Bank Statements: Large Language Models as Monetary Policy Forecasters

Jens Hilscher[a], Yevgeny Mugerman[b], Alon Raviv[c]

June 2026

Abstract

We study whether leading public large language models can forecast central bank policy statements in real time. We generate and archive 140 time-stamped forecasts from five public LLMs during the weeks leading up to three December 2025 policy meetings at the Federal Reserve, the ECB, and the Bank of England. We evaluate forecasts along three dimensions: the realized policy decision, the accuracy of the models' probability forecasts, and the hawkish dovish communication tone of the projected statements. Communication tone is measured using prespecified rubric that captures inflation language, labor market language, risk balance, forward guidance, and voting disagreement. Across the full sample, individual model date accuracy is 87.9 percent, while the equal weight consensus is correct on 92.9 percent of forecast dates and improves probabilistic accuracy by lowering the mean multiclass Brier score from 0.227 to 0.201. The main result is that models agree on the likely rate decision earlier than they agree on the communication surrounding that decision. This matters for finance because markets respond not only to policy actions, but also to the language used to explain them. Generative AI therefore provides a richer ex ante panel of probabilities and projected statements around low frequency monetary policy events.

Keywords: *Large Language Models; Monetary Policy; Central Bank Communication; Forecast Combination; Text-as-Data.*
JEL Codes: *E52, E58, G14, C53, C55*

This project was supported by the David Goldman Data-Driven Innovation Research Center.

[a] University of California, Davis (jhilscher@ucdavis.edu). [b] Graduate School of Business Administration, Bar-Ilan University, and the Federmann Center for the Study of Rationality, The Hebrew University of Jerusalem (mugi@huji.ac.il). [c] Graduate School of Business Administration, Bar-Ilan University (Alon.Raviv@biu.ac.il).

1. Introduction

Central bank announcements are information events for financial markets. A large event-study literature shows that asset prices respond not only to the contemporaneous policy action but also to information conveyed about the economic outlook and the expected path of future policy (Kuttner, 2001; Bernanke and Kuttner, 2005; Gürkaynak, Sack, and Swanson, 2005). The post-meeting statement is therefore more than a record of the decision. It is the standardized narrative through which the committee communicates its interpretation of incoming data, its assessment of risks, and its guidance about how policy may evolve.

This paper asks whether leading public large language models (LLMs) can forecast central bank communications in real time. Specifically, can an LLM generate, before a meeting occurs, a projected policy statement that resembles what the central bank ultimately publishes? The question is distinct from forecasting the policy rate alone. A plausible projected statement must get the policy action right, but it must also match the institution's language conventions and convey a coherent rationale, risk assessment, and guidance. Forecasting the full statement is therefore a joint prediction problem over a discrete policy action and a high-dimensional text object.

The question is especially relevant in finance because central bank communication is economically important yet empirically sparse. For any one institution, there are only a small number of scheduled policy meetings in a year, even though each meeting can move interest rates, exchange rates, equity prices, and volatility. This creates a natural data problem for statement-level forecasting and for the study of narrative variation in monetary policy communication. Generative AI can help close part of this gap. It does not create new realized policy decisions, but it can generate a dense ex ante panel of time-stamped probabilities and projected statements around each meeting. That richer panel makes it possible to study not only whether the eventual action was anticipated, but also how alternative narrative framings emerge, converge, and disperse before the announcement. In that

sense, LLMs are useful not only for forecasting policy decisions, but also for creating a richer real time panel of projected statements that helps measure uncertainty about future central bank communication.

This paper should be read as a real time forecast evaluation and proof of concept rather than as a broad claim about all central bank meetings or all LLM forecasting settings. The December 2025 meeting set allows us to examine whether LLMs can anticipate both policy decisions and the language used to communicate them. It also allows us to study whether cross model agreement and disagreement are informative before the official announcement. In particular, disagreement across projected statements may provide a useful measure of ex ante uncertainty about the tone, guidance, risk assessment, and committee balance of the eventual communication.

Our empirical design treats each LLM as a forecaster. We run a standardized prompt (Appendix A) across five publicly available LLM products (ChatGPT, Claude, Gemini, Grok, and Perplexity) repeatedly during the pre-meeting window. The prompt is discipline-imposing. It requires the model to declare a precise data-cutoff timestamp in U.S. Eastern Time, provide transparent sourcing for numerical claims, and separate verified data from forecasts in a structured evidence table. It also requests an explicit probability distribution over discrete policy outcomes and a draft statement written in the institutional voice, together with a short list of proposed changes relative to the previous official statement.

Chronological integrity is a central challenge in evaluating LLM forecasts. A purely retrospective exercise that prompts a model after a meeting has occurred risks contamination: the model, or its retrieval tools, may implicitly rely on information that entered the public record after the evaluation date. Recent work emphasizes the importance of chronological consistency when LLM outputs are used for prediction or inference and proposes methods to reduce look-ahead problems (He, Lv, Manela, and Wu, 2025; Ludwig, Mullainathan, and Rambachan, 2025). Our contribution is complementary. We preserve the real time nature of the exercise by prompting models before the meetings, time stamping each output, and requiring models to state their information cutoff.

We evaluate each forecast against the policy decision that was ultimately announced by the central bank and the text of the official statement. Our sample covers three December 2025 meetings for which we collected real-time forecasts and the corresponding official releases: the FOMC statement issued on December 10, 2025, which lowered the target range by 25 basis points; the ECB monetary policy decisions release dated December 18, 2025, which left the key rates unchanged; and the BoE Monetary Policy Summary published on December 18, 2025, which reports a 5–4 vote to reduce Bank Rate by 25 basis points. The sample yields 140 model-date observations across different forecast dates.

We evaluate forecasts along three dimensions. First, we assess policy-step accuracy: whether the model predicts the realized discrete policy action. Second, when models report probabilities for each possible policy outcome, we evaluate probabilistic accuracy using the multiclass Brier score, where lower values indicate better calibrated probability forecasts. Third, and centrally, we quantify the policy communication tone of each projected statement using a prespecified hawkish dovish rubric that maps statement language into an index on a 1 to 5 scale. The rubric captures not only the projected rate action, but also how the statement describes inflation, labor market conditions, the balance of risks, forward guidance, and voting disagreement.

Our main findings are as follows. The ECB is the easiest case in our sample: the equal-weight consensus selects a hold from the first recorded date and remains there throughout, with an average realized-outcome probability of 74.8 percent. The BoE cut is also recognized early, but the consensus probability on the realized cut averages only 63.7 percent, and the projected vote split varies materially across models. The FOMC cut is the hardest case. The consensus initially leans toward a hold, misses the realized action on two early dates, and converges only on November 24, 2025; by December 9, however, it assigns 84.4 percent probability to the realized 25 basis point cut. Across the full sample, the equal-weight consensus improves date-level accuracy from 87.9 percent at the model-date level to 92.9 percent and lowers the mean multiclass Brier score from 0.227 to 0.201.

A central result is that narrative convergence is slower than policy-step convergence. Even when models begin to agree on the most likely rate action, they continue to differ in how they characterize inflation persistence, labor-market slack, the balance of risks, and the appropriate strength of forward guidance. The hawkish dovish analysis is therefore central to the paper. It captures information in the statement that goes beyond the rate decision, including the central bank's guidance about future policy. Cross-model dispersion in tone, vote language, and guidance is informative about the degree of pre-meeting communication uncertainty that remains after the modal rate call has largely converged.

This paper contributes to literature in three ways. First, we provide a prospective, real-time evaluation of LLM-generated central bank statements. Existing work in monetary economics and finance uses text methods to measure sentiment and semantic content in realized communications (for example, Lucca and Trebbi, 2009; Hansen and McMahon, 2016), and a growing literature applies transformer models and LLMs to classify and summarize central bank texts (Hilscher, Nabors and Raviv, 2026; Gössi et al., 2023; Kim et al., 2024). We instead treat the next statement itself as the object to be forecast and evaluate projected language against the realized release.

Second, we develop a transparent real time prompting protocol. Each forecast is generated before the meeting, archived with a timestamp, and accompanied by a declared information cutoff, source references for numerical claims, and a separation between verified facts and forecasts. These features address a key barrier to integrating LLM outputs into research designs: without careful documentation of the information available at the time of each forecast, it is difficult to know whether the model is genuinely forecasting or relying on information that became available later.

Third, we make a finance-specific methodological contribution by using generative AI to study an inherently low-frequency empirical object. Central bank communication offers relatively few meeting level observations, but each announcement is economically important, and the statement's language affects asset prices by shaping expectations about future policy. Generative AI helps close part of this gap by producing a dense ex ante panel of probabilities and projected statements around each event. That panel allows researchers to study convergence, disagreement, and narrative uncertainty

at a much richer frequency than the meeting calendar alone would permit. Our hawkish dovish rubric and cross model comparisons turn projected statement language into measurable differences in policy tone, forward guidance, and internal committee balance.

The results do not imply that LLMs can replace central bank committees. Official statements reflect internal deliberation, judgment, strategic communication, and institutional drafting conventions that are not fully observable from public data. Instead, LLM forecasts are useful as benchmarks for what could reasonably be predicted before the meeting from public information.

The remainder of the paper proceeds as follows. Section 2 reviews related work on monetary policy communication, text-as-data methods, and LLMs in economics and finance. Section 3 describes the data and research design. Section 4 presents the results for the December 2025 meeting set. Section 5 discusses limitations. Section 6 concludes.

2. Related Literature

This paper contributes to four related literatures: monetary policy announcements and central bank communication, quantitative text analysis in economics and finance, the use of large language models in economic and financial prediction, and forecast combination. Our distinctive feature is that we use LLMs prospectively. We prompt models in real time, with an explicit information cutoff, to generate the next policy statement before it is released, rather than applying language tools retrospectively to already published text.

A first literature studies the market effects of monetary policy announcements. High frequency event studies show that asset prices respond not only to the policy rate decision, but also to the information conveyed about future policy and the economic outlook. Kuttner (2001) identifies policy surprises using federal funds futures, while Bernanke and Kuttner (2005) document equity market responses to unexpected policy actions. Gürkaynak, Sack, and Swanson (2005) show that markets react separately to the target rate surprise and to the path or communication component of the announcement. These studies motivate treating the post meeting statement as part of the policy event, not merely as a description of the rate decision.

A second literature measures central bank communication directly using text methods. Lucca and Trebbi (2009) develop an automated measure of FOMC statement content, and Hansen and McMahon (2016) show that central bank language has macroeconomic and financial effects beyond the contemporaneous policy instrument. Ehrmann and Talmi (2020) further show that semantic similarity in central bank press releases affects market volatility, highlighting the importance of wording changes. Our paper builds on this literature but shifts the timing of the question. Instead of measuring realized communication after the fact, we ask whether the next communication can be forecast before the meeting.

More broadly, the paper relates to the text as data literature in economics and finance. Gentzkow, Kelly, and Taddy (2019) provide the general methodological foundation for converting text into quantitative variables. In finance, Tetlock (2007) shows that media tone contains information relevant for returns, while Loughran and McDonald (2011) demonstrate the importance of domain specific dictionaries for measuring financial tone.

Baker, Bloom, and Davis (2016) use newspaper text to construct a policy uncertainty index. Recent work also applies language based tools to financial disclosure, including Lesmy, Muchnik, and Mugerman (2025) on the readability of 10 K filings and Muchnik, Mugerman, Noiman, and Sade (2025) on LLM based measures of disclosure readability. Our contribution differs because the text being evaluated is projected rather than realized.

A fourth and more recent literature applies transformer models and LLMs to central bank communication. Hilscher, Nabors, and Raviv (2026) construct a sentiment measure for published Federal Reserve and ECB communications using FinBERT and show that central bank sentiment contains information for policy rates and Taylor rule variables. Gössi, Chen, Kim, Bermeitinger, and Handschuh (2023) introduce a FinBERT based model for FOMC minutes, while Kim, Spörer, Lee, and Handschuh (2024) compare different model classes for sentiment detection in central bank statements. Taskin and Akal (2025) apply BERT based methods to FOMC minutes during the COVID period. Silva, Moriya, and Veyrune (2025) use a fine tuned LLM to classify central bank communications across a large multilingual dataset, and Kanelis, Kranzmann, and Siklos (2025) use LLM methods to separate monetary policy and financial stability sentiment in FOMC minutes. These studies show the value of modern language models for measuring central bank communication. We instead examine whether LLMs can forecast the next official statement itself.

The closest forecasting related paper is Kazinnik and Sinclair (2025), who develop a multi agent framework for FOMC decision making that combines an LLM based simulation with a Bayesian voting model. Their focus is committee decision making and counterfactual policy outcomes. Our focus is different. We treat the full official statement as the forecast object and evaluate whether model generated statements anticipate not only the rate decision, but also the language used to justify it.

Our design also connects to methodological literature using LLMs for prediction and inference. Recent methodological work highlights the challenges of using LLMs for economic prediction. Ludwig, Mullainathan, and Rambachan (2025) distinguish between using LLMs to predict future outcomes and using them to measure text or other economic concepts, and emphasize that predictive evaluation must avoid information leakage from the future.

Crane, Karra, and Soto (2025) show that LLMs can produce unstable or systematically incorrect answers to macroeconomic questions, which reinforces the need for careful evaluation. Glasserman and Lin (2023) discuss look ahead bias in LLM based financial prediction, while He, Lv, Manela, and Wu (2025) study methods for making language models more chronologically consistent. Our design addresses these concerns directly by prompting models before the policy meetings, archiving each output with its generation time, requiring an explicit information cutoff, and evaluating forecasts only against statements released later.

Finally, the paper relates to the forecast combination literature. Bates and Granger (1969) show that combining forecasts can improve accuracy when individual models contain complementary information. Stock and Watson (2004) and Timmermann (2006) show that simple forecast combinations are often strong benchmarks in macroeconomic forecasting. In our setting, forecast combination is useful because different LLMs may rely on different information sources and may interpret the same pre meeting evidence differently.. The equal weight consensus therefore serves both as a forecast and as a measure of cross model agreement about the likely decision and communication tone.

3. Data and Design

This section describes the empirical design used to evaluate LLM forecasts of central-bank policy communications. The design is prospective rather than retrospective: prompts are issued before the meetings occur, outputs are archived as generated, and evaluation is conducted against the realized official communications. This is important in the present setting because the underlying empirical object is low-frequency. Each central bank holds only a small number of scheduled meetings per year, yet each meeting produces economically meaningful information for financial markets. Our approach therefore turns three realized policy events into a denser panel of time-stamped ex ante forecasts and projected statements.

3.1. Sample, meetings, and benchmark texts

Our empirical design evaluates forecasts that are generated and archived before the policy meetings and then compared with the actual decisions and official statements released afterward. We study three December 2025 meetings: the Federal Open Market Committee (FOMC) meeting of December 9–10, 2025, with the statement released on December 10, 2025; the European Central Bank (ECB) Governing Council meeting of December 18, 2025, with the monetary policy decisions release issued the same day; and the Bank of England (BoE) Monetary Policy Committee (MPC) meeting ending on December 17, 2025, with the Monetary Policy Summary and Minutes published on December 18, 2025.

The realized policy outcomes differ across institutions. The FOMC lowered the target range by 25 basis points. The ECB left its key policy rates unchanged. The BoE reduced Bank Rate by 25 basis points, and the published Minutes report a 5-4 vote. For comparability across institutions, we map policy outcomes into a common five-point discrete grid,

$$\mathcal{K} = \{-50, -25, 0, +25, +50\},$$

where the entries denote basis-point changes relative to the policy rate in place immediately before the meeting. For the ECB, this grid refers to the common movement in the key policy rates, which move jointly in the meeting studied here. For the BoE, we additionally record projected and realized vote splits when they are reported, but the primary forecast object remains the bank rate decision.

3.2. LLM forecasters and real-time prompting

We treat each LLM as a distinct forecaster. The model set comprises five publicly available products: ChatGPT, Claude, Gemini, Grok, and Perplexity, as available during the forecast window. We do not fine-tune any model, and we do not regenerate outputs after the meeting outcome is known. Instead, on each forecast date we submit the same standardized prompt to each model, varying only the institution-specific inputs.

The dataset contains 140 model-date observations: 45 for the FOMC, 55 for the ECB, and 40 for the BoE. These correspond to 9 FOMC forecast dates, 11 ECB forecast dates, and 8 BoE forecast dates. Although we archive all model outputs, some forecasts cannot be used for every metric. For example, a forecast can be used to evaluate the predicted rate decision, but not the Brier score, if it does not report a complete probability distribution.

Timing is central to our design. We save every prompt and every model response with the exact date and time it was produced. This shows that the forecasts were made before the official statement was released and helps prevent the forecasts from being influenced by information that became available later. Appendix A reproduces the FOMC prompt verbatim. The ECB and BoE prompts follow the same structure and differ only in institution-specific details such as the policy instrument, wording conventions, and meeting dates.

3.3. Prompt protocol

The prompt is designed to make model outputs interpretable, auditable, and as comparable as possible across products. It imposes four disclosure requirements. First, the model must declare a precise information cutoff timestamp in U.S. Eastern Time. Any statement referring to information beyond that cutoff must be labeled explicitly as a forecast or preview.

Second, the model must disclose the sources used for numerical claims, including document titles and page or section references. This requirement does not guarantee perfect traceability, but it makes retrieval differences across products more observable and discourages unverifiable numeric assertions.

Third, the model must separate verified data from forecasts in a two-column evidence table. Each row identifies the series, the reference period, the value, and the corresponding source reference. This structure is intended to distinguish the model's view of the current information set from its forecasted interpretation of that information.

Fourth, the model must report an explicit probability distribution over the five discrete policy outcomes in $\mathcal{K}$. The prompt also asks the model to explain the main drivers of the distribution, describe the implicit reaction function it is using, and run internal consistency checks.

Beyond these disclosure rules, the prompt requests a projected policy statement written in the institutional voice, a short "changes versus previous statement" section, a validation log, and a one-sentence confidence rating. The prompt is therefore discipline-imposing but not mechanically restrictive: it standardizes disclosure and output structure without forcing models to reason in the same way.

3.4. Equal-weight cross-model consensus

In addition to individual model forecasts, we construct a simple cross-model consensus. The rationale is the standard forecast-combination intuition: models may contain partially independent information and may make idiosyncratic errors, so averaging can improve robustness.

For probabilities, the consensus assigns each outcome k the equal-weight average probability across models with valid distributions on date t:

$$\bar{p}_{bmt}(k) = \frac{1}{|\mathcal{J}_{bmt}|} \sum_{j \in \mathcal{J}_{bmt}} p_{jbmt}(k)$$

where $\mathcal{L}_{bmt}$ is the set of models with valid probabilities. The consensus point forecast is the outcome with the highest $\bar{p}_{bmt}$(k). For each central bank, meeting, forecast date, and possible policy outcome, we calculate the consensus probability as the simple average of the probabilities reported by the models. For example, if five models assign probabilities of 70, 70, 40, 60, and 33 percent to a 25 bp cut, the consensus probability for a 25 bp cut is $(70 + 70 + 40 + 60 + 33)/5 = 54.6$percent. We repeat this calculation for each possible outcome. The consensus point forecast is the outcome with the highest average probability. If a model does not provide a valid probability distribution, the average is calculated only over the models with valid probabilities on that date.

3.5. Evaluation metrics

We evaluate forecasts along three dimensions.

Policy-step accuracy. Let a_{jbmt} denote model j's predicted discrete action and a_{real} the realized action. Policy-step accuracy is the indicator $1[a_{jbmt} = a_{real}]$. For the BoE, we also record the predicted vote split when provided and compare it to the realized split, but we treat the policy rate decision as the primary outcome.

Probabilistic accuracy. When a model reports probabilities $p_{jbmt}(k)$ over the five outcomes, we evaluate them using proper scoring rules. Our baseline is the multiclass Brier score,

$$\text{Brier}_{jbmt} = \sum_{k} (p_{jbmt}(k) - y_{bm}(k))^2$$

where $y_{bm}(k) = 1$ for the realized outcome and zero otherwise. Lower values indicate better probabilistic calibration. We compute the same scores for the consensus distribution $\bar{p}_{bmt}(k)$.

Hawkish–dovish tone. Hawkish-dovish tone is measured using a rule-based, eight-criterion rubric described in Appendix B. Each criterion is evaluated on a 1-5 scale with substantive; the final tone index is the equal-weight average across the eight criterion scores. Appendix B also defines within-date tone bias as the deviation of a model's tone score from the cross-model mean on the same bank-date pair and provides the code used to aggregate criterion-level scores into the reported tables.

3.6. Replicability and scope conditions

The design is transparent at the prompt-and-output level. Every archived forecast includes the full prompt, the time of the run, the declared data cutoff, and the raw model response. When available, we also retain the model version displayed by the platform. Forecasts are therefore interpreted as products of the public system available on that specific date, not as timeless properties of an abstract model family.

The real-time protocol eliminates the most direct form of look-ahead bias that arises when models are prompted only after the realized statement has entered the public record. At the same time, the design does not attempt to equalize all underlying retrieval environments across platforms. Some products may have built-in browsing or richer live retrieval, while others may rely more heavily on parametric training data. Rather than assuming away these differences, our protocol makes them more visible by requiring explicit cutoffs, source disclosure, and a separation between verified data and forecasted content.

Finally, the design should be interpreted as a disciplined forecast-evaluation framework rather than as an attempt to estimate a structural monetary-policy rule from a large panel. The number of realized meetings is necessarily small. That is precisely why the real-time panel of LLM outputs is useful: it adds rich ex ante variation around low-frequency policy events while preserving a transparent link to the realized decision and statement.

4. Results

This section reports the performance of the large language model (LLM) forecasts and the equal-weight cross-model consensus for the December 2025 meeting set. We evaluate forecasts against the realized policy decisions and published communications for three meetings: the FOMC meeting of December 9–10, 2025 (25 bp cut), the ECB Governing Council meeting of December 18, 2025 (hold), and the Bank of England Monetary Policy Committee meeting ending on December 17, 2025 and published on December 18, 2025 (25 bp cut to 3.75 percent on a 5–4 vote). Figure 1 (Panels

A through C) traces the probability that each model, and the equal-weight consensus, assigns to the realized outcome over the pre-meeting window. Table 1 reports selected probability distributions for the three central banks – ECB, Fed and the BoE. Tables 2 and 3 summarize the hawkish–dovish tone index computed from the projected statements using the pre-specified rubric in Appendix B. Table 4 then turns to the BoE case, where disagreement persists not only over the policy step, but also over the perceived balance of the Committee and the language used to justify the decision. Taken together, the results reveal substantial heterogeneity across institutions. The ECB is the cleanest case in the sample, with stable early agreement on the realized hold. The FOMC is the most difficult case, as the consensus initially leans toward a hold before converging to the realized cut. The BoE lies between those extremes: the realized cut is recognized relatively early, but the meeting remains a close call in both probability space and narrative framing.

4.1. Policy decision forecasts and convergence

Table 1 reports probability distributions over five discrete outcomes (cut 50 bp, cut 25 bp, hold, hike 25 bp, hike 50 bp) for selected dates in the pre-meeting window. Figure 1 complements Table 1 by tracing, for each model and for the consensus, the probability assigned to the policy decision that was ultimately realized. It also shows how much uncertainty remains before the meeting by revealing whether models still assign substantial probability to the main alternative outcome, such as a hold when the realized decision is a 25 bp cut. The key empirical distinction is between settings in which the public information set points clearly to one outcome and settings in which the same information can still rationalize multiple plausible decisions.

Panel A of Figure 1 and Panel A of Table 1 show that the FOMC meeting is the hardest case in the sample. On November 20, 2025, the equal-weight consensus still assigns more probability to a hold (54.0 percent) than to the realized 25 bp cut (40.2 percent). Three models (ChatGPT, Grok, and Perplexity) remain hold-leaning, while Claude and Gemini already place the highest probability on the cut. By December 9, 2025, however, all five

models assign the highest probability to the realized 25 bp cut, and the consensus probability on that outcome rises to 84.4 percent, with only 12.6 percent remaining on a hold. The FOMC case therefore illustrates late convergence after a materially ambiguous early window.

Panel B shows that the ECB is the clearest case. The modal forecast is a hold from the beginning of the pre-meeting window, and the selected dates in Table 1 make that stability transparent. On November 24, 2025, model hold probabilities range from 55 to 75 percent and the consensus assigns 67.0 percent probability to unchanged rates. By December 17, 2025, the day before the meeting, all models still select a hold and the consensus hold probability rises to 79.8 percent. Relative to the FOMC and BoE, disagreement in the ECB case is concentrated more in the tails than in the modal decision itself.

Panel C places the BoE between those two extremes. The realized 25 bp cut is recognized relatively early, but confidence remains moderate and the projected internal balance of the Committee varies across models. On November 19, 2025, the consensus assigns 54.6 percent probability to the realized cut and 41.4 percent to a hold. Gemini and Perplexity still lean toward holding, while ChatGPT, Claude, and Grok favor a cut. By December 8, 2025, all five models place the highest probability on a 25 bp cut, but the consensus still retains 27.4 percent probability on a hold, which is materially larger than in the ECB case and consistent with a genuinely close meeting.

The point predictions reported in the “Most likely” column are therefore more uniform than the full distributions. For the ECB, every model and the consensus select the realized hold on the reported dates. For the FOMC and the BoE, by contrast, the more informative variation lies in how much probability remains on the main alternative even after the modal outcome shifts to the realized action. For monetary policy communication, that residual mass is economically meaningful because it captures how close the decision still appears in real time.

4.2. Probabilistic accuracy and the role of the ensemble

The probability distributions in Table 1 support two related conclusions. First, when the policy decision is well signaled in the pre-meeting information set, as in the ECB case, models rapidly concentrate probability on the realized outcome and preserve that ranking throughout the window. Second, when the decision is more finely balanced, as in the FOMC and especially the BoE, meaningful probability mass remains on nearby alternatives even late in the pre-meeting period.

Across all 140 model-date observations, top-event accuracy is 87.9 percent at the individual-model level. The equal-weight consensus improves date-level accuracy to 92.9 percent and lowers the mean multiclass Brier score from 0.227 to 0.201. The selected dates in Table 1 illustrate the same pattern in a compact way. The consensus multiclass Brier score falls from 0.651 to 0.041 across the two reported FOMC dates, from 0.168 to 0.057 across the two ECB dates, and from 0.379 to 0.178 across the two BoE dates. These declines are economically intuitive: as the meeting approaches and the information set becomes richer, the probability distribution moves closer to the realized outcome.

The ensemble improves robustness by attenuating idiosyncratic forecasts, but it does not mechanically dominate the best individual model on every date. The FOMC panel makes this especially clear. Early in the window, hold-leaning forecasts dilute the information contained in the cut-leaning models, so the consensus remains on the wrong side of the decision before eventually converging. In the ECB panel, by contrast, the ensemble largely stabilizes an already strong common signal. The BoE case is again intermediate: the consensus improves robustness, but it deliberately preserves substantial hold risk because several models continue to view the meeting as a narrow and contingent call. In that sense, the ensemble's value is not that it eliminates disagreement, but that it summarizes the center of a distribution of plausible ex ante views.

4.3. Hawkish–dovish tone and narrative heterogeneity

The hawkish–dovish analysis is economically important because markets price not only the current policy step, but also the communicated path of policy, the balance of risks, and the strength of forward guidance. Because this tone measure is criterion-by-criterion rather than a black-box sentiment classifier, differences in tone can be traced to specific language about inflation, labor markets, risk balance, financial conditions, forward guidance, and dissents; Appendix B provides the worked examples. Two projected statements can therefore agree on the same rate decision and still differ materially in the information they convey to market participants. We accordingly treat textual stance as a separate forecast object rather than as a stylistic add-on.

Tables 2 and 3 report the hawkish–dovish tone index for projected FOMC and ECB statements. The index is the equal-weight average of eight rubric components defined in Appendix B and summarizes how the projected statements frame inflation, labor-market conditions, risk balance, financial conditions, forward guidance, and voting agreement. Three patterns stand out.

First, tone dispersion remains even when models increasingly agree on the policy action. In the FOMC panel, the cross-model range is 0.49 points on November 20, 2025, widens to 0.61 on November 24, and remains elevated at 0.62 on December 1, even though the modal policy call is moving toward the realized cut. Only by December 9 does the range be compressed materially, to 0.16. Perplexity is the most hawkish model on average in the FOMC panel (3.08), while Claude, ChatGPT, and Grok sit at the more dovish end of the distribution. This difference is substantively useful: Claude is relatively dovish in tone on November 20 even though it is already on the correct side of the rate call, which shows that policy-step accuracy and narrative stance are distinct dimensions of forecast performance.

The ECB panel exhibits the same logic, but at a higher and more stable tone level. The cross-model range is only 0.11 on November 11 and 0.15 on November 24, but it widens to 0.33 on December 5 and 0.37 by December 17 as some projected statements stress persistence and caution more

strongly than others. Gemini is the most hawkish model on average in the ECB panel (3.32), while Grok is the most dovish (3.19). Claude lies close to the cross-model mean overall but becomes more hawkish by the final date. Thus, even in the cleanest policy-step case, the projected narrative remains meaningfully heterogeneous.

Second, average tone moves over time within an institution even when the modal action is stable. In the FOMC panel, the cross-model average declines from 3.14 on November 20 to 2.67 on December 9 as cut language becomes more accepted and the projected guidance becomes less defensive. In the ECB panel, the average remains above 3 throughout: 3.28, 3.11, 3.24, and 3.37 across the four reported dates, consistent with a hold decision accompanied by continued emphasis on persistence, patience, and meeting-by-meeting calibration. These within-institution shifts reinforce the value of the textual analysis: the rate decision and the communicated stance do not converge at the same speed.

Third, the direction of model-specific tone bias is persistent enough to be informative. Viewed through the within-date bias definition in Appendix B, Perplexity tends to display a positive hawkish bias in the FOMC panel, while Gemini tends to display a positive hawkish bias and Grok a negative bias in the ECB panel. This is useful for interpretation because it suggests that some part of the cross-model heterogeneity reflects stable narrative tendencies rather than pure noise. In short, the textual layer adds information that is not contained in policy-step accuracy alone.

4.4. Disagreement under close calls: evidence from the BoE

Table 4 reports a side-by-side comparison of model-generated BoE communications on November 19, 2025, which is the most informative snapshot of the BoE case because the rate decision is still genuinely contested. On that date, the equal-weight consensus in Table 1 assigns 54.6 percent probability to the realized 25 bp cut and 41.4 percent to a hold, indicating that the meeting is perceived as finely balanced rather than settled. The model-level projections in Table 4 reflect that ambiguity directly: ChatGPT, Claude, and Grok project a 25 bp cut to 3.75 percent, whereas Gemini and Perplexity still project a hold at 4.00 percent.

The disagreement is not limited to the policy step itself. The projected vote splits differ materially across models and provide an additional window into perceived committee balance. ChatGPT and Claude both project a 3–6 split in hold-versus-cut notation, Grok projects 4–5, and Gemini and Perplexity retain 5–4 hold-majority formulations. Narrative framing also diverges. The cut-leaning projections emphasize disinflation, a softer labour-market backdrop, and scope for gradual easing, although even within that group the language is not identical: Claude is more explicit about slack building and conditional movement toward neutral, while Grok retains more caution about residual services-inflation risk. By contrast, Gemini and Perplexity place greater weight on inflation persistence and a still-restrictive stance, which leads them to preserve a hold-first narrative. The vote notation itself is also not fully standardized across outputs, which is informative about model-compliance limits when the prompt requires institution-specific formatting.

This dispersion is economically meaningful. The realized BoE outcome was indeed a 25 bp cut, but on a narrow 5–4 vote, so the internal balance of the Committee was genuinely close. Table 4 therefore shows that LLM disagreement is most informative not when the policy decision is obvious, but when the same public information can support multiple plausible combinations of rate action, vote split, and guidance. In that sense, the BoE case reinforces the value of treating projected statement language as a forecast object in its own right: the rate call and the communication call do not converge at the same speed, and the remaining dispersion in narrative framing provides a transparent proxy for pre-meeting communication uncertainty.

4.5. Summary

Taken together, the December 2025 meeting set provides suggestive evidence rather than definitive conclusions. First, forecast performance appears to vary across institutions: the ECB is the clearest case in our sample, the FOMC is more difficult early in the forecast window, and the BoE remains a close call even after the modal forecast shifts toward the realized cut. Second, the equal weight consensus improves average performance in this

sample, including probabilistic accuracy, but it does not eliminate disagreement when the policy decision is ambiguous. Third, the remaining disagreement is especially informative in the communication dimension. Even when models begin to agree on the likely rate decision, they may still differ in projected tone, risk assessment, forward guidance, vote split language, and the explanation of the decision. This cross model dispersion should therefore not be interpreted only as noise or forecast failure. It may also provide a useful pre meeting measure of uncertainty about how the central bank will communicate its decision. Larger samples are needed to test whether this interpretation generalizes beyond the December 2025 meetings.

5. Discussion and Limitations

Four points are central for interpreting the results.

First, agreement on the rate decision does not imply agreement on the communication. Across the December sample, the policy-step call often converges before the surrounding language about inflation persistence, labor-market slack, risk balance, and forward guidance. This is why the hawkish–dovish analysis is not a stylistic add-on to the rate forecast. From a finance perspective, the relevant object is the full announcement package: markets respond not only to the current policy step, but also to what the statement implies about the future path of policy. The persistent dispersion in the tone scores therefore indicates that models can agree on the same action while still transmitting materially different policy signals.

Second, the paper speaks to a data problem that is especially important in finance and monetary economics. Central-bank decisions are economically consequential but low-frequency. Even for major institutions, there are only a small number of policy meetings each year, which limits the effective sample for studying statement-level forecasting and communication dynamics. Generative AI does not create additional realized policy decisions, but it can help close part of this gap by generating a dense ex ante panel of time-stamped probability forecasts and projected statements around each

meeting. That richer panel makes it possible to study how forecasts converge, where disagreement persists, and how communication uncertainty evolves before the announcement. In this sense, the contribution is not only predictive; it is also measurement-based.

Third, several features of realized statements reflect institution-specific implementation choices and committee-level drafting judgment that are not reducible to a mechanical mapping from a small set of public macroeconomic series. The BoE vote split is an obvious example, but the same point applies to the FOMC's balance-of-risks language and to the ECB's calibration of persistence and patience. Official releases are committee documents that embed strategic communication choices, drafting conventions, and institutional preferences. For that reason, even a well-performing LLM forecast should be interpreted as a benchmark for what is predictable from the public information set, not as a substitute for committee judgment.

Fourth, compliance itself is part of the evaluation. Some outputs use inconsistent vote notation, mix verified data with previews, or leave probabilities only partially explicit. These features do not make the forecasts uninformative; rather, they reveal differences in how well models translate their internal judgment into disciplined, institution-style policy communication. The prospective design with archived raw outputs is therefore valuable for two reasons: it captures predictive content, and it makes the quality of model-generated communication auditable.

These points also define the main limitations of the design. Our baseline sample remains small, the models themselves evolve over time, and retrieval environments differ across platforms. The disclosure requirements in the prompt make those differences more visible, but they do not fully harmonize the underlying information sets. In addition, the hawkish–dovish rubric is transparent and replicable, yet it still contains an element of researcher judgment. More extensive samples, repeated scoring exercises, and complementary automated text measures would help assess the stability of the narrative results. Finally, although real-time prompting materially reduces the most direct form of look-ahead bias, it cannot rule out all forms of inadvertent contamination if a system accesses post-cutoff information that is not disclosed.

6. Conclusion

This paper studies whether frontier LLMs can forecast central-bank policy statements before they are released. Using a real-time design with archived prompts and outputs, we evaluate five models (ChatGPT, Claude, Gemini, Grok, and Perplexity) ahead of three December 2025 meetings at the FOMC, ECB, and BoE. The evidence shows that forecast performance is institution-specific: the ECB is the cleanest case, the BoE cut is identified early but remains a close-call setting, and the FOMC cut is the hardest case early in the window before the consensus converges.

A second conclusion is that narrative convergence is slower than policy-step convergence. Even when models ultimately agree on the most likely action, they continue to differ in how they frame inflation, labor markets, the balance of risks, and the strength of forward guidance. The hawkish–dovish analysis is therefore central to the paper's contribution, because it shows that the communication dimension contains information that is not captured by the rate call alone.

The broader contribution is methodological and finance-oriented. Monetary policy communication is a low-frequency empirical object, and that creates a natural data constraint for statement-level forecasting. Generative AI can help close part of this gap by producing a structured ex ante panel of probability forecasts and projected statements around each event. That panel does not replace realized data, but it makes it possible to study convergence, disagreement, and narrative uncertainty at a much richer frequency than the meeting calendar alone would allow. In this sense, LLMs are useful not only as forecasters, but also as scalable tools for measuring ex ante communication uncertainty.

The equal-weight ensemble reinforces this point. Averaging across models improves robustness and reduces average probabilistic loss, but it does not eliminate disagreement in ambiguous environments. Residual dispersion across models is informative in its own right, because it provides a transparent proxy for uncertainty about both the decision and the language used to justify it.

Future work can extend the sample across time, institutions, and communication formats; compare LLM probabilities with market-implied and survey-based expectations; and test whether ex ante LLM disagreement helps explain announcement-day asset-price reactions or realized volatility. A related direction is to formalize chronological auditing standards for time-sensitive prediction tasks and to combine rubric-based measures, such as the hawkish–dovish index, with automated text-similarity and embedding-based diagnostics. Overall, the evidence suggests that disciplined LLM forecasts can serve as a useful benchmark for what is predictable from public information and as a new empirical tool for studying central-bank communication in finance.

References

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

Bates, J. M., & Granger, C. W. J. (1969). The combination of forecasts. *Journal of the Operational Research Society, 20*(4), 451–468.

Bernanke, B. S., & Kuttner, K. N. (2005). What explains the stock market's reaction to Federal Reserve policy? *The Journal of Finance, 60*(3), 1221–1257.

Crane, L. D., Karra, A., & Soto, P. E. (2025). Total recall? Evaluating the macroeconomic knowledge of large language models (Finance and Economics Discussion Series 2025-044). Board of Governors of the Federal Reserve System.

Ehrmann, M., & Talmi, J. (2020). Starting from a blank page? Semantic similarity in central bank communication and market volatility. *Journal of Monetary Economics, 111*, 48–62.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature, 57*(3), 535–574.

Glasserman, P., & Lin, C. (2023). Assessing look-ahead bias in stock return predictions generated by GPT sentiment analysis. arXiv:2309.17322.

Gössi, S., Chen, Z., Kim, W., Bermeitinger, B., & Handschuh, S. (2023). FinBERT-FOMC: Fine-tuned FinBERT model with Sentiment Focus method for enhancing sentiment analysis of FOMC minutes. In *Proceedings of the International Conference on AI in Finance (ICAIF 2023)* (pp. 357–364). Association for Computing Machinery.

Gürkaynak, R. S., Sack, B., & Swanson, E. T. (2005). Do actions speak louder than words? The response of asset prices to monetary policy actions and statements. *International Journal of Central Banking, 1*(1), 55–93.

Hansen, S., & McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. *Journal of International Economics, 99*(S1), S114–S133.

He, S., Lv, L., Manela, A., & Wu, J. (2025). Instruction tuning chronologically consistent language models. arXiv:2510.11677.

Hilscher, J., Nabors, K., & Raviv, A. (2026). Information in central bank sentiment: An analysis of Fed and ECB communication. *Journal of International Financial Markets, Institutions and Money*, 110.

Kanelis, D., Kranzmann, L. H., & Siklos, P. L. (2025). The financial instability–monetary policy nexus: Evidence from the FOMC minutes (Deutsche Bundesbank Discussion Paper No. 13/2025).

Kazinnik, S., & Sinclair, T. M. (2025). FOMC In Silico: A Multi-Agent System for Monetary Policy Decision Modeling. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5424097.

Kim, W., Spörer, J. F., Lee, C. L., & Handschuh, S. (2024). Is small really beautiful for central bank communication? Evaluating language models for finance. In *Proceedings of the International Conference on AI in Finance (ICAIF 2024)* (pp. 626–633). Association for Computing Machinery.

Kuttner, K. N. (2001). Monetary policy surprises and interest rates: Evidence from the Fed funds futures market. *Journal of Monetary Economics, 47*(3), 523–544.

Lesmy, D., Muchnik, L., & Mugerman, Y. (2025). Lost in the fog: Growing complexity in financial reporting: A comparative study. *Humanities and Social Sciences Communications, 12*, Article 1813.

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

Lucca, D. O., & Trebbi, F. (2009). Measuring central bank communication: An automated approach with application to FOMC statements (NBER Working Paper No. 15367). National Bureau of Economic Research.

Ludwig, J., Mullainathan, S., & Rambachan, A. (2025). Large language models: An applied econometric framework (NBER Working Paper No. 33344). National Bureau of Economic Research.

Muchnik, L., Mugerman, Y., Noiman, B., & Sade, O. (2025). Opaque by design? Leveraging large language models to analyze financial reports readability under intellectual property litigation. SSRN.

Silva, T. C., Moriya, K., & Veyrune, R. M. (2025). From text to quantified insights: A large-scale LLM analysis of central bank communication (IMF Working Paper No. 2025/109). International Monetary Fund.

Stock, J. H., & Watson, M. W. (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting, 23*(6), 405–430.

Taskin, B., & Akal, F. (2025). Tales of turbulence: BERT-based multimodal analysis of FED communication dynamics amidst COVID-19 through FOMC minutes. *Computational Economics, 65*, 117–146.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance, 62*(3), 1139–1168.

Timmermann, A. (2006). Forecast combinations. In G. Elliott, C. W. J. Granger, & A. Timmermann (Eds.), *Handbook of Economic Forecasting* (Vol. 1, pp. 135–196). Elsevier.

Figure 1, Panel A. FOMC (realized outcome: 25 bp cut). Convergence of LLM forecast probabilities toward the realized policy outcome. Each marker shows the probability that an individual model assigns to the outcome that was ultimately realized; the bold black line is the equal-weight consensus. The shaded band spans the range of individual-model probabilities. The dashed vertical line marks the meeting date.

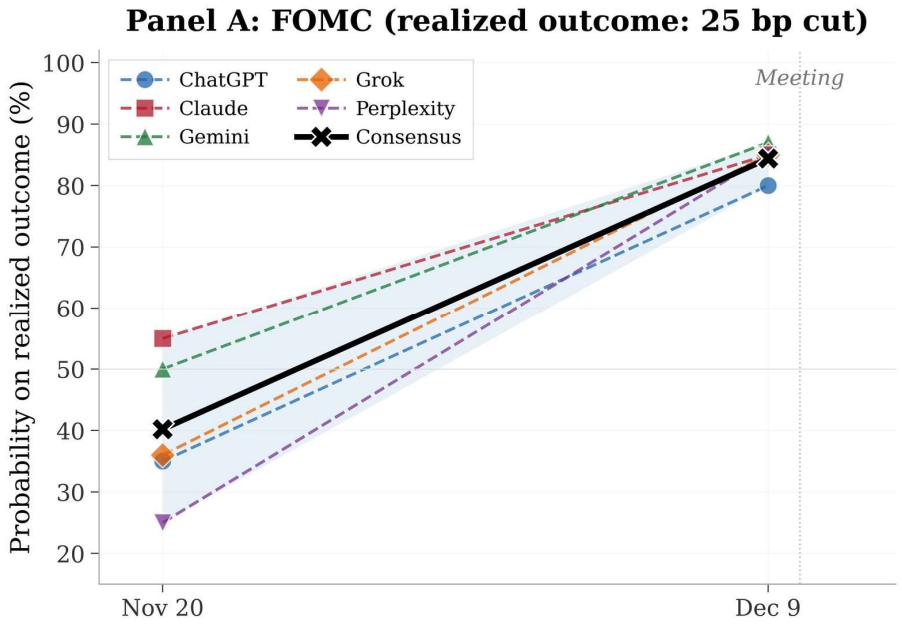

Figure 1, Panel B. ECB (realized outcome: hold). Convergence of LLM forecast probabilities toward the realized policy outcome. Each marker shows the probability that an individual model assigns to the outcome that was ultimately realized; the bold black line is the equal-weight consensus. The shaded band spans the range of individual-model probabilities. The dashed vertical line marks the meeting date.

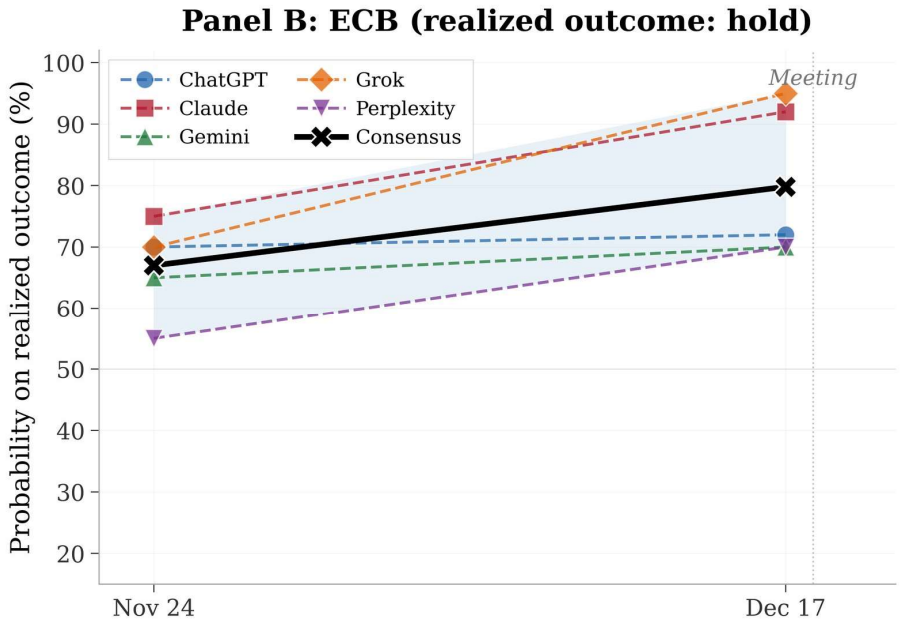

Figure 1, Panel C. BoE (realized outcome: 25 bp cut). Convergence of LLM forecast probabilities toward the realized policy outcome. Each marker shows the probability that an individual model assigns to the outcome that was ultimately realized; the bold black line is the equal-weight consensus. The shaded band spans the range of individual-model probabilities. The dashed vertical line marks the meeting date.

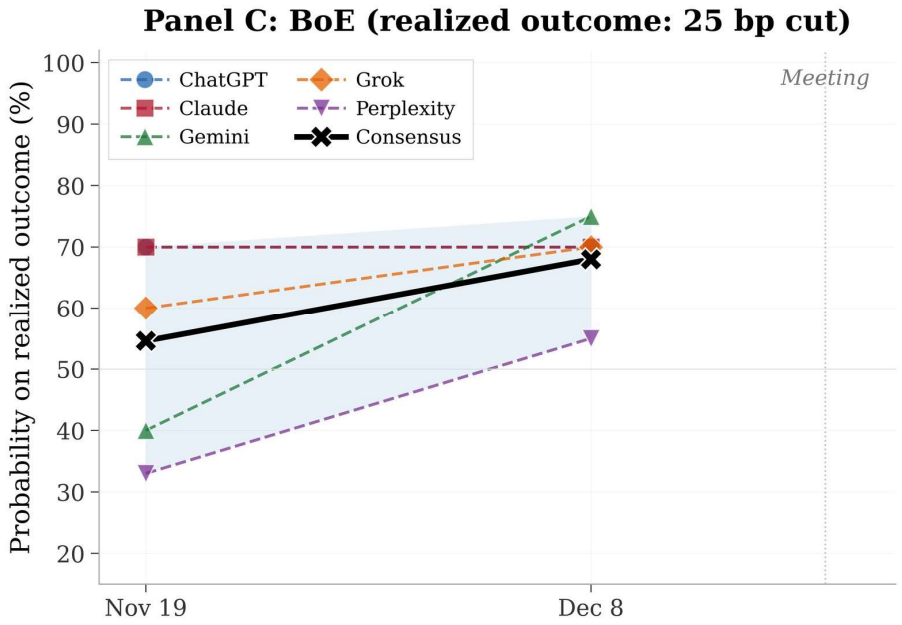

Table 1. LLM Probability Distributions (Selected Dates)

Panel A: FOMC (December 9–10, 2025). Realized outcome: 25 bp cut.

Date	Model	50 bp cut	25 bp cut	Hold	25 bp hike	50 bp hike	Most likely
November 20, 2025	ChatGPT	3	35	57	4	1	Hold
November 20, 2025	Claude	2	55	42	1	0	25 bp cut
November 20, 2025	Gemini	5	50	40	5	0	25 bp cut
November 20, 2025	Grok	0	36	64	0	0	Hold
November 20, 2025	Perplexity	2	25	67	5	1	Hold
November 20, 2025	***Consensus***	***2.4***	***40.2***	***54.0***	***3.0***	***0.4***	***Hold***
December 9, 2025	ChatGPT	5	80	14	1	0	25 bp cut
December 9, 2025	Claude	2	85	13	0	0	25 bp cut
December 9, 2025	Gemini	3	87	10	0	0	25 bp cut
December 9, 2025	Grok	0	85	15	0	0	25 bp cut
December 9, 2025	Perplexity	1.5	85	11	2	0.5	25 bp cut
December 9, 2025	***Consensus***	***2.3***	***84.4***	***12.6***	***0.6***	***0.1***	***25 bp cut***

Panel B: ECB (December 18, 2025). Realized outcome: hold.

Date	Model	50 bp cut	25 bp cut	Hold	25 bp hike	50 bp hike	Most likely
November 24, 2025	ChatGPT	3	22	70	4	1	Hold
November 24, 2025	Claude	3	20	75	1	1	Hold
November 24, 2025	Gemini	5	20	65	10	0	Hold
November 24, 2025	Grok	1	25	70	3	1	Hold
November 24, 2025	Perplexity	5	30	55	8	2	Hold
November 24, 2025	***Consensus***	***3.4***	***23.4***	***67.0***	***5.2***	***1.0***	***Hold***
December 17, 2025	ChatGPT	1	20	72	6	1	Hold
December 17, 2025	Claude	0	7	92	1	0	Hold
December 17, 2025	Gemini	0	5	70	25	0	Hold
December 17, 2025	Grok	0	5	95	0	0	Hold
December 17, 2025	Perplexity	5	10	70	10	5	Hold
December 17, 2025	***Consensus***	***1.2***	***9.4***	***79.8***	***8.4***	***1.2***	***Hold***

Panel C: BoE (meeting ending December 17, 2025; summary published December 18, 2025). Realized outcome: 25 bp cut.

Date	Model	50 bp cut	25 bp cut	Hold	25 bp hike	50 bp hike	Most likely
November 19, 2025	ChatGPT	0	70	30	0	0	25 bp cut
November 19, 2025	Claude	5	70	25	0	0	25 bp cut
November 19, 2025	Gemini	0	40	60	0	0	Hold
November 19, 2025	Grok	10	60	30	0	0	25 bp cut
November 19, 2025	Perplexity	0	33	62	5	0	Hold
November 19, 2025	***Consensus***	***3.0***	***54.6***	***41.4***	***1.0***	***0.0***	***25 bp cut***
December 8, 2025	ChatGPT	1	70	25	3	1	25 bp cut
December 8, 2025	Claude	3	70	27	0	0	25 bp cut
December 8, 2025	Gemini	5	75	20	0	0	25 bp cut
December 8, 2025	Grok	3	70	25	1	1	25 bp cut
December 8, 2025	Perplexity	1	55	40	3	1	25 bp cut
December 8, 2025	***Consensus***	***2.6***	***68.0***	***27.4***	***1.4***	***0.6***	***25 bp cut***

Notes: Entries report model-forecast probability mass over five discrete outcomes; "Consensus" is the simple cross-model average. Realized decisions and key policy settings are taken from official central bank releases: the December 10, 2025 FOMC statement; the December 18, 2025 ECB monetary policy decisions release; and the BoE Monetary Policy Summary and Minutes for the meeting ending December 17, 2025 and published on December 18, 2025.

Table 2. FOMC Hawkish–Dovish Index by Model and Date (1 = dovish, 5 = hawkish)

Date	ChatGPT	Claude	Gemini	Grok	Perplexity	Average
November 20, 2025	3.25	2.84	2.96	3.33	3.31	3.14
November 24, 2025	2.75	2.81	3.24	2.63	2.99	2.88
December 1, 2025	2.69	2.94	2.94	2.79	3.31	2.93
December 9, 2025	2.70	2.75	2.59	2.60	2.71	2.67
Overall	*2.85*	*2.83*	*2.93*	*2.84*	*3.08*	*2.91*

Notes: The hawkishness index is the simple average of eight rubric components defined in Appendix B; scores are computed from the projected statement text. "Overall" is the mean across the four dates.

Table 3. ECB Hawkish–Dovish Index by Model and Date (1 = dovish, 5 = hawkish)

Date	ChatGPT	Claude	Gemini	Grok	Perplexity	Average
November 11, 2025	3.22	3.25	3.33	3.29	3.33	3.28
November 24, 2025	3.08	3.09	3.21	3.06	3.11	3.11
December 5, 2025	3.45	3.22	3.19	3.22	3.12	3.24
December 17, 2025	3.23	3.38	3.56	3.19	3.49	3.37
Overall	*3.25*	*3.24*	*3.32*	*3.19*	*3.26*	*3.25*

Notes: Same rubric logic as Table 2, with the ECB-specific substitution described in Appendix B (balance sheet programs in place of the "data delays" criterion). Benchmark ECB decision language and policy settings are from the official ECB monetary policy decisions release dated December 18, 2025.

Table 4. Bank of England Cross-Model Projection Comparison (November 19, 2025)

Criterion	ChatGPT	Claude	Gemini	Grok	Perplexity
Rate decision	Cut 25 bp to 3.75%	Cut 25 bp to 3.75%	Hold at 4.00%	Cut 25 bp to 3.75%	Hold at 4.00%
Vote split	3–6 (hold versus cut)	3–6 (hold versus cut)	5–4 (hold versus cut)	4–5 (hold versus cut)	5–4 (hold versus cut)
Inflation	Easing inflation; weaker labour market	Disinflation progressing; slack building; wage momentum easing	Services inflation and expectations still persistent	Inflation pressures easing but services remain a risk	Persistent inflation risks; services inflation emphasized
Guidance	Data-dependent; gradual downward path	Gradual moves toward neutral; conditional on disinflation	Hold with easing bias; data-dependent	Measured easing but cautious; data-dependent	Restrictive stance; easing conditional

Notes: This is a descriptive comparison of model-generated projections on a single forecast date; it reports what models stated, including inconsistent vote notation when it appeared. The realized decision benchmark for the BoE meeting ending December 17, 2025 is the official Monetary Policy Summary and Minutes published on December 18, 2025.

Appendix A. Prompt Template

Prompt reproduced verbatim (FOMC version):

Task

Generate an up-to-date projected FOMC policy statement for the upcoming meeting on December 9-10, 2025. Your output must read like an FOMC statement and clearly explain your method and evidence.

Method and evidence rules

1. Declare your data cutoff
   • Give a precise timestamp in US Eastern Time used by you.
   • Anything after that timestamp is forecast or preview only. Label it as such.
2. Source transparency
   • List every source you used with the full document or page title and the exact page or section.
   • For each numeric value add a bracketed citation that points to the specific page or section.
   • If a value is a forecast or nowcast name the forecaster or model.
3. Separate facts and forecasts
   • Create a table with two columns: Verified data and forecast or preview.
   • For each item includes series name period value and a bracketed page or section reference.
4. Probabilistic decision
   • Report a probability distribution for the meeting outcome: cut 25 bp, hold, hike 25 bp and for hike 50bp and 50bp. The probabilities must sum to 100 percent.
   • State the main drivers that move probability mass ranked by impact.
5. Reaction function disclosure
   • Explain how you have reached the decision and what the reaction function that led to this interest rate decision.

• Show how the decision is based on current data and how do they map into the different possible outcomes.

• If you use any weights or thresholds print them.

6. Consistency and self-check

• Run three checks and report the result:

A. All cited numbers trace to a page or section.

B. No forecast is presented as verified data.

C. All dates are in month day-year format and refer to the latest release as of your cutoff.

Required output sections

A) Header

• Meeting dates.

• Your data cutoff with time zone.

• One line summary of your base case and the probability distribution.

B) Method

• Explanation on how you gather data and form the decision.

• If you have used a simple reaction function please show it, otherwise explain your method.

C) Evidence table

• Two column table: Verified data and forecast or preview.

• Include at least these series when available

• Each row includes series name latest period value and a bracketed page or section reference.

D) Draft statement

• One to three paragraphs styled like an FOMC post decision statement.

• Include the projected decision, the rationale the balance of risks and guidance language.

• Do not invent names of officials. Do not state that a document exists unless you cited it.

E) Changes versus previous statement

• Bullet list of specific textual changes you propose relative to the prior statement you used including a brief why for each change.

• Include bracketed references to the sections you altered.

F) Probability and drivers

• A bullet list with the outcome probabilities.

• Three to five bullets that rank the drivers by importance each with a short explanation and a source reference.

G) Validation log

• List every numeric claim with its source page or section.

• Note any conflicts across sources and how you resolved them.

H) Caveats

• One short paragraph on what new data before the meeting could shift the decision and how.

Output constraints

Appendix B. Hawkish–Dovish Tone and Bias Scoring

B.1. Rubric structure.

Each projected statement is scored on eight criteria. Each criterion is mapped to a 1–5 score, where 1 is maximally dovish and 5 is maximally hawkish. Half-point increments are permitted. The eight criteria are: (1) policy action and rate range; (2) inflation language; (3) labor market characterization; (4) risk-balance assessment; (5) financial-conditions line; (6) data delays or uncertainty note (ECB: replaced by balance sheet programs language); (7) forward-guidance tone; and (8) voting disagreement.

The scoring model used here is a rule-based, criterion-by-criterion annotation model. We do not use a generic sentiment model as the baseline scoring rule because generic positive or negative tone is not the same as monetary-policy hawkishness. A statement can sound negative in a general linguistic sense while still be dovish, or vice versa.

The criterion set was chosen to mirror the components of central-bank communication that matter most for finance. Criterion 1 captures the current policy action. Criteria 2–4 capture the macroeconomic rationale, especially inflation, labor-market conditions, and the balance of risks. Criteria 5–7 capture the operational and forward-looking communication layer: financial conditions, institution-specific uncertainty or balance-sheet language, and forward guidance. Criterion 8 captures voting or dissent signals where they exist. Together, these eight dimensions reflect the aspects of policy communication that are most likely to affect expectations about the future path of policy.

B.2. Formal definitions.

Let s_{j,b,t,c} denote the score assigned to model j, central bank b, forecast date t, and criterion c ∈ {1,…,8}. Each criterion is scored on a 1–5 scale, where larger values are more hawkish. Half-point increments are permitted.

$$s_{j,b,t,c} \in \{1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0\}. \tag{B.1}$$

The baseline hawkish–dovish index is the simple average across the eight criteria, where ŝ_{j,b,t,c} is the operational score after handling structurally absent criteria and genuine truncation.

$$H_{j,b,t} = \frac{1}{8} \sum_{c=1}^{8} \tilde{s}_{j,b,t,c}. \tag{B.2}$$

$$\tilde{s}_{j,b,t,c} = \begin{cases} s_{j,b,t,c}, & \text{if criterion } c \text{ is observed directly;} \\ 3.0, & \text{if criterion } c \text{ is structurally absent and the neutral anchor applies;} \\ 3.0, & \text{if criterion } c \text{ is mechanically truncated, at most two criteria are missing, and the statement remains scoreable;} \\ \text{NA}, & \text{if the projected statement is too incomplete to score reliably.} \end{cases}$$

For a given bank-date pair, the within-date tone bias of model j is the model's tone score minus the cross-model mean score on that same date. Negative values indicate a more dovish tone than the cross-model mean; positive values indicate a more hawkish tone.

$$\text{Bias}_{j,b,t} = H_{j,b,t} - \frac{1}{N_{b,t}} \sum_{i=1}^{N_{b,t}} H_{i,b,t}.$$

B.3. Scoring rules for the eight criteria.

Table B1 gives the operational scoring rules. The scoring logic is intentionally transparent. The coder identifies the most relevant evidence span for each criterion, maps it to the closest anchor, and uses a half-point value when the text lies between two adjacent anchors.

Table B1. Hawkish–dovish rubric with explicit anchors

Criterion	Dove anchor	Neutral anchor	Hawk anchor	Explicit decision rule
Policy action and rate range	1.0 = cut 50 bp or stronger easing language; 2.0 = cut 25 bp	3.0 = hold	4.0 = hike 25 bp; 5.0 = hike 50 bp or stronger tightening language	Start from the modal action. Add +0.5 if the statement explicitly frames policy as still restrictive or insufficiently restrictive after the action. Subtract 0.5 if the statement explicitly frames the action as support for growth or employment or as the start or continuation of easing. Clip to [1,5].
Inflation language	1.0 = inflation at or below target, broad disinflation, downside inflation risk; 2.0 = clear disinflation progress	3.0 = mixed language, still above target but improving	4.0 = inflation somewhat elevated or persistent; 5.0 = strong persistence or upside inflation risk	Use the dominant descriptor. If the statement contains both strong progress and explicit persistence, use 3.5.
Labor-market characterization	1.0 = clear slack or weakening, rising unemployment, layoffs; 2.0 = moderate softening	3.0 = balanced or neutral	4.0 = labor market still solid or tight; 5.0 = overheating	Score hawkishness, not 'good news' tone. Stronger labor markets imply less easing pressure and therefore a more hawkish score.
Risk-balance assessment	1.0 = downside growth or employment risks dominate; 2.0 = risks shifted toward weaker activity or employment	3.0 = risks broadly balanced or two-sided with no clear tilt	4.0 = upside inflation risks dominate; 5.0 = strong inflation-risk tilt	If both sides are explicitly mentioned, assign 3.0 or 3.5 depending on which side receives more emphasis.
Financial-conditions line	1.0 = tighter credit or financial stress explicitly weighs on activity;	3.0 = no dedicated financial-conditions line	4.0 = easing financial conditions seen as inflationary; 5.0 = strong	If the statement says nothing specific about financial conditions, assign 3.0.

Criterion	Dove anchor	Neutral anchor	Hawk anchor	Explicit decision rule
	2.0 = restrictive conditions cited as growth drag	or purely generic monitoring language	concern that looser conditions undermine disinflation	
Uncertainty / data delays / balance-sheet programs	FOMC or BoE: 1.0 = exceptional uncertainty or missing data pushes toward caution or easing; 2.0 = moderate uncertainty emphasized	FOMC or BoE: 3.0 = no material uncertainty note. ECB: 3.0 = no non-standard balance-sheet or programs language, or standard operational language only	FOMC or BoE: 4.0–5.0 = explicit confidence or diminished uncertainty. ECB: 4.0–5.0 = explicit restrictive balance-sheet runoff, QT, or programs tightening	This is the institution-specific criterion. For the ECB, replace the uncertainty note with balance-sheet or programs language. If absent, assign 3.0.
Forward-guidance tone	1.0 = explicit easing bias or openness to further cuts; 2.0 = conditional easing bias	3.0 = data-dependent, no preset path	4.0 = cautious or higher-for-longer language; 5.0 = explicit tightening bias	'Not pre-committing' combined with strong anti-inflation language usually maps to 3.5 or 4.0.
Voting or dissent signal	1.0 = large dovish majority or dovish dissents; 2.0 = clear easing majority	3.0 = no vote language, no dissent signal, or genuinely neutral presentation	4.0 = narrow split or hawkish dissent against easing; 5.0 = multiple hawkish dissents or majority resisting easing	If vote language is absent, assign 3.0. If a cut is projected but several members are described as preferring hold, assign 3.5 or 4.0 depending on the split.

Ambiguous-case rules.

If adjacent anchors both apply, use the midpoint. For example, 'inflation has eased but remains somewhat elevated' maps naturally to 3.0 rather than forcing a pure 2.0 or 4.0.

If two opposing anchors appear with roughly equal force in the same criterion, assign 3.0 unless the statement clearly resolves the tension. For example, 'two-sided risks' without an evident tilt is 3.0 on risk balance.

If the statement is silent on a criterion that is commonly omitted in institutional templates, use the neutral anchor of 3.0. This applies especially to financial conditions and vote language.

If the statement is truncated mechanically, rather than substantively silent, impute 3.0 only when no more than two criteria are affected and the rest of the statement is complete enough to identify the action, the macro rationale, and the guidance. Otherwise exclude the observation from the tone sample.

If the projected statement and the surrounding explanation conflict, the draft statement itself governs the tone score. The surrounding narrative may be used for diagnostics, but the tone rubric is a statement-level measure.

B.4. Worked example using ChatGPT projected FOMC text from December 1, 2025.

Evidence spans used.

The projected statement includes the following key language from the December FOMC forecast archive:

- 'Inflation has eased over the past year but remains above the Committee's 2 percent objective.'
- '… which increases uncertainty about current conditions.'
- '… risks … still skewed modestly toward weaker employment.'
- 'The Committee decided to lower the target range …'

Table B2. End-to-end scoring example: ChatGPT, FOMC, December 1, 2025

Criterion	Evidence used	Score	Reason
Policy action and range	'decided to lower the target range …'	2.5	Base score 2.0 for a 25 bp cut, plus +0.5 because the statement frames the move as only a small reduction in restraint rather than a sharp easing pivot.

Criterion	Evidence used	Score	Reason
Inflation language	'Inflation has eased … but remains above the Committee's 2 percent objective.'	3.0	Mixed language: clear progress, but still above target.
Labor market characterization	'… skewed modestly toward weaker employment.'	2.5	Labor conditions are softening, which is dovish.
Risk-balance assessment	'… risks … still skewed modestly toward weaker employment.'	2.5	Explicit downside tilt to employment risks.
Financial-conditions line	No dedicated financial-conditions sentence beyond generic monitoring language	3.0	Structural absence, so neutral anchor applies.
Uncertainty / data delays	'… increases uncertainty about current conditions.'	2.5	Explicit uncertainty note that weakens confidence in a hawkish reading.
Forward-guidance tone	'… will carefully assess …'	2.5	Conditional, cautious, and mildly dovish guidance.
Voting or dissent signal	No explicit vote language inside the projected statement	3.0	Structural absence, so neutral anchor applies.

$$S = 2.5 + 3.0 + 2.5 + 2.5 + 3.0 + 2.5 + 2.5 + 3.0 = 21.5.$$

$$H_{\text{ChatGPT,FOMC,2025-12-01}} = \frac{21.5}{8} = 2.6875 \approx 2.69. \tag{B.4}$$

This example shows why the overall tone score need not be as dovish as the policy action alone. The projected statement cuts rates, but it still retains above-target inflation language and a cautious, "carefully assess" guidance formulation.

B.5. Worked example using Claude projected ECB text from December 17, 2025.

Evidence spans used.

The projected statement includes the following key language from the December ECB forecast archive:

- 'The Governing Council today decided to keep the three key ECB interest rates unchanged.'
- '… persistent price pressures in services.'
- 'The labor market remains robust, with unemployment holding at 6.4 percent.'
- 'The outlook remains subject to two-sided risks.'
- '… looser domestic financial conditions could sustain price pressures …'

• 'We are not pre-committing to a particular rate path.'

Table B3. End-to-end scoring example: Claude, ECB, December 17, 2025

Criterion	Evidence used	Score	Reason
Policy action and range	'decided to keep the three key ECB interest rates unchanged.'	3.5	Base score 3.0 for hold, plus +0.5 because the hold is framed in a still restrictive anti-inflation posture rather than as an easing hold.
Inflation language	'persistent price pressures in services' and inflation outlook 'broadly unchanged'	3.5	Inflation is improving only partially; services persistence keeps the language on the hawkish side of neutral.
Labor market characterization	'The labor market remains robust …'	3.5	Robust labor-market language reduces immediate easing pressure.
Risk-balance assessment	'The outlook remains subject to two-sided risks' with explicit upside inflation risk	3.5	Two-sided frame, but wording leaves meaningful upside inflation concern.
Financial-conditions line	'… looser domestic financial conditions could sustain price pressures …'	3.5	Explicit concern that easier financial conditions could keep inflation pressure alive.
Balance-sheet / programs criterion	No non-standard balance-sheet or programs language in the projected press release	3.0	ECB-specific criterion defaults to neutral when the statement is silent or merely standard.
Forward-guidance tone	'We are not pre-committing to a particular rate path.'	3.5	Data-dependent, but with a hawkishly cautious refusal to pre-commit to easing.
Voting or dissent signal	No vote language in the projected press release	3.0	Structural absence, so neutral anchor applies.

$$S = 3.5 + 3.5 + 3.5 + 3.5 + 3.5 + 3.0 + 3.5 + 3.0 = 27.0.$$

$$H_{\text{Claude,ECB,2025-12-17}} = \frac{27.0}{8} = 3.375 \approx 3.38. \tag{B.5}$$

This example illustrates why a simple hold does not imply a neutral tone. The rate decision is unchanged, but the narrative is more hawkish than neutral because the projected statement continues to emphasize services-price persistence, robust labor-market conditions, and caution in forward guidance.