Disagreement and Errors in Professional Policy-Rate Expectations: Evidence from New York Fed Policy-Rate Paths

A-Young Park, Ph.D.

Senior Researcher, Supervisory Strategy Research Team

Financial Supervisory Service

Seoul 03171, South Korea

Phone: +82-10-2100-1617

Email: ayoungp@fss.or.kr

The views expressed in this paper are those of the author and do not necessarily reflect the official views or policies of the Financial Supervisory Service.

Disagreement and Errors in Professional Policy-Rate Expectations: Evidence from New York Fed Policy-Rate Paths

Abstract

Professional forecasters may outperform public benchmarks on average while still making large policy-path errors in predictable states. I study whether disagreement with a public-information benchmark constructed from cutoff-dated inputs reveals those errors in the reported New York Fed Market Participant aggregate of modal policy-rate forecasts. The professional path has a mean absolute error of 32.86 basis points, compared with 49.61 for the Ridge challenger. Yet the absolute professional–model gap ranks 50-basis-point forecast errors: its ROC-AUC is 0.712 across 388 horizon cells. The signal does not deliver a stable advantage over a flat no-change path. When one-quarter of cells are flagged, model disagreement captures 47.3% of large errors versus 43.9% for no change on the full support, but 47.8% versus 52.8% on a later support; paired differences are imprecise. A decomposition traces part of the capture difference to same-direction professional and model errors, including cases in which the model is closer to the realization. ECB and Bank of England policy-path applications show similar raw ranking, but the evaluated disagreement scores have no established fixed-share capture advantage over no-change distance. Public-information benchmarks can reveal forecast difficulty without supplying a uniformly stronger diagnostic of professional expectations.

JEL classification: C53; E43; E47; E52; E58.

Keywords: monetary policy expectations; professional forecasts; forecast disagreement; policy-rate paths; real-time data; forecast errors.

The views expressed in this paper are those of the author and do not necessarily reflect the official views or policies of the Financial Supervisory Service.

1 Introduction

Expected policy-rate paths are central to the pricing of interest-rate-sensitive assets and to assessments of the stance and transmission of monetary policy. Professional surveys provide a direct measure of those expectations, but even well-informed forecasts can incur large errors when policy paths change across meetings. The economically relevant question is therefore not only which forecast has the lowest average loss. It is also whether a separately constructed public-information benchmark reveals when professional expectations are unusually fragile.

This paper studies that question using the Market Participant policy-rate paths collected by the Federal Reserve Bank of New York. I compare each professional path with a Ridge forecast constructed from real-time macroeconomic vintages, lagged financial-market information, and Federal Reserve text available by the survey cutoff. The setting provides a stringent comparison of diagnostic information. The professional aggregate may incorporate institutional knowledge and judgment absent from the model, while the model offers a fixed path based on cutoff-dated public inputs. A large gap between the two may indicate a difficult forecasting state even if the model is not the better forecast.

A simple monetary-policy benchmark makes this test demanding. Target rates are persistent, policy changes occur on a discrete grid, and the largest errors are concentrated at more distant horizons and in particular policy cycles. The distance of a professional path from no change can therefore predict subsequent loss without any model. Model disagreement may merely repackage this persistence signal. Conversely, a small professional–model gap need not imply accuracy: both paths may make large errors in the same direction. The analysis consequently compares professional–model disagreement with no-change distance and forecast horizon on identical observations and at common flagged shares.

The real-time U.S. panel contains 388 horizon cells from 60 survey paths and 60 target Federal Open Market Committee (FOMC) meetings. The primary outcome is an absolute professional forecast error of at least 50 basis points, equivalent to two standard 25-basis-point policy moves. Market observations strictly precede each survey deadline and macroe-

conomic series use the vintages then available. The Ridge is estimated in expanding windows with training labels restricted to outcomes already realized at each forecast origin. Imputation and scaling are fitted within the corresponding training samples, and penalty selection follows the documented recursive schedule. Six stance measures use a fixed text classifier trained on pre-2015 sentences from a labeled corpus released in 2023; these measures are retrospective transformations of cutoff-eligible documents rather than historically available components at every origin.

The first result establishes the forecasting benchmark. The Market Participant path has a mean absolute error of 32.86 basis points, compared with 49.61 for the targeted-text Ridge, 44.27 for federal-funds futures, and 62.11 for no change. The professional aggregate is therefore the better point forecast under the reported realized-point loss criteria. The model's possible contribution is diagnostic rather than substitutive.

The second result shows raw ranking content but no established advantage over no-change distance. Across the full 388-cell support, absolute professional–model disagreement has a receiver operating characteristic area under the curve (ROC-AUC) of 0.712 for 98 large errors. Market Participant loss rises sharply across disagreement quartiles, and the ranking survives controls for horizon, policy cycle, lead time, within-profession disagreement, and distance from no change. Yet no-change distance has an AUC of 0.700 on the same observations. When one-quarter of cells are flagged, model disagreement captures 47.3% of large errors and no change captures 43.9%.

The relative ordering is not stable across evaluation supports. On the later 250-cell support required for recursive probability estimation, model disagreement has an AUC of 0.709 and captures 47.8% of large errors at a 24.8% flagged share. No change has an AUC of 0.777 and captures 52.8%. The paired comparisons remain imprecise. The change in point-estimate ordering coincides with a sharp change in policy-cycle composition; the design does not identify a regime effect.

A decomposition of the observations weighted differently by the two scores gives the main

economic interpretation. Absolute disagreement measures the distance between forecasts, not their distance from the realization. A large professional error can therefore receive little disagreement-based weight when the model errs in the same direction and remains close to the professional path. The same-direction, model-closer category contains 14.3% and 18.5% of the large errors on the full and nested later supports. Its net contributions to model-minus-no-change capture are -3.75 and -4.10 percentage points. These are fractional-weight contributions, not counts of forecasts wholly excluded from the flagged set. The decomposition accounts for part of the observed selection difference; it does not identify why the forecasts err together.

The comparison extends to professional policy-rate paths from the European Central Bank and the Bank of England. Professional–challenger disagreement again has positive raw-ranking content, but no change has the higher capture point estimate at the stated fixed shares in all three policy-rate settings. A Philadelphia Fed inflation application does not reproduce the raw-ranking result. The evidence supports a bounded regularity in policy-rate forecasting rather than a universal property of model disagreement.

The contribution is empirical: the paper quantifies how a fixed public-information benchmark changes the composition of large professional policy-path errors selected relative to no change and horizon. The shared-direction decomposition locates the errors behind that contrast, while the nested-sample and cross-survey comparisons show its sensitivity to the evaluation setting. These findings concern the measurement and diagnosis of policy-rate expectations, not an identified model of how those expectations are formed. They connect evidence on monetary-policy expectations and information rigidity with real-time forecast evaluation, unstable forecast comparisons, professional judgment, and monetary-policy disagreement (Kuttner, 2001; Coibion and Gorodnichenko, 2015; Faust and Wright, 2009; Giacomini and Rossi, 2010; Sastry, 2026; Pedersen, 2025).

The remainder of the paper is organized as follows. Section 2 describes the policy-path surveys, public-information model, real-time cutoffs, forecast loss, and diagnostic scores.

Section 3 presents the empirical comparisons and inference. Section 4 reports forecast accuracy and large-error ranking. Section 5 studies matched-support performance and shared-direction errors. Section 6 presents the European Central Bank and Bank of England applications. Section 7 concludes.

2 Professional Policy-Rate Paths and Public-Information Benchmarks

2.1 New York Fed Policy-Path Surveys

The professional paths come from the Market Participant and Primary Dealer panels of the New York Fed's policy-expectations surveys. The separate surveys were combined into the Survey of Market Expectations in January 2025, while the underlying data continued to report the two panel breakdowns (Federal Reserve Bank of New York, 2026). I reconstruct the legacy PDF archive and parse the standardized spreadsheet releases directly. The resulting archive contains 90 survey dates from March 2015 through June 2026. The 2014 questionnaires reported period-end paths rather than a distinct next-FOMC mode and are retained only in the archive check.

For survey date t, professional panel j, and scheduled-FOMC horizon $h \in \{1, \ldots, 7\}$, let $\widetilde{H}^{j}_{t,h}$ denote the reported modal target-rate midpoint. The primary professional forecast is the Market Participant aggregate path. The Primary Dealer aggregate provides a second professional path and the distance between the two aggregates supplies a within-profession disagreement control. Survey cutoff dates define the information available at each forecast origin. The realized target-rate midpoint is attached only after the corresponding target meeting occurs.

The main comparison uses the reported Market Participant aggregate of modal policy-rate forecasts under a stylized post-response access assumption. It does not assume that

individual respondents or outside market participants observe that aggregate at the survey deadline. The separate public-release exercise changes the information set and retains only targets unresolved at its documented publication-date proxy; its timing construction is reported in the Appendix.

The releases also report the 25th and 75th percentiles of respondents' modal paths. Their difference is an observed measure of cross-respondent dispersion, not an individual respondent's subjective uncertainty (Rich and Tracy, 2017). This interquartile range (IQR) is available for all 388 evaluation cells, but it equals zero in 213 cells and its granularity changes when standardized spreadsheets replace the legacy PDF archive. I therefore treat the IQR as a competing observed statistic rather than a structural belief parameter; detailed source-format diagnostics appear in the Appendix.

2.2 Real-Time Public-Information Forecasts

The reference public-information benchmark is a targeted-text Ridge that combines real-time macroeconomic variables, lagged financial-market variables, and targeted measures extracted from Federal Reserve documents. It produces an expanding-window forecast of scheduled-FOMC target-rate changes. The first out-of-sample test survey is in January 2019. A macro-and-market Ridge, labeled "Macro Ridge" in the tables, omits the targeted Federal Reserve text block. A separate macro-only ablation, a broad-text Ridge, an Elastic Net, and additional information-block comparisons are reported in the Appendix. The purpose is not to select the ex post best architecture, but to evaluate a fixed and auditable public-information challenger.

At each survey deadline, macroeconomic inputs use release-dated ALFRED vintages, market observations are dated strictly before the cutoff, and Federal Reserve documents are included only if they were available by the cutoff. Survey-day closes are excluded because daily closing timestamps are not uniformly earlier than survey responses. Delayed Tealbooks and transcripts are also excluded. The Ridge is estimated in expanding windows with

training labels restricted to outcomes already realized at each forecast origin. Imputation and scaling are fitted within the corresponding training samples, and penalty selection is revisited every eight test surveys under the documented recursive schedule. Every training row satisfies

$$\text{realized target date}_{i,h} < \text{survey cutoff}_t.$$

No target realized after a test survey enters the corresponding training sample or recursive risk calculation.

The Federal Reserve documents satisfy the historical cutoffs, but the sentence-level stance classifier was trained on pre-2015 sentences from a labeled corpus released in 2023. The resulting six stance measures are therefore retrospective fixed transformations of cutoff-eligible text rather than components of a system available at the earliest survey origins. A classifier-free sensitivity preserves the substantive comparison. The Appendix documents the feature blocks, transformations, labeled corpus, penalty selection, and sensitivity in detail.

The targeted-text path $M_{t,h}$ is the horizon-h forecast based on information available at survey t. The model begins out-of-sample evaluation only after at least 30 prior survey dates are available for estimation. Recursive large-error probability estimation begins later and requires 20 already realized out-of-sample survey paths in addition to the model's training history. This requirement generates the later 250-cell support from 40 survey paths.

2.3 Monetary-Policy Persistence and Forecast Errors

Let b_t be the target-rate midpoint at the survey origin. For exposition, write $H_{t,h}$, $M_{t,h}$, $N_{t,h}$, and $r_{t,h}$ for the professional, model, no-change, and realized changes from b_t, expressed in basis points. Thus $N_{t,h} = 0$. The canonical pipeline forms these changes before converting each grid-native U.S. path to policy ticks. For a path change x, define

$$\mathcal{T}(x) = \text{round}_{1/2 \text{ away from } 0}(x/25).$$

Writing $q^H_{t,h} = \mathcal{T}(H_{t,h})$ and analogously for q^M, q^N, and q^r, the loss and scores are

$$L^H_{t,h} = 25\big|q^H_{t,h} - q^r_{t,h}\big|,$$

$$D^M_{t,h} = 25\big|q^H_{t,h} - q^M_{t,h}\big|, \qquad D^N_{t,h} = 25\big|q^H_{t,h} - q^N_{t,h}\big|.$$

Here D^M is model disagreement and D^N is no-change distance. The large-error indicator is $Y_{t,h} = \mathbf{1}\{L^H_{t,h} \geq 50\}$, corresponding to two standard policy moves. The 50-basis-point cutoff is a fixed statistical classification, not an estimated welfare or institutional-loss threshold.

The policy-tick convention applies to the Market Participant and Primary Dealer paths, the Ridge challenger, the no-change path, and the realized U.S. midpoint. Federal-funds-futures paths, recursively estimated probabilities, and the foreign applications retain their native continuous values. Frequent ties are economically meaningful in a discrete policy process. All fixed-share comparisons assign uniform fractional weight within a cutoff tie rather than breaking ties arbitrarily.

No-change distance is available at the forecast origin, but on cells where the realized target is unchanged it equals the professional forecast's realized absolute loss, $D^N_{t,h} = L^H_{t,h}$. Such no-move cells constitute 40.5% of the full support and 30.4% of the later support, with four large professional errors in each support. Some of the score's ranking performance therefore follows from an accounting identity. The existing moved-target comparison shows whether the common-share contrast depends on those cells; conditioning on realized movement is an ex post diagnostic, not a forecasting rule available at the origin.

The professional object is a modal path, while the Ridge need not target the same statistical functional. Part of D^M can reflect the difference between a grid-constrained mode and a mean-oriented prediction when predictive mass is asymmetric across the policy grid. The reported results order forecasts under realized-point loss and do not isolate this functional difference from other sources of disagreement.

3 Empirical Comparisons

3.1 Forecast Accuracy

Average point accuracy is summarized by mean absolute error (MAE), root mean squared error (RMSE), bias, and out-of-sample loss relative to no change. Errors follow the common sign convention forecast minus realization. Because the survey path is modal, the Ridge prediction need not target the same functional, and a futures price can include a risk premium, these statistics are comparisons under the displayed realized-point loss criteria rather than a claim that all forecasts target an identical functional. Federal-funds futures are retained as the standard near-term market benchmark (Kuttner, 2001; Gürkaynak et al., 2007; Piazzesi and Swanson, 2008).

3.2 Large-Error Ranking

Raw ranking asks whether larger pre-outcome values of a diagnostic score tend to be associated with large subsequent professional errors. Linear-probability regressions relate $Y_{t,h}$ to standardized model disagreement, adding horizon fixed effects, policy-cycle fixed effects, lead time, Dealer–Participant disagreement, and no-change distance in stages. These regressions describe predictive associations, not causal effects.

ROC-AUC is the probability that a randomly drawn large-error cell receives a strictly higher score than a randomly drawn non-large-error cell, plus one half of the probability of a tied score. Precision–recall AUC (PR-AUC), loss gradients across disagreement quartiles, cross-unit comparisons, and whole-path randomization provide complementary evidence. An estimated AUC above one half indicates favorable raw ranking in the evaluated sample; inferential uncertainty is assessed separately. It does not establish better probability calibration or superior detection relative to a feasible monetary-policy benchmark (Hand, 2009; Shmueli, 2019).

The recursive probability comparison asks a further question. A parsimonious risk speci-

fication uses horizon, policy-cycle indicators, lead time, professional-path magnitude, within-profession disagreement, no-change distance, and prior same-horizon difficulty. Model disagreement is then added to that information set. Policy-cycle indicators are retrospective calendar controls rather than a real-time state classifier, and proper-score comparisons are reported separately from direct score ranking.

3.3 Benchmark-Relative Detection on Common Support

For diagnostic score j and flagged share s, let $A^j_{t,h}(s) \in [0,1]$ denote membership in the top-score flagged set. The weight is binary away from the cutoff and fractional within a cutoff tie. Large-error capture is

$$K_j(s) = \frac{\sum_{t,h} A^j_{t,h}(s) Y_{t,h}}{\sum_{t,h} Y_{t,h}},$$

and the mean absolute error among unflagged forecasts is

$$\mathrm{MAE}_j^{\mathrm{unflagged}}(s) = \frac{\sum_{t,h}\{1 - A^j_{t,h}(s)\} L^H_{t,h}}{\sum_{t,h}\{1 - A^j_{t,h}(s)\}}.$$

Each comparison fixes the support, large-error definition, flagged share, and tie treatment. The 25.0% and 24.8% shares on the two U.S. supports both correspond to retaining three quarters of the cells, with the latter reflecting the integer count of 62 out of 250.

The fixed-share exercise ranks scores jointly within each stated historical evaluation sample; its cutoff is not an origin-by-origin decision rule. Recursive probability estimation and within-survey selection are separate exercises reported on their respective supports.

Global ranking and detection at a stated share are distinct objects. A scalar AUC ordering need not order capture at any particular point of the ranked list, although pointwise receiver operating characteristic dominance is sufficient for capture dominance at every share. The Appendix gives the finite-sample counterexample, the sufficient condition, probability-score comparisons, and the complete fixed-share sensitivity.

The historical exercise evaluates the information contained in alternative pre-outcome scores. It does not observe a warning shown to a forecaster, a subsequent revision, or the effect of additional analysis on the final forecast. Fixed-share capture therefore describes historical error detection, not the causal value of an assisted forecasting process.

3.4 Inference and Support

Two fixed supports appear throughout. The full out-of-sample support contains 388 cells from 60 Market Participant paths and 60 target FOMCs. The later support contains 250 cells from 40 paths and 40 target FOMCs and begins when expanding-window large-error probabilities can be formed after the 20-path burn-in. The later support is nested in the full support. Fixed scores can be evaluated on both; recursive probability specifications require the later support.

Horizon cells are dependent along both survey-path and target-FOMC dimensions. Paired benchmark-relative ranking, probability-score, and fixed-share intervals use 4,999 survey-origin-by-target-FOMC pigeonhole-bootstrap draws, with scores and cutoff-tie weights recomputed within each draw (Owen, 2007; Cameron et al., 2011). Survey-path proper-score differences are averaged within path and evaluated with path-level wild inference. Horizon and policy-cycle comparisons use 4,999 survey-cluster draws that resample complete paths, and complete-path randomization provides a complementary ranking comparison. The methods are not treated as interchangeable.

The two supports differ sharply in policy-cycle composition. The 138 cells present only in the full support are from the pre-pandemic and pandemic/effective-lower-bound periods, whereas 86.0% of the later support falls in rapid tightening or plateau/easing. The support comparison is descriptive. Calendar time and policy state move together, and the design does not identify a regime effect. A normal-reference 27.1-percentage-point detectable difference at the reported operating point is an approximate summary of current-sample precision, not a design-stage power calculation or an equivalence bound. The paired intervals permit

economically meaningful advantages for either score; failure to reject a difference is not evidence of equivalence.

4 Professional Forecast Accuracy and Large Policy-Path Errors

4.1 Professional Forecasts Dominate Average Point Loss

Professional forecasts dominate the public benchmarks in the average-loss comparison (Table 1). The Market Participant MAE is 32.86 basis points and the Primary Dealer MAE is 33.51, compared with 44.27 for federal-funds futures, 49.42 for the Macro Ridge, and 49.61 for the targeted-text Ridge. The professional paths also have lower RMSE. The Ridge nevertheless improves on a flat path: its MAE is 49.61 basis points, compared with 62.11 for no change, and its out-of-sample R^2 relative to no change is 0.326.

The professional aggregate is thus the better point forecast under the reported loss. This comparison matters for interpretation: disagreement with the Ridge is not evidence that the Ridge should replace the professional path. It asks whether a distinct public-information path reveals states in which the more accurate professional forecast incurs a large error.

4.2 Public-Model Disagreement Ranks Large Errors

Model disagreement contains raw ranking information for large professional errors (Table 2). A one-standard-deviation increase in disagreement is associated with a 17.0-percentage-point increase in the probability of a 50-basis-point loss without controls. The coefficient falls to 12.6 with horizon fixed effects, to 9.6 with policy-cycle fixed effects, and to 5.6 after adding no-change distance, Dealer–Participant disagreement, and lead time. No-change distance has a 12.3-point coefficient in the final specification. The attenuation shows that the public-model gap overlaps substantially with transparent measures of forecast difficulty.

The unconditional ranking is economically visible. Model disagreement has ROC-AUC 0.712 and PR-AUC 0.478 for 98 large errors across 388 cells and 60 paths. Market Participant MAE rises from 17.52 basis points in the lowest disagreement quartile to 61.52 in the highest. A cross-unit score that omits the evaluated professional path has AUC 0.705, and human-free and leave-one-professional-out comparisons preserve the qualitative ranking. Whole-path randomization likewise rejects the stated path-randomization null. These results show that the ranking is not generated solely by reusing the evaluated path.

The evidence remains predictive. Absolute disagreement supplies no direction for correction, and its coefficient attenuates once horizon, policy cycle, and no-change distance enter. The question is therefore whether the raw model score adds stable detection beyond the no-change benchmark, not whether professional errors are predictable in an unrestricted sense.

4.3 Comparison with the No-Change Benchmark

Model disagreement has slightly higher AUC and large-error capture than no-change distance on the full sample, but the later sample reverses both orderings (Table 3). When approximately one-quarter of forecasts are flagged, capture is 47.3% versus 43.9% on the full sample and 47.8% versus 52.8% on the later sample. The paired comparisons do not provide clear evidence that either score performs better.

The no-change AUC advantage partly reflects the no-move identity, but the fixed-share ordering remains support-sensitive after the analysis is restricted to cells where the policy rate moved. Moved cells contain 94 of the 98 full-support large errors and 61 of the 65 later-support large errors. There, the no-change AUC falls to 0.523 and 0.691 on the two supports, compared with 0.636 and 0.686 for model disagreement. Recomputing the three-quarter-coverage comparison within moved cells yields capture rates of 37.1% versus 34.6% on the full support and 43.6% versus 50.8% on the later support. Both paired intervals include zero.

Adding the observed modal-path IQR worsens the point estimates relative to the parsimonious horizon-and-no-change probability specification. Adding model disagreement to that IQR-augmented specification does not restore performance. This comparison concerns the published dispersion statistic and the stated recursive models; it does not establish that professional uncertainty is uninformative. The detailed recursive and source-format results appear in the Appendix.

Appendix Figure IA.1 reports the moving-start-date sensitivity. Its point estimate crosses zero as early policy-cycle observations leave the support and later rebounds, while the paired band includes zero throughout. This pattern reinforces support sensitivity without identifying the source of the change.

5 Shared-Direction Errors in Policy-Rate Expectations

5.1 Common-Support Comparisons

Table 3 provides the common-support comparison used in the decomposition. The analysis preserves each support and separates the observations according to their realized forecast-error geometry.

5.2 Which Policy-Path Errors Remain Hidden?

Absolute disagreement measures the distance between forecasts, not their distance from the realization. A large professional error can therefore receive a small disagreement score when the model errs in the same direction and remains close to the professional path. The decomposition measures how such cases contribute to the difference between model-disagreement and no-change selection.

For cell i, let R_i^M and R_i^N denote the fractional assignment weights under model disagree-

ment and no-change distance and define $\delta_i = R_i^M - R_i^N$. Equal capacity implies $\sum_i \delta_i = 0$. With large-error indicator Y_i, professional loss L_i^H, n cells, and common total assignment weight K, the two reported differences satisfy

$$\Delta \,\text{capture} = \frac{\sum_i \delta_i Y_i}{\sum_i Y_i}, \qquad \Delta \,\text{MAE}^{\text{unflagged}} = -\frac{\sum_i \delta_i L_i^H}{n - K}.$$

Positive and negative parts of δ_i generate the respective added masses. Added mass denotes a difference in fractional assignment weight; it does not necessarily identify a forecast selected by one score and entirely excluded by the other.

At the common flagged share on the 388-cell support, each score relatively adds 47.11 cell-equivalents (Table 4). The model-added and no-change-added masses have similar mean horizons, 4.11 and 4.14, but different path directions. The professional and Ridge paths in the model-added mass call on average for a 20.35-basis-point decline and an 11.07-basis-point increase. In the no-change-added mass they call for increases of 20.27 and 24.95 basis points.

The model-added mass has a mean professional absolute error of 37.98 basis points and a 32.7% large-error rate, compared with 27.16 basis points and 25.5% for the no-change-added mass. These differences account arithmetically for the model score's 3.49-point capture advantage on the full support. They do not imply that the Ridge supplies the correct alternative path. The Ridge is closer to the realization in only 7.4% of model-added mass, and 80.2% of the large errors in that mass place the two forecasts on the same side of the realization.

The composition reverses on the later support, where 29.14 cell-equivalents differ between the 62 expected flags. The model-added mass has a 22.4% large-error rate, compared with 33.7% for the no-change-added mass. Among their large errors, the professional and Ridge paths miss in the same direction 52.6% and 99.2% of the time, respectively. No-change distance consequently assigns more weight to the shared-direction errors associated with large tightening paths in that sample.

Let $e_i^H = q_i^H - q_i^r$ and $e_i^M = q_i^M - q_i^r$ denote signed tick errors. The identity $D_i^M = 25|e_i^H - e_i^M|$ shows that when the two nonzero errors have the same sign and $|e_i^M| < |e_i^H|$, disagreement equals $25(|e_i^H| - |e_i^M|)$. The identity does not imply that every same-direction case has a small score or falls outside the flagged set.

The same-direction, model-closer category contains 14 large professional errors on the full support and 12 on the nested later support. Its net contribution to model-minus-no-change capture is -3.75 and -4.10 percentage points, respectively. These are contributions from differences in assignment weights, not counts of wholly missed forecasts. The decomposition identifies which realized error configurations account for the observed selection difference; it does not identify why professional and model forecasts err on the same side of the realization.

5.3 Horizon and Policy-Cycle Composition

Large errors concentrate at longer horizons and vary across policy cycles (Figure 1). Large-error prevalence is 9.4% in H2–H3 and 40.8% in H4–H7; H1 contains only one large error. Model-minus-no-change AUC differences change sign across the displayed horizon and policy-cycle groups and remain imprecise. The joint horizon interaction has $p = 0.593$. Although the policy-cycle interaction is statistically nonzero, it describes variation in the raw association rather than stable benchmark dominance.

The support composition is consequential. The 138 cells added by the full support lie entirely in the pre-pandemic and pandemic/effective-lower-bound periods, with 63 and 75 cells, respectively. Thirty-two of the 33 large errors in this added block are pre-pandemic. The later support contains no pre-pandemic cells and assigns 86.0% of its observations to rapid tightening or plateau/easing. The corresponding additive contributions locate the change in capture: on the full support, H4–H7 and pre-pandemic cells contribute 4.57 and 10.14 percentage points to the model-minus-no-change difference, while rapid tightening contributes -8.47 points; on the later support, rapid tightening contributes -12.52 points. The contributions sum to the pooled differences by construction and do not identify a policy-

cycle mechanism.

An exploratory within-survey comparison in the Appendix flags the top one or two eligible horizons within each path and confirms that horizon is a demanding transparent benchmark.

6 Evidence from Other Central-Bank Policy-Rate Surveys

The U.S. result is not a universal machine effect. The European Central Bank (ECB) and Bank of England applications ask whether the same distinction between raw disagreement and the no-change benchmark appears in other professional policy-rate surveys.

The ECB application uses the published median Survey of Monetary Analysts path and a public-history challenger available at each origin. Professional–challenger disagreement has AUC 0.668 for 64 large errors among 308 cells, compared with 0.695 for no change. The Bank of England application uses modal Bank Rate paths from 37 Market Participants Surveys. Its corresponding AUCs are 0.777 and 0.795 for 64 large errors among 261 cells, and its challenger-to-professional MAE ratio is 1.70. In both settings the professional path has lower realized-point MAE than the challenger.

The fixed-share comparison gives the same capture ordering (Table 5). At the separately calculated approximately 25% flagged shares, professional–challenger and no-change capture are 47.8% and 52.8% in the United States, 43.8% and 49.1% in the ECB application, and 56.2% and 57.9% in the Bank of England application. MAE among unflagged forecasts is not ordered uniformly: model disagreement has the lower point estimate in the ECB application, while no change has the lower estimate in the United States and Bank of England applications.

The surveys, horizons, challengers, and policy regimes differ, so the three environments are not a pooled validation sample and the rows do not estimate one structural effect. A Philadelphia Fed inflation application, reported in the Appendix, does not reproduce the

raw ranking. The evidence supports a bounded regularity in policy-rate forecasting: model disagreement can rank professional errors, but the paired comparisons do not establish an incremental capture advantage for the evaluated scores over no-change distance.

7 Conclusion

Professional policy-rate forecasts can dominate a public-information model in average accuracy while disagreement with that model still ranks large forecast errors. In the New York Fed Market Participant paths, the professional aggregate has substantially lower mean absolute error than the Ridge challenger, but the absolute gap between the two identifies difficult horizon cells. Average forecast quality and diagnostic disagreement are therefore distinct properties.

The model signal does not, however, have an established advantage over a no-change policy path. Its relative AUC and fixed-share capture change across the full and later evaluation supports, and the paired comparisons are imprecise. The error decomposition traces part of those differences to cases in which professional and model paths fail in the same direction, including cases in which the model is closer to the realization. Similar raw-ranking evidence, without an established fixed-share capture advantage relative to no change, appears in the ECB and Bank of England applications.

The implication for monetary-policy expectations is that external models should be evaluated against no-change distance and horizon on common support. A model may reveal forecast difficulty without supplying a direction of correction or a superior classification rule. The historical data do not observe warnings, revisions, or the effect of additional analysis on final forecasts. The contribution is instead to identify which professional policy-path errors public-information benchmarks reveal and which remain hidden when forecasts fail together.

References

Agarwal, N., A. Moehring, P. Rajpurkar, and T. Salz (2023). Combining human expertise with artificial intelligence: Experimental evidence from radiology. Working Paper 31422, National Bureau of Economic Research.

Agarwal, N., A. Moehring, and A. Wolitzky (2025, June). Designing Human–AI collaboration: A sufficient-statistic approach. Working Paper 33949, National Bureau of Economic Research.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2011). Robust inference with multiway clustering. *Journal of Business & Economic Statistics 29*(2), 238–249.

Coibion, O. and Y. Gorodnichenko (2015). Information rigidity and the expectations formation process: A simple framework and new facts. *American Economic Review 105*(8), 2644–2678.

El-Yaniv, R. and Y. Wiener (2010). On the foundations of noise-free selective classification. *Journal of Machine Learning Research 11*(53), 1605–1641.

Faust, J. and J. H. Wright (2009). Comparing Greenbook and reduced form forecasts using a large realtime dataset. *Journal of Business & Economic Statistics 27*(4), 468–479.

Federal Reserve Bank of New York (2026). Survey of Market Expectations.

Giacomini, R. and B. Rossi (2010). Forecast comparisons in unstable environments. *Journal of Applied Econometrics 25*(4), 595–620.

Gürkaynak, R. S., B. Sack, and E. T. Swanson (2007). Market-based measures of monetary policy expectations. *Journal of Business & Economic Statistics 25*(2), 201–212.

Hand, D. J. (2009). Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Machine Learning 77*(1), 103–123.

Kar, P., H. Narasimhan, and P. Jain (2015). Surrogate functions for maximizing precision at the top. In *Proceedings of the 32nd International Conference on Machine Learning*, Volume 37 of *Proceedings of Machine Learning Research*, pp. 189–198. PMLR.

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.

Owen, A. B. (2007). The pigeonhole bootstrap. *The Annals of Applied Statistics 1*(2), 386–411.

Pedersen, M. (2025). Judgment in macroeconomic output growth predictions: Efficiency, accuracy and persistence. *International Journal of Forecasting 41*(2), 475–486.

Piazzesi, M. and E. T. Swanson (2008). Futures prices as risk-adjusted forecasts of monetary policy. *Journal of Monetary Economics 55*(4), 677–691.

Rich, R. W. and J. Tracy (2017, February). The behavior of uncertainty and disagreement and their roles in economic prediction: A panel analysis. Staff Report 808, Federal Reserve Bank of New York.

Sastry, K. A. (2026). Disagreement about monetary policy. *American Economic Journal: Macroeconomics 18*(2), 486–522.

Shah, A., S. Paturi, and S. Chava (2023). Trillion dollar words: A new financial dataset, task & market analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6664–6679. Association for Computational Linguistics.

Shmueli, G. (2019). Lift up and act! classifier performance in resource-constrained applications.

Tables

Table 1: Accuracy of Professional and Public Policy-Rate Forecasts

Forecast source	Cells	Surveys	MAE	RMSE	Bias	MAE / Participant MAE
Market Participant	388	60	32.86	64.14	-6.57	1.00
Primary Dealer	388	60	33.51	64.64	-5.67	1.02
Fed funds futures	388	60	44.27	67.47	-21.38	1.35
Macro Ridge	388	60	49.42	86.26	-1.35	1.50
Targeted-text Ridge	388	60	49.61	86.97	-0.64	1.51

Notes: The common out-of-sample panel contains 388 horizon cells across 60 survey paths. Errors are the forecast minus the realized target midpoint in basis points: $H - r$ for professional forecasts and $M - r$ for model forecasts. Because each forecast and the realized target are expressed as rate midpoints, that difference equals the error in the implied change from the rate in force. This is not a comparison of forecasts elicited for a common statistical functional. The professional paths have lower realized-point MAE and RMSE than the public-information benchmarks. The subsequent analysis asks whether disagreement with a less accurate benchmark nevertheless identifies large professional policy-path errors. "Macro Ridge" denotes the macroeconomic-and-lagged-market benchmark without the targeted Federal Reserve text block; the macro-only ablation is reported separately in Appendix Table IA.17.

Table 2: Public-Model Disagreement and Large Professional Policy-Path Errors

	(1)	(2)	(3)	(4)
Targeted disagreement	16.995***	12.582***	9.647**	5.612**
	(2.967)	(3.706)	(3.912)	(2.838)
No-change distance				12.311***
				(3.813)
Horizon fixed effects	No	Yes	Yes	Yes
Policy-cycle fixed effects	No	No	Yes	Yes
Other forecast controls	No	No	No	Yes
Observations	388	388	388	388
Survey paths	60	60	60	60
Target FOMCs	60	60	60	60
Adjusted R^2	0.151	0.260	0.334	0.379

Notes: The dependent variable equals one when the Market Participant forecast incurs an absolute error of at least 50 basis points. Coefficients from linear-probability regressions are percentage-point changes per one-standard-deviation increase in the indicated distance. Parentheses report standard errors clustered by survey path and target FOMC. Columns add H1–H7 fixed effects, policy-cycle fixed effects, Dealer–Participant disagreement, days to the target FOMC, and no-change distance as shown. Underlying forecasts and source documents satisfy the survey-deadline cutoff; stance-derived features and policy-cycle indicators are retrospective transformations or calendar controls. The estimates describe a predictive association, not a contemporaneously deployable historical system or a causal effect; Appendix Table IA.10 evaluates recursive probability performance relative to transparent horizon, persistence, and survey-dispersion variables. The attenuation from Column (1) to Column (4), together with the larger no-change coefficient in Column (4), shows that the model-derived diagnostic score contains conditional ranking information but overlaps substantially with transparent forecast difficulty. *, **, and *** denote statistical significance at the 10, 5, and 1 percent levels, respectively, based on two-sided tests using those clustered standard errors.

Table 3: Model Disagreement versus a No-Change Policy Path

Panel A. Point-forecast accuracy

Support	Cells	Surveys	Large losses	Professional MAE	Model MAE	No-change MAE
Development	388	60	98	32.86	49.61	62.11
Recursive-aligned	250	40	65	33.70	50.00	73.00

Panel B. Large-error ranking

Support	Model AUC	No-change AUC	Difference	95% interval
Development	0.712	0.700	0.012	[−0.103, 0.124]
Recursive-aligned	0.709	0.777	-0.067	[−0.208, 0.050]

Panel C. Large-error capture at a common flagged share

Support	Flagged share (%)	Model capture (%)	No-change capture (%)	Difference [95% interval] (percentage points)
Development	25.0	47.3	43.9	3.5 [−13.6, 17.4]
Recursive-aligned	24.8	47.8	52.8	−5.0 [−26.2, 11.8]

Panel D. Mean absolute error among unflagged forecasts

Support	Model (bp)	No change (bp)	Difference [95% interval] (bp)
Development	23.31	25.06	−1.75 [−7.30, 3.22]
Recursive-aligned	23.52	21.90	1.63 [−2.85, 6.99]

Notes: Panel A evaluates the professional path, the targeted-text Ridge path, and a no-change path on identical observations; MAE is in basis points. Panels B–D compare model disagreement, the absolute professional-minus-model path gap, with no-change distance, the absolute professional path change. A large loss is an absolute professional forecast error of at least 50 basis points. At the reported flagged shares, uniform fractional weights are assigned within cutoff ties. Differences are model disagreement minus no change. Intervals use 4,999 paired survey-origin-by-target-FOMC pigeonhole-bootstrap draws. The development-sample AUC interval comes from the exact-score ranking calculation; the recursive AUC and fixed-share intervals use the common allocation draw bank. Membership in a flagged set is a historical classification and does not denote an observed forecast revision.

Table 4: Forecast-Error Geometry and Shared-Direction Failures

Panel A. Composition of differently weighted cases

Support	Relatively added by	Replacement mass	Mean horizon	H4–H7 share (%)	Professional path (bp)	Model path (bp)	Professional MAE (bp)
Development	Model disagreement	47.11	4.11	59.4	-20.35	11.07	37.98
Development	No change	47.11	4.14	63.1	20.27	24.95	27.16
Recursive-aligned	Model disagreement	29.14	4.28	64.1	-23.32	-2.67	23.74
Recursive-aligned	No change	29.14	4.16	64.0	60.62	62.08	34.23

Panel B. Error characteristics of differently weighted cases

Support	Relatively added by	Large-loss rate (%)	Model closer share (%)	Common-direction share (%)	Common direction among large losses (%)
Development	Model disagreement	32.7	7.4	41.1	80.2
Development	No change	25.5	27.2	38.7	100.0
Recursive-aligned	Model disagreement	22.4	7.2	21.3	52.6
Recursive-aligned	No change	33.7	32.3	47.1	99.2

Panel C. Additive contribution by forecast-error geometry

Support	Error geometry	Cells	Losses	Model-added mass	No-change-added mass	Δ capture (pp)	Δ MAE among unflagged forecasts (bp)
Development	Human exact Model exact;	211	0	17.31	20.15	0.00	0.00
Development	human nonzero	29	5	3.08	8.74	2.41	0.28
Development	Opposite directions Same direction;	14	3	7.37	0.00	0.70	-0.69
Development	model closer Same direction;	14	14	0.39	4.07	-3.75	0.51
Development	model not closer	120	76	18.97	14.15	4.13	-1.86
Recursive-aligned	Human exact Model exact;	117	0	12.32	8.67	0.00	0.00
Recursive-aligned	human nonzero	28	4	2.10	6.75	3.10	0.35
Recursive-aligned	Opposite directions Same direction;	14	3	8.53	0.00	1.54	-1.27
Recursive-aligned	model closer Same direction;	12	12	0.00	2.67	-4.10	0.71
Recursive-aligned	model not closer	79	46	6.19	11.06	-5.59	1.83

Notes: Let R_i^M and R_i^N denote the fractional assignment weights generated by model disagreement and no-change distance, respectively, and let $\delta_i = R_i^M - R_i^N$. Model-added and no-change-added masses are the positive and negative parts of δ_i. Added mass is a difference in fractional assignment weight and does not necessarily identify a forecast selected by one score and entirely excluded by the other. Professional, model, and realized path changes are first converted to 25-basis-point policy ticks using round-half-away-from-zero. Common-direction cases have professional and model errors with the same nonzero sign; "model closer" requires a strictly smaller absolute tick error. In Panel C, capture contributions equal $\Delta\,\text{Capture}_{g,\text{pp}} = 100\sum_{i\in g}\delta_i Y_i / \sum_{i=1}^{n} Y_i$. Contributions to MAE among unflagged forecasts equal $\Delta\,\text{MAE}_{\text{unflagged},g} = -\sum_{i\in g}\delta_i L_i^H/(n-K)$, where $L_i^H = 25|e_i^H|$ is professional loss in basis points, g indexes the displayed error category, and n and K are the cell count and common total assignment weight for the stated evaluation support. The rows therefore sum to the corresponding matched-support differences in Table 3. The decomposition is descriptive and does not identify a common information shock or a causal effect.

Table 5: Professional Policy-Rate Forecasts across Survey Environments

Panel A. Forecast accuracy and raw loss ranking

Environment	Ranking support cells / origins	Challenger / professional MAE	Professional–challenger AUC	No-change AUC	Professional–challenger minus no change AUC [95% CI]
United States	388 / 60	1.51	0.712	0.700	0.012 [-0.103, 0.124]
ECB SMA	308 / 42	1.79	0.668	0.695	-0.027 [-0.203, 0.112]
BoE MaPS	261 / 37	1.70	0.777	0.795	-0.017 [-0.128, 0.092]

Panel B. Large-error detection at an approximately 25% flagged share

Environment	Detection support cells / origins	Flagged N (share)	Professional–challenger capture (%)	No-change capture (%)	Professional–challenger MAE among unflagged forecasts	No-change MAE among unflagged forecasts
United States	250 / 40	62 (24.8%)	47.8	52.8	23.5	21.9
ECB SMA	308 / 42	77 (25.0%)	43.8	49.1	23.91	25.84
BoE MaPS	261 / 37	66 (25.3%)	56.2	57.9	17.27	16.95

Notes: A large policy-rate loss is an absolute professional forecast error of at least 50 basis points. Panel A reports average-loss and ranking statistics on each environment's stated ranking support. The MAE ratio is challenger MAE divided by professional MAE. Professional–challenger AUC uses absolute professional–challenger disagreement to rank subsequent large professional losses; no change uses the professional forecast's distance from an unchanged policy path. AUC-difference intervals use 4,999 pigeonhole-bootstrap draws over forecast origins and target meetings. Panel B is a separate common-share comparison within each environment. It shows the detection support, expected flagged count, realized flagged share, large-loss capture, and professional MAE among unflagged forecasts. The displayed U.S. row is reported on the 250-cell recursive support to align it with the recursive probability exercise. The same fixed-score comparison is feasible on all 388 cells and is reported in Main Table 3. ECB and BoE use one support for ranking and fixed-share detection. Standalone professional–challenger AUC intervals are reported in Appendix Tables IA.7, IA.22, and IA.23 on the corresponding supports. The rows are not a common-sample comparison. The ECB and BoE detection rows contain 46 and 36 target meetings, respectively. Uniform fractional assignment at cutoff ties preserves the stated flagged share. The U.S. professional, challenger, and no-change paths use the path-first integer policy-tick convention, matching Main Table 3; the ECB and BoE applications retain their native continuous values. Flagging is a historical classification, not an observed correction. The environments differ in survey design, horizon structure, challenger, and policy regime; the rows are descriptive scope comparisons rather than estimates of one pooled structural effect. Raw ranking recurs in the three policy-rate settings, but the no-change rule has the higher capture point estimate in each, separating a recurring diagnostic pattern from benchmark-relative fixed-share detection.

Figures

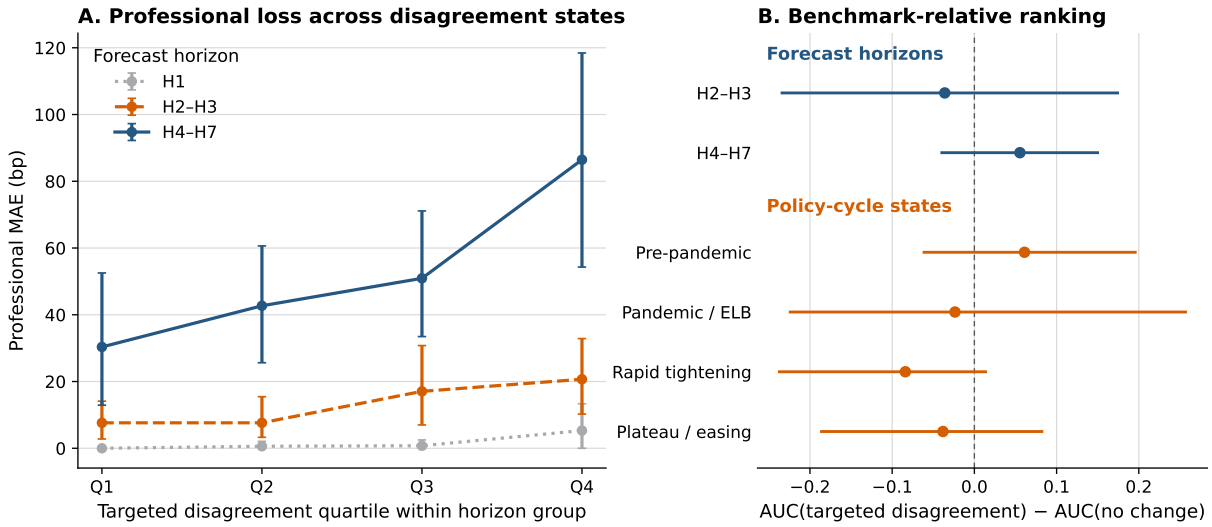


Figure 1: Professional Policy-Path Errors and Benchmark-Relative Ranking across Horizons and Policy Cycles

Notes: Panel A reports professional MAE across quartiles of path-first integer-tick targeted disagreement formed separately within H1, H2–H3, and H4–H7. Score ties that cross a quartile boundary receive uniform fractional membership. Vertical bars are percentile intervals from 4,999 survey-cluster bootstrap draws that resample complete professional paths. Panel B reports targeted-minus-no-change AUC differences and survey-cluster bootstrap 95% confidence intervals that resample complete survey paths for the two supported horizon groups and all four policy-cycle states. Positive values favor targeted disagreement. H1 is shown descriptively in Panel A but is excluded from Panel B because it contains only one 50-basis-point loss. The development sample has 388 cells across 60 survey paths. The comparisons describe forecast difficulty and ranking heterogeneity; they do not identify causal effects.

Appendix to

"Disagreement and Errors in Professional Policy-Rate Expectations: Evidence from New York Fed Policy-Rate Paths"

A-Young Park, Ph.D.

Financial Supervisory Service

September 2026

Part A. Real-Time Data and Forecast Construction

A1. Sample Construction

Table IA.1: Sample Flow for the U.S. Analyses

Panel A. Analysis samples

Analysis	Cells	Survey paths	Targets	Origin dates	Inclusion rule
Common forecast panel	388	60	60	2019-01-22–2026-06-08	Common professional, market, and model support; target dates 2019-01-30–2026-06-17.
Recursive probability/common-support allocation sample	250	40	40	2021-07-19–2026-06-08	Twenty realized out-of-sample survey paths provide the minimum recursive risk history; all labels must be realized before each test survey. Fixed-score allocation uses this support for matched comparisons. Target dates 2021-07-28–2026-06-17.

Panel B. Path parser and common-support construction

Stage	Deterministic rule
Legacy PDF text	Convert official result PDFs with layout preservation; identify panel and response deadline; locate the modal target-rate question.
Legacy path tables	Read the reported 25th, median, and 75th percentile rows and response counts; map dated columns to scheduled FOMC meetings; reject missing or duplicate meeting dates.
Range-era and image tables	Average reported top- and bottom-of-range percentiles when the table covers all respondents. Flag range-only respondent summaries as ineligible. Admit image-only tables only through page-cited visual entries.
Standardized releases	Parse the released spreadsheet dates and percentiles, classify FOMC-meeting versus quarter-, half-year-, and year-end columns, sort by target date, and number FOMC columns within survey and panel.
Analysis join	Retain H1–H7 cells with Market Participant and Primary Dealer paths, realized target midpoint, model forecasts, and cutoff-safe market path all observed.

Panel C. Exact complement of the 388-cell common panel in the 60×7 *grid*

Survey deadline	Excluded horizons	Survey deadline	Excluded horizons
2020-06-01	H7	2024-06-03	H7
2021-06-07	H7	2024-07-22	H6, H7
2022-06-06	H7	2025-10-20	H7
2022-10-24	H7	2025-12-01	H6, H7
2023-06-05	H7	2026-01-20	H5–H7
2024-01-22	H7	2026-03-09	H4–H7
2024-03-11	H7	2026-04-20	H3–H7
2024-04-22	H7	2026-06-08	H2–H7

Notes: Panel C lists 32 excluded (t, h) pairs. The 60 observed survey deadlines define 420 potential H1–H7 cells; removing exactly those pairs leaves 388. The list is the set complement after the common-support join and does not assign an unobserved reason to a missing cell. Recursive probability forecasts require the later 250-cell support after the 20-path burn-in; fixed-score allocation is reported on both the 250-cell and full 388-cell score-only supports. The reduction from 388 to 250 cells reflects the requirement of 20 realized out-of-sample survey paths for recursive risk estimation, in addition to the challenger model's own training history. The 250-cell allocation comparison aligns the fixed-score rules with the recursive probability support; fixed-score allocation itself does not require this burn-in. It is not an outcome-selected subsample.

A2. Public-Information Challenger Specification

Table IA.2: Public-Information Challenger Specification

Component	Construction	Timing or estimation rule
Forecast target	Scheduled-FOMC target-rate midpoint change, with reported path values placed on the 25-basis-point policy grid	Horizon cells H1–H7
Macroeconomic block	168 transformations from 14 release-dated real-time series covering output, housing, production, retail sales, prices, employment, participation, earnings, and claims	ALFRED vintage by survey date; release date no later than the survey cutoff
Financial-market block	113 transformations from 10 policy and market series covering the effective funds rate, Treasury rates and slope, credit spread, equities, dollar, volatility, oil, and financial conditions	Observation date strictly precedes the survey cutoff; survey-day closes excluded
Federal Reserve text block	14 targeted measures from statements, minutes, press-conference transcripts, SEP, Beige Book, implementation notes, strategy statements, speeches, and testimony	Latest-document and 30-day measures available by the cutoff; delayed Tealbooks and transcripts excluded
Text formula	Net hawkishness equals mean posterior hawkish probability minus mean posterior dovish probability; topic measures are dictionary counts per 1,000 tokens	Exponential half-life is 15 days for 30-day speech/public-remarks aggregates and 90 days for the core bundle
Horizon and context block	FOMC horizon, squared horizon, days to target meeting, and current target-rate midpoint	Four features observed at the forecast origin
Estimator	Ridge with LSQR solver, tolerance 10^{-4}, median imputation, and standardization	All preprocessing is fitted within the relevant training fold
Named inputs and indicators	299 named inputs: 281 macro and market, 14 targeted text, and four horizon or context variables	Missingness indicators are generated fold by fold and sit outside the 299 named-input count; zero, one, or two indicators enter across outer folds
Penalty selection	$\alpha \in \{1, 10, 100, 1000\}$ chosen by mean absolute validation loss	Three expanding time-series validation blocks; retuned every eight test surveys
Training sample	Expanding window with at least 30 prior survey dates	A label enters only when its target FOMC date precedes the test survey cutoff
Elastic Net comparison	$(\alpha, \ell_1 \text{ ratio}) \in \{(0.01, 0.05), (0.1, 0.05), (1, 0.05), (1, 0.25)\}$; cyclic coordinate descent, 5,000 iterations, tolerance 10^{-3}	Same folds, timing rules, and retuning frequency as Ridge
Multi-model comparison	Cell-wise median of macro-and-market Ridge, targeted-text Ridge, broad-text Ridge, and Elastic Net macro-plus-text paths	Reported separately from the reference targeted disagreement score

The 14 macro series are GDPC1, HOUST, INDPRO, RSAFS, CPIAUCSL, CPILFESL, PCEPI, PCEPILFE, CES0500000003, CIVPART, EMRATIO, ICSA, PAYEMS, and UNRATE. The 10 market series are BAA10Y, the stitched broad-dollar series, DCOILWTICO, DFF, DGS10, DGS2, NASDAQCOM, NFCI, T10Y2Y, and VIXCLS. The targeted block contains the core-document net-hawkishness measure; eight core-document topic frequencies for inflation, labor, growth, financial conditions, uncertainty, forward guidance, tightening, and easing; latest-statement, latest-minutes, and latest-press- conference-transcript net hawkishness; and 30-day speech and public-remarks net hawkishness.

The 168 macro transformations apply the following twelve quantities to each of the 14 release-dated series. Let x_0 be the latest cutoff-eligible observation, x_1 and x_2 the preceding two releases, and x_d the latest observation dated no later than d calendar days before the date of x_0. The features are x_0, x_1, x_2, $x_0 - x_1$, $100(x_0 - x_1)/|x_1|$, and, for $d \in \{90, 180, 365\}$, both $x_0 - x_d$ and $100(x_0 - x_d)/|x_d|$, together with the calendar days between the survey cutoff and the date of x_0. A percentage change is missing when its denominator is zero. The financial-market block applies the same twelve transformations to nine series. The stitched broad-dollar index contributes only the four percentage changes and the release lag because the pre- and post-February-2019 component indexes have different bases. Thus $14 \times 12 = 168$ and $9 \times 12 + 5 = 113$.

The stance classifier uses the FOMC communication corpus released by Shah et al. (2023). The source files contain 1,984 training and 496 test sentences; retaining observations dated no later than 2014 leaves 1,283 training and 325 test sentences. The sentence-level TF–IDF logistic regression is fit only on the 1,283 training sentences, while the 325 test sentences are used only for classification diagnostics. Labels 0, 1, and 2 denote dovish, hawkish, and neutral sentences. The vectorizer lowercases text, strips Unicode accents, uses one- and two-word ngrams, requires document frequency of at least two, excludes terms appearing in more than 98% of labeled sentences, caps the vocabulary at 40,000 features, applies sublinear term frequency, and uses ℓ_2 normalization. The balanced logistic regression uses $C = 2$, the LBFGS solver, a 3,000-iteration limit, and random state 2026. For a document, posterior probabilities are averaged across sentence chunks and net hawkishness is the mean hawkish posterior minus the mean

dovish posterior.

The labeled corpus was released in 2023, after the forecast-evaluation sample begins. The stance mapping is therefore a retrospective fixed text transformation applied to Federal Reserve documents available by each historical cutoff, rather than a classifier that was operationally available at the earliest survey dates.

The topic dictionary uses case-insensitive whole-word matches. Inflation contains inflation, price stability, price pressure(s), core PCE, PCE, CPI, inflation expectation(s), disinflation, and deflation. Labor contains labor market, employment, unemployment, payroll(s), job(s), wage(s), participation rate, and maximum employment. Growth contains growth, economic activity, output, GDP, demand, consumer spending, consumption, investment, recession, and slowdown. Financial conditions contains financial conditions, credit conditions, lending, liquidity, Treasury yield(s), yield curve, credit spread(s), balance sheet, and asset price(s). Uncertainty contains uncertain or uncertainty, risk(s), may, might, could, contingent, depend or depends, monitor or monitoring, data-dependent, and range of outcomes. Forward guidance contains forward guidance, future path, policy path, target range, remain/remains/remained, continue/continues/continued, until, expect/expects, anticipate/anticipates, appropriate, and will be prepared. Tightening contains tighten/tightening/tightened, restrictive, rate hike(s), raise rates, higher rates, reduce the balance sheet, and quantitative tightening. Easing contains ease/easing/eased, accommodative, rate cut(s), lower rates, reduce rates, quantitative easing, and asset purchase(s). Counts are divided by the document's alphabetic-token count and multiplied by 1,000.

For the two targeted 30-day speech aggregates, the exponential weight on a document aged a days is proportional to $\exp[-\log(2)a/15]$, so the half-life is 15 days. The latest-core-document bundle uses the same formula with a 90-day half-life. Latest-statement, latest-minutes, and latest-press-conference measures select one document and therefore receive no age weighting.

The first outer test fold (January 22, 2019) begins its three validation blocks on June 6, 2016, April 24, 2017, and March 12, 2018. Because the splits expand with the outer training sample, the three cut dates are recomputed when tuning recurs; for the last retuning window

(January 20, 2026) they begin on March 12, 2018, October 26, 2020, and June 5, 2023. The two possible fold-local indicators correspond to missing T10Y2Y three-month percentage change and missing 30-day speech net hawkishness. The challenger is fixed for the reported evaluation and is not selected as the most accurate model in the sample.

An exploratory sensitivity removes all six net-hawkishness measures that depend on the labeled stance corpus while retaining the eight dictionary topic frequencies, 281 macro and lagged-market variables, four context variables, and the original estimation and evaluation procedures. The resulting 293-input Ridge has an MAE of 49.36 basis points, compared with 49.61 for the reference challenger, on the 388-cell support. On the 250-cell common support, its disagreement score has ROC-AUC 0.690 and assigns 45.6% of large losses to the flagged set at the 24.8% flagged share, leaving 23.64 basis points of mean absolute error among unflagged cells. The corresponding values are 0.709, 47.8%, and 23.52 basis points for the reference challenger, and no change assigns 52.8% of large losses to the flagged set while leaving 21.90 basis points of error among unflagged cells. The classifier-free-minus-no-change differences are -7.2 percentage points for capture (95% paired interval $[-30.4, 11.2]$) and $+1.74$ basis points for MAE among unflagged cells ($[-2.75, 7.64]$). Thus removing the retrospectively fitted stance block does not produce a common-share detection advantage over no change. These estimates were developed after the main comparisons and should not be interpreted as independent validation; the intervals establish neither equivalence nor population inferiority.

A3. Integer Policy-Tick Outcome Construction

Table IA.3 applies the path-first transformation in the main text: each grid-native U.S. forecast and realization is converted from basis points to integer 25-basis-point policy ticks before losses or disagreement scores are formed. The resulting tick scores, or their equivalent basis-point multiples, are the primary inputs to U.S. ranking, fixed-share detection, bootstrap, and path-randomization analyses. The stored floating-point calculation is reported only as a numerical sensitivity. Futures paths, probability and residual scores, and the foreign applications retain their native continuous values; the median of four tick-valued model paths may lie on half-ticks.

Table IA.3: Numerical Definitions and Ranking Performance

Definition	Cells	Large losses	Model-disagreement AUC	No-change AUC	Difference
Path-first integer ticks	388	98	0.712	0.700	0.012
Grid-tolerant labels, stored scores	388	98	0.733	0.694	0.039
Literal floating-point labels and scores	388	77	0.695	0.620	0.075

Notes: The path-first construction converts each grid-native professional, model, and realized path change to integer 25-basis-point ticks using round-half-away-from-zero before forming forecast errors or disagreement scores. The grid-tolerant construction retains the 98-event primary label but applies the stored continuous scores. The literal construction applies the unrounded stored scores and the literal floating-point comparison $|e_i^H| \geq 50$. Twenty-one economically grid-equivalent observations lie infinitesimally below the literal threshold. Difference is model-disagreement AUC minus no-change AUC. Each row uses internally consistent event labels and scores.

A4. Point-Forecast Performance Relative to No Change

Table IA.4 supplements Main Table 1 with benchmark-relative realized-point loss. The targeted-text Ridge is less accurate than the Market Participant forecast but more accurate than a flat no-change path. This is a forecast-production comparison; it does not evaluate the disagreement score used for fixed-share large-error detection. The adjusted-loss calculation reported in the underlying result files is not presented as a formal Clark–West test because the evaluated regularized cell-panel forecast is not a classical nested-model comparison.

Table IA.4: Point-Forecast Performance Relative to Professional and No-Change Forecasts

Panel A. Realized-point forecast performance

Forecast	MAE	RMSE	Bias	OOS R^2 vs. Participant	OOS R^2 vs. no change
Market Participant	32.86	64.14	-6.57	0.000	0.633
Primary Dealer	33.51	64.64	-5.67	-0.016	0.628
Fed funds futures	44.27	67.47	-21.38	-0.106	0.594
Macro Ridge	49.42	86.26	-1.35	-0.809	0.337
Targeted-text Ridge	49.61	86.97	-0.64	-0.839	0.326
No change	62.11	105.91	-10.18	-1.727	0.000

Panel B. Paired loss differences for the targeted-text Ridge

Benchmark	Loss	Mean loss difference	95% interval		Cells
			Two-way	Pigeonhole	
Market Participant	Absolute error	16.75	[9.37, 24.13]	[9.38, 25.85]	388
Market Participant	Squared error	3,450.39	[1,047.66, 5,853.11]	[1,408.51, 6,733.63]	388
No change	Absolute error	-12.50	[-24.63, -0.37]	[-26.65, -0.19]	388
No change	Squared error	-3,653.35	[-7,131.66, -175.04]	[-8,019.84, -642.92]	388

Notes: All rows use the common 388-cell out-of-sample panel across 60 survey paths and 60 target FOMCs. Errors are forecast minus realized policy-path change in basis points; squared-error differences are in squared basis points. Out-of-sample R^2 is $1 - \sum e^2_{\text{candidate}} / \sum e^2_{\text{benchmark}}$. Negative paired loss differences favor the candidate. Two-way intervals use survey-path and target-FOMC clustering; pigeonhole intervals use 4,999 paired draws. The comparison concerns realized-point forecast production and does not establish warning-rule or allocation value. "Macro Ridge" denotes the macroeconomic-and-lagged-market benchmark without the targeted Federal Reserve text block; the macro-only ablation is reported separately in Table IA.17.

Part B. Benchmark-Relative Forecast Evaluation

B1. Threshold and Numerical Sensitivity

Table IA.5 reports the fixed operational definition together with literal-comparison and stored-value sensitivities. These alternatives can change numerical ties but do not alter the information set or the reference targeted-text model.

Table IA.5: Numerical Sensitivity of the Operational 50-Basis-Point Threshold

Definition	Misses	Targeted AUC	Median AUC	No-change AUC	Dealer gap AUC
Path-first integer ticks	98	0.712	0.709	0.700	0.551
Grid-tolerant labels / stored scores	98	0.733	0.711	0.694	0.555
Literal floating-point $\geq$ 50bp	77	0.695	0.691	0.620	0.479

Notes: Twenty-one theoretically exact 50-basis-point losses are stored infinitesimally below 50 because of floating arithmetic. The primary row converts each grid-native U.S. path to integer policy ticks before forming either losses or ranking scores. The second row applies a grid tolerance to the event label but retains stored floating-point scores; the third uses a literal stored-value comparison for both. The differences are numerical sensitivities, not changes in the information set, model, or economic threshold.

B2. Pre-Survey Whole-Path Randomization

Table IA.6 reports the cutoff-feasible randomization, which permutes complete pre-survey targeted-machine forecast-change paths and never uses the recipient realized outcome. The observed ROC-AUC exceeds both reference-distribution means, with two-sided p-values of 0.034 and 0.031; the PR-AUC evidence is weaker under the support-only construction. The matched-support comparison in Table 3 does not show a reliable advantage over no change. The exercise therefore supports raw loss ranking rather than machine-specific incremental value. Both the observed statistic and every randomized path use the same path-first integer policy-tick convention. Stored floating-point scores enter only the numerical sensitivity in Section A3. Survey-path, target-FOMC, and pigeonhole intervals are the inference methods used in the main text; a second, geometry-conditioned path randomization reaches the same qualitative conclusion and is omitted.

Table IA.6: Pre-Survey Whole-Path Randomization

Null / statistic	Actual	Null mean	Actual-minus-null	Favorable p	Two-sided p
Cycle/support path permutation / ROC-AUC	0.712	0.654	0.058	0.019	0.034
Cycle/support path permutation / PR-AUC	0.478	0.394	0.084	0.008	0.008
Support-only path permutation / ROC-AUC	0.712	0.631	0.081	0.015	0.031
Support-only path permutation / PR-AUC	0.478	0.410	0.069	0.046	0.086

Notes: The actual operational 50-basis-point statistic uses 388 cells. Each of 9,999 draws permutes complete targeted-machine forecast-change paths; it never uses the recipient realized outcome. Human forecasts and outcomes remain fixed. Both the observed and reassigned paths are converted to integer policy ticks before disagreement is formed. The cycle/support null conditions on policy regime and identical horizon support; the support-only sensitivity removes the cycle conditioning. These tests reject random date alignment but do not establish superiority over no change, which is evaluated in Main Table 3.

B3. Incremental Information beyond Forecast Difficulty

Table IA.7 reports the development-panel ranking comparisons for Market Participant, Primary Dealer, cross-unit, and pooled professional losses.

Table IA.7: Raw Loss Ranking Relative to No Change

Evaluated loss	Ranking score	Cells	ΔAUC	Multiway 95% CI	Targeted AUC	PR-AUC	No-change AUC
Market Participant	Own-unit targeted disagreement	388	0.012	[-0.103, 0.124]	0.712	0.478	0.700
Market Participant	Cross-unit targeted disagreement	388	0.009	[-0.129, 0.127]	0.705	0.486	0.695
Primary Dealer	Cross-unit targeted disagreement	388	0.041	[-0.073, 0.154]	0.733	0.504	0.692
Pooled professional	Cross-unit targeted disagreement	776	0.025	[-0.085, 0.138]	0.718	0.494	0.693

Notes: The common development panel contains 388 cells across 60 survey paths. The pooled row stacks Market Participant and Primary Dealer losses on the same cells and does not treat 776 rows as independent. ΔAUC is professional–model disagreement minus the geometry-matched no-change score. Multiway intervals preserve survey-path and target-FOMC dependence. The own-unit Market Participant AUC has a survey-origin-by-target-FOMC interval of $[0.583, 0.828]$ on the same 388-cell support. The primary score is targeted disagreement formed after each grid-native U.S. path is converted to integer policy ticks, so identical 25 bp policy-grid gaps form exact ties. The 50-basis-point event is an operational loss threshold rather than an estimated classification cutoff. Because the scores are absolute distances, they do not encode a revision direction.

Table IA.8 reports the expanding-window risk models, discrimination and proper scores, risk-coverage loss, retained error, miss capture, and survey-path comparisons used as supplementary probability diagnostics. The survey-cluster intervals are reported in the same table family rather than repeated in a separate table.

Table IA.8: Incremental Information Beyond Forecast Difficulty

Panel A. Expanding-window risk models

Risk model	Cells	ROC-AUC	PR-AUC	Brier	RC-MAE area	Retained MAE
Human + no-change + state	250	0.600	0.413	0.2288	31.89	34.04
+ targeted disagreement	250	0.610	0.421	0.2294	31.65	33.78
+ residual targeted disagreement	250	0.579	0.372	0.2625	33.93	33.38

Panel B. Survey-cluster sensitivity for targeted disagreement beyond observable forecast difficulty

Metric	Favorable difference	CI lower	CI upper	Two-sided p
ROC-AUC	+0.010	+0.004	+0.016	0.002
PR-AUC	+0.008	+0.000	+0.020	0.040
Brier score	-0.001	-0.005	+0.004	0.792
RC-MAE area (50–100%)	+0.235	-0.050	+0.731	0.134
Retained MAE at 75%	+0.266	-0.258	+2.821	0.670
Miss capture at 75%	+0.000	-0.016	+0.058	1.000

Panel C. Residual gap after predictable-difficulty removal

Metric	Favorable difference	CI lower	CI upper	Two-sided p
ROC-AUC	-0.021	-0.063	+0.022	0.351
Brier score	-0.034	-0.069	-0.006	0.013
RC-MAE area (50–100%)	-2.042	-5.053	+0.291	0.129

Panel D. Survey-path proper-score comparison for Participant large-loss probabilities

Added score	Score	Mean diff.	Wild 95% CI	Equal-ability p	Conditional p
Own-unit targeted disagreement	brier	0.0007	[-0.0033, 0.0047]	0.747	0.233
Own-unit targeted disagreement	log loss	-0.0114	[-0.0306, 0.0078]	0.276	0.191
Cross-unit targeted disagreement	brier	0.0033	[-0.0026, 0.0091]	0.323	0.374
Cross-unit targeted disagreement	log loss	0.0500	[0.0043, 0.0961]	0.026	0.224

Notes: The recursive common support contains 250 cells across 40 survey paths and 40 target FOMCs. Every prediction is expanding-window; training labels must be realized before the test survey deadline, and imputation, indicators, and scaling are fitted within each training fold. The baseline includes Dealer–Participant disagreement, horizon, the existing policy-cycle indicators, lead time, professional-path magnitude, no-change distance, and lagged same-horizon difficulty. The policy-cycle indicators are retrospective calendar-state controls, not a deployable real-time state classifier. Positive favorable differences always favor the augmented model; signs are reversed for losses. Intervals use 4,999 survey-cluster bootstrap draws. These survey-cluster intervals are a dependence sensitivity. Main Table 3 reports the direct matched-support score comparison under survey-by-target-FOMC pigeonhole inference, and Appendix Table IA.10 reports the parsimonious recursive comparison that adds within-participant IQR and then model disagreement. RC-MAE area integrates retained professional MAE over 50–100% forecast coverage. The 50-basis-point event and 75% coverage are reported as a descriptive operating point. Panel D reports augmented-minus-baseline proper-score differentials, so negative values favor augmentation; none establishes a favorable survey-path proper-score gain. Neither augmented model improves all three reported AUC, Brier, and RC-MAE measures; these historical-sample results do not establish causal or uniquely machine-specific information.

The same recursively estimated probabilities can also be ranked at the reported flagged share. Table IA.9 compares the state-and-persistence probability specification, its targeted-disagreement augmentation, targeted disagreement, and no-change disagreement on identical support. This exploratory comparison was designed after the 250-cell allocation comparison and uses the same historical sample; it does not supply independent evidence from another setting.

Table IA.9: Recursive State-and-Persistence Specifications at a Common Flagged Share

Panel A. Common-share point estimates

Rule or score	AUC	PR-AUC	Capture (%)	MAE among unflagged	Loss, $\rho = 2$	RC-MAE area
State and persistence	0.600	0.413	33.8	34.04	0.592	31.89
+ targeted disagreement	0.610	0.421	33.8	33.78	0.592	31.65
Targeted disagreement	0.709	0.516	47.8	23.52	0.520	24.66
No-change disagreement	0.777	0.626	52.8	21.90	0.493	22.59

Panel B. Paired differences with multiway 95% intervals

Candidate minus comparator	ΔAUC	ΔCapture (pp)	Δ MAE among unflagged (bp)
State and persistence minus targeted disagreement	-0.109 [-0.362, 0.149]	-13.9 [-41.1, 18.4]	10.52 [-3.07, 24.27]
State and persistence minus no change	-0.176 [-0.413, 0.052]	-19.0 [-44.8, 4.6]	12.15 [0.29, 25.54]
+ targeted disagreement minus state and persistence	0.010 [-0.001, 0.023]	0.0 [-2.5, 7.4]	-0.27 [-3.81, 0.60]

Notes: The support contains 250 cells, 40 survey paths, 40 target FOMCs, and 65 professional losses of at least 50 basis points. Every rule flags 62 cells (24.8%). The state-and-persistence specification includes professional disagreement, horizon, policy-cycle indicators, lead time, professional-path magnitude, no-change distance, and prior same-horizon difficulty; the next row adds targeted disagreement. Existing probabilities are evaluated without refitting. The allocation-loss column uses the illustrative calibration $\rho = c_L/c_A = 2$, where c_L is the cost of leaving a large-loss cell unreviewed and c_A is the cost of assigning one cell to review. Panel B reports candidate minus comparator; its intervals use 4,999 paired survey-origin-by-target-FOMC pigeonhole-bootstrap draws and recompute each allocation and cutoff tie. Table IA.8 reports the survey-cluster sensitivity for the same incremental AUC comparison. This same-sample comparison is exploratory. The AUC and capture intervals in Panel B include zero. The MAE-among-unflagged interval for the state-and-persistence rule relative to no change is $[0.29, 25.54]$ and excludes zero; the other MAE-among-unflagged intervals include zero. These comparisons do not establish a causal effect of review, and equivalence is not tested. The displayed loss uses one common ρ for illustration; actual unreviewed-loss costs may differ by horizon and policy state.

B4. Cross-Respondent Modal-Path Dispersion

The Market Participant IQR is the difference between the 75th and 25th percentiles of respondents' modal policy-rate paths at each survey–horizon cell. It is a measure of cross-respondent dispersion, not individual subjective uncertainty. The recursive exercise uses an ℓ_2-penalized logit with fold-specific median imputation, missing-value indicators, and standardization. The inverse penalty is fixed at $C = 1$, the LBFGS solver is allowed up to 5,000 iterations, and the penalty is not retuned across test origins. The comparison uses the same 250 cells, 65 large errors, and 24.8% flagged share throughout.

Table IA.10: Cross-Respondent Modal-Path Dispersion and Recursive Large-Error Detection

Panel A. Out-of-sample performance on the 250-cell support

Recursive score	ROC-AUC	PR-AUC	Brier loss	Log loss	Large-loss capture (%)	MAE among unflagged (bp)
Prior prevalence	0.266	0.182	0.2061	0.6081	12.3	39.69
Horizon + no change	0.824	0.592	0.1532	0.4541	53.8	22.07
Horizon + no change + IQR	0.583	0.423	0.2193	0.8367	43.1	30.59
Horizon + no change + IQR + model disagreement	0.566	0.395	0.2196	0.8694	41.5	31.52

Panel B. Increment from model disagreement relative to the IQR specification

Metric	Difference	95% interval	p-value	Draws
ROC-AUC	-0.0178	[−0.0502, 0.0123]	0.240	4,999
PR-AUC	-0.0285	[−0.1143, 0.0342]	0.389	4,999
Brier loss	0.0003	[−0.0142, 0.0145]	0.977	4,999
Log loss	0.0327	[−0.0175, 0.0942]	0.228	4,999
Large-loss capture (pp)	-1.54	[−11.11, 8.78]	1.000	4,999
MAE among unflagged (bp)	0.93	[−3.06, 4.86]	0.944	4,999

Panel C. Measurement of modal-path dispersion on the 388-cell panel

Source format	Cells	Zero-IQR cells (share, %)	Median IQR (bp)	Mean IQR (bp)	Median respondents
Legacy PDF	247	153 (61.9)	0.0	11.26	28
Standardized XLSX	141	60 (42.6)	25.0	17.69	27

Notes: IQR is the 75th minus the 25th percentile of Market Participant respondents' modal policy-rate forecasts. It measures cross-respondent modal-path dispersion, not individual subjective uncertainty. Each recursive logit uses only training observations whose policy outcomes were realized before the test survey. Predictors are standardized within each training fold, and the L_2 penalty is fixed at $C = 1$ without outcome-based tuning. All Panel A models are evaluated on the same 250 cells and 65 large losses; 62 cells receive fractional flag assignments. Panel B reports the IQR-plus-model-disagreement specification minus the IQR specification. Higher AUC, PR-AUC, and capture are favorable, whereas lower Brier loss, log loss, and MAE among unflagged forecasts are favorable. Intervals and two-sided p-values use 4,999 paired survey-origin-by-target-FOMC pigeonhole-bootstrap draws. The recursive comparisons are exploratory.

B5. Loss Gradients and Forecast Heterogeneity

A within-survey swap that ranks Market Participant losses with Dealer–model disagreement, and conversely, preserves positive raw ranking and is omitted as a table. It removes own-path arithmetic but not common-date dependence. Table IA.11 reports the unconditional increase in professional forecast losses across disagreement quartiles.

Table IA.11: Professional Forecast Losses across Disagreement Quartiles

Quartile	Expected cells	Mean targeted disagreement	Professional MAE	Median AE	Large-loss rate
Q1	97	0.00	17.52	0.00	11.6%
Q2	97	12.11	23.19	0.00	17.6%
Q3	97	25	29.22	0.00	24.0%
Q4	97	63.40	61.52	25	47.8%

Notes: The table uses the common 388-cell, 60-survey development panel. Grid-native professional and targeted-model paths are converted path-first to integer 25-basis-point policy ticks before disagreement and loss are formed. Quartiles each receive one-fourth of the sample. When a tied score crosses a quartile boundary, every observation in the tie receives the same fractional membership in the adjacent quartiles. Expected cells are therefore membership weights rather than an arbitrary division of tied observations. Targeted disagreement, mean absolute error (MAE), and median absolute error (AE) are in basis points. A large loss is a professional absolute error of at least two policy ticks. The table describes an unconditional loss gradient and does not identify a causal effect of disagreement.

Table IA.12 collects the complete horizon and policy-cycle evidence. The raw disagreement–loss slope varies across policy states, while no-change-relative AUC differences change sign and remain imprecise. All groups are reported; none is selected from the realized estimates.

Table IA.12: Forecast-Horizon and Policy-Cycle Heterogeneity

Panel A. Forecast-horizon groups

Horizon group	Cells	Misses (rate)	Targeted AUC	No-change AUC	Δ AUC	95% CI
H1	60	1 (1.7%)	0.924	0.924	0.000	Not reported
H2–H3	117	11 (9.4%)	0.689	0.725	-0.036	[-0.236, 0.176]
H4–H7	211	86 (40.8%)	0.696	0.640	+0.055	[-0.041, 0.152]

Panel B. Existing policy-cycle states

Policy-cycle state	Cells	Misses (rate)	Targeted AUC	No-change AUC	Δ AUC	95% CI
Pre-pandemic	63	32 (50.8%)	0.525	0.464	+0.061	[-0.063, 0.197]
Pandemic / ELB	110	16 (14.5%)	0.704	0.728	-0.024	[-0.226, 0.259]
Rapid tightening	81	33 (40.7%)	0.814	0.898	-0.084	[-0.239, 0.015]
Plateau / easing	134	17 (12.7%)	0.721	0.759	-0.038	[-0.188, 0.084]

Panel C. Formal interaction tests

Interaction family	Reference group	Joint cluster-robust p
Targeted disagreement × horizon group	H2–H3	0.593
Targeted disagreement × policy-cycle state	Pre-pandemic	< 0.001

Notes: The common out-of-sample development panel contains 388 horizon cells across 60 surveys. The outcome is a professional absolute error of at least 50 basis points; the threshold represents two standard 25-basis-point policy moves. Targeted disagreement and no-change disagreement are formed path-first on the integer policy-tick grid. Horizon groups (H1, H2–H3, H4–H7) and the four policy-cycle labels were fixed before this analysis and no null group is suppressed. AUC intervals are percentile intervals from 4,999 survey-cluster bootstrap draws that resample complete policy paths. H1 contains one large miss, below the five-event support rule, so its descriptive AUC is shown but its interval is not reported. Panel C reports joint tests from linear probability models with survey-cluster robust covariance using the same integer-tick targeted score; the horizon model controls for policy-cycle state, and the policy-cycle model controls for exact horizon. These development-sample estimates are descriptive rather than causal.

Table IA.13: Sample Composition and Direct Score Ordering across Supports

Panel A. Policy-cycle composition of development and recursive support

Sample	Cells	Large losses	Pre-pandemic	Pandemic / ELB	Rapid tightening	Plateau / easing
Development support	388	98	16.2	28.4	20.9	34.5
Recursive support	250	65	0.0	14.0	32.4	53.6

Panel B. Direct targeted-disagreement and no-change ranking on exact support

Support or policy state	Cells	Large losses	Targeted AUC	No-change AUC	Difference	Survey-cluster 95% interval
Development support	388	98	0.712	0.700	0.012	[-0.063, 0.090]
Recursive support	250	65	0.709	0.777	-0.067	[-0.148, 0.010]
Recursive: Pandemic / ELB	35	15	0.627	0.750	-0.123	[-0.365, 0.114]
Recursive: Rapid tightening	81	33	0.814	0.898	-0.084	[-0.229, 0.017]
Recursive: Plateau / easing	134	17	0.721	0.759	-0.038	[-0.179, 0.087]

Panel C. Direct score ordering in disjoint survey-origin blocks

Block	Cells	Survey origins	Target FOMCs	Large losses	Targeted AUC	No-change AUC	Difference
Earlier complement	138	20	25	33	0.725	0.622	0.103
Later recursive block	250	40	40	65	0.709	0.777	-0.067

Panel D. Inference for the later-minus-earlier AUC difference

Inference	Difference	95% interval	p-value	Valid draws
Survey × target FOMC (primary)	-0.171	[-0.415, 0.043]	0.110	4,998
Survey only	-0.171	[-0.337, -0.024]	0.022	4,999
Target FOMC only	-0.171	[-0.289, -0.061]	0.004	4,999

Notes: Panel A reports policy-state cell shares (percent). The development panel contains 388 cells across 60 survey paths; the recursive panel contains 250 cells across 40 paths. Panel B evaluates both scores directly on each stated support. Its intervals are percentile intervals from 4,999 survey-path cluster bootstrap draws and are reported only when a group contains at least five large losses. Pre-pandemic does not appear on recursive support. Panel C partitions the 388 cells into the exact 138-cell earlier complement and the 250-cell later recursive block. The cell sets and survey origins are disjoint, but five target FOMCs appear in both blocks. Panel D reports the later-minus-earlier difference in targeted-minus-no-change AUC. The primary interval and two-sided p-value use a stratified survey-by-target-FOMC pigeonhole bootstrap that resamples survey paths separately within each block and uses common target-FOMC multiplicities across blocks. The one-way rows are sensitivity checks. Of 4,999 requested draws, 4,998 are valid in the primary row. This comparison was developed after the support-specific point estimates were observed and is exploratory. Its primary interval includes zero, so it does not establish a population change in relative ranking or identify policy state as its cause. Pooled AUC is a pairwise ranking probability and is not an additive average of the policy-state or survey-origin-block rows.

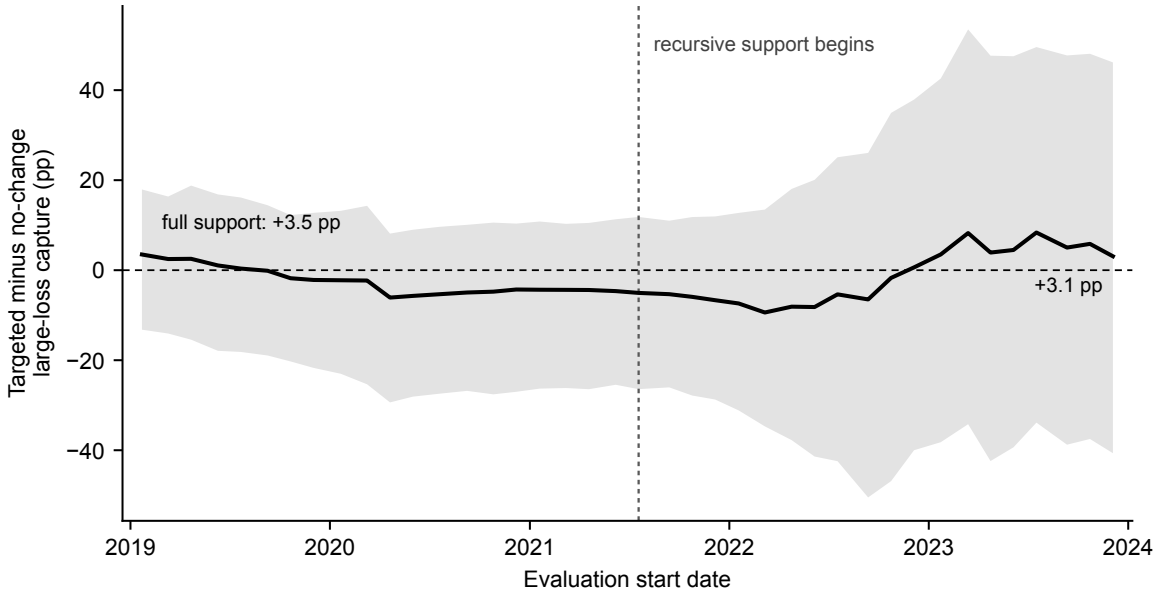


Figure IA.1: Fixed-Share Detection Difference as the Evaluation Start Date Advances

Notes: Each point recomputes the targeted-minus-no-change large-error capture difference at a requested 25% flagged share on cells whose survey origin falls on or after the stated start date, using the path-first integer policy-tick scores without refitting, the 50-basis-point event, and fractional assignment at cutoff ties. The shaded band is the paired survey-origin-by-target-FOMC pigeonhole-bootstrap 95% interval from 4,999 draws. Windows with fewer than 15 large errors are not displayed. The leftmost point is the full 388-cell support; the vertical line marks the start of the nested 250-cell support. Only the evaluation window moves. This support sensitivity was examined after the pooled comparison, and calendar time and policy state move together; the figure does not identify a policy-state mechanism.

B6. Fixed-Share Robustness and Alternative Signals

The expanding-window residual of D^M on D^N has recursive AUC below one half because the projection subtracts a positive multiple of the stronger score; it is not used as a warning rule. Unique selection by the four-model median versus no change is small and imprecise and is omitted as a separate table. Table IA.14 varies one element of the allocation comparison at a time: the size of the loss event, the professional panel whose losses are evaluated, and the policy cycles present in the sample. Defining the event as one or three policy ticks instead of two leaves the sign of the capture difference unchanged within each support, so the threshold is not what separates the two supports. Evaluating the Primary Dealer path in place of the Market Participant path is a separate aggregate forecast and loss object, not an interchangeable version of the Market Participant path. It yields a targeted-minus-no-change difference of -0.1 percentage point on 250 cells and $+11.1$ points on 388 cells, and the orthogonal residual remains below the no-change rule under both panels. The recursive gap stays negative after each cycle is dropped. It is most negative when the pandemic/ELB state is omitted (-9.4 percentage points) and closest to zero when plateau/easing is omitted (-1.2). Every interval for the primary model-disagreement comparisons includes zero. The residual-rule intervals exclude zero after dropping the pandemic/ELB and plateau/easing periods, in both cases in the direction favoring no change. Three of the four development-sample subsamples retain a positive point estimate, while dropping the pre-pandemic block turns the development difference slightly negative. No perturbation yields a difference that favors model disagreement with an interval excluding zero, and the residual remains below no change in every reported subsample. These perturbations were chosen after the reported estimates were known. They bound the sensitivity of the comparison to the threshold, the evaluated panel, and any single policy cycle, and they are exploratory rather than independent confirmation.

Table IA.14: Fixed-Share Detection across Thresholds, Forecast Units, and Policy Cycles

Sample and event	Rule	Cells	Losses	Rule capture (%)	No change capture (%)	Difference (pp)	95% CI	p
Panel A. Large-loss threshold								
250 cells, 25bp event	D^M	250	133	33.3	36.1	-2.8	[-13.0, 6.0]	0.466
250 cells, 25bp event	Residual	250	133	25.6	36.1	-10.5	[-27.2, 3.2]	0.129
250 cells, 50bp event	D^M	250	65	47.8	52.8	-5.0	[-26.2, 11.8]	0.470
250 cells, 50bp event	Residual	250	65	24.6	52.8	-28.2	[-58.7, 1.1]	0.058
250 cells, 75bp event	D^M	250	35	52.9	62.9	-9.9	[-38.2, 5.9]	0.332
250 cells, 75bp event	Residual	250	35	22.9	62.9	-40.0	[-87.5, 10.7]	0.118
388 cells, 25bp event	D^M	388	177	38.9	37.3	+1.6	[-9.6, 10.9]	0.874
388 cells, 50bp event	D^M	388	98	47.3	43.9	+3.5	[-13.6, 17.4]	0.736
388 cells, 75bp event	D^M	388	63	48.2	41.0	+7.2	[-12.2, 26.2]	0.446
Panel B. Professional forecast panel								
250 cells, Dealer panel	D^M	250	60	55.8	55.9	-0.1	[-19.5, 18.5]	0.904
250 cells, Dealer panel	Residual	250	60	25.0	55.9	-30.9	[-63.0, 1.3]	0.061
388 cells, Dealer panel	D^M	388	96	54.7	43.6	+11.1	[-6.5, 27.4]	0.260
Panel C. Removal of a single policy cycle								
250 cells, drop pandemic ELB	D^M	215	50	55.0	64.4	-9.4	[-34.7, 14.1]	0.439
250 cells, drop pandemic ELB	Residual	215	50	24.0	64.4	-40.4	[-70.6, -4.3]	0.035
250 cells, drop plateau/easing	D^M	116	48	47.9	49.1	-1.2	[-24.4, 9.9]	0.796
250 cells, drop plateau/easing	Residual	116	48	18.8	49.1	-30.3	[-61.3, -2.7]	0.034
250 cells, drop rapid tightening	D^M	169	32	36.7	39.0	-2.3	[-28.4, 23.2]	0.756
250 cells, drop rapid tightening	Residual	169	32	31.2	39.0	-7.8	[-50.7, 28.7]	0.600
388 cells, drop pandemic ELB	D^M	278	82	44.3	41.8	+2.5	[-17.7, 17.2]	0.934
388 cells, drop plateau/easing	D^M	254	81	48.3	44.0	+4.4	[-13.9, 18.4]	0.658
388 cells, drop pre-pandemic	D^M	325	66	54.2	56.5	-2.3	[-25.6, 15.5]	0.629
388 cells, drop rapid tightening	D^M	307	65	41.4	37.1	+4.3	[-14.3, 24.4]	0.613

Notes: Each row applies the same flagged share on the stated sample. D^M is the absolute professional–model gap; Residual is its expanding-window remainder after projection on no-change distance, estimated from earlier survey origins. No change is the comparison rule at the identical flagged share. Events use integer 25-basis-point policy ticks. Panel B evaluates the Dealer path and its own losses and is not independent evidence. Panel C recomputes the flagged count after removing the stated cycle. Intervals and p-values use 4,999 survey-by-target-FOMC pigeonhole-bootstrap draws that recompute flagged counts and cutoff ties. These post-result perturbations are exploratory rather than independent confirmation.

Table IA.15 imposes a common flagged count within each survey path rather than ranking cells jointly across historical origins. Six late origins are excluded because their apparent candidate sets are truncated only by outcomes not yet realized at the evaluation endpoint; genuinely shorter historical paths remain. The resulting sample contains 54 survey paths and 367 cells. Distant-horizon selection has the highest capture point estimate for both the top-one and top-two comparisons. With two cells selected, model disagreement leaves 6.62 more basis points of unreviewed error than the horizon rule, with a 95% interval of $[1.34, 13.53]$. Because the exercise was developed after the pooled result, it is exploratory and should not be interpreted as independent confirmation.

Table IA.15: Within-Survey Fixed-Share Detection on Complete Eligible Choice Sets

Panel A. Fixed-share detection outcomes

Horizons flagged	Rule	Flagged mass	Flagged share (%)	Large-loss capture (%)	MAE among unflagged (bp)
Top one	Model disagreement	54	14.7	24.7	31.15
Top one	No change	54	14.7	23.3	31.32
Top one	Distant horizon	54	14.7	30.6	27.96
Top two	Model disagreement	108	29.4	44.1	27.95
Top two	No change	108	29.4	44.4	27.57
Top two	Distant horizon	108	29.4	58.2	21.33

Panel B. Paired differences relative to alternative rules

Horizons assigned	Comparator	Outcome	Difference	95% interval	p-value
Top one	No change	Capture (pp)	1.38	[−7.09, 11.46]	0.766
Top one	No change	MAE among unflagged (bp)	-0.17	[−3.12, 3.04]	0.884
Top one	Distant horizon	Capture (pp)	-5.95	[−19.37, 5.87]	0.312
Top one	Distant horizon	MAE among unflagged (bp)	3.19	[−0.45, 7.53]	0.083
Top two	No change	Capture (pp)	-0.30	[−12.12, 11.48]	0.956
Top two	No change	MAE among unflagged (bp)	0.38	[−4.28, 5.04]	0.844
Top two	Distant horizon	Capture (pp)	-14.10	[−30.72, 1.19]	0.066
Top two	Distant horizon	MAE among unflagged (bp)	6.62	[1.34, 13.53]	0.013

Notes: The sample contains 367 horizon cells, 54 survey paths, and 98 large losses. It excludes six late survey origins for which some contemporaneously eligible H1–H7 forecasts lack realized outcomes at the evaluation endpoint; genuinely shorter legacy paths are retained. Each rule assigns the top one or top two eligible cells within a survey. Cutoff ties receive uniform fractional assignment. Panel B reports model disagreement minus the stated comparator. Intervals and two-sided p-values use 4,999 paired survey-origin-by-target-FOMC pigeonhole-bootstrap draws. The exercise is an offline historical comparison and does not observe a warning, review action, or forecast revision.

Table IA.16 reports the no-move identity and the moved-cell allocation. Panel A documents that on cells where the realized midpoint equals the rate in force at the survey origin the no-change distance equals the realized professional loss, so within no-move cells its AUC is 1 by construction; restricting to the cells where the target moved lowers the no-change AUC to 0.523 on the development panel and 0.691 on the recursive panel, against 0.636 and 0.686 for targeted disagreement. Panel B reruns the 75%-coverage allocation inside the moved cells: the capture orderings match the full-support orderings and both paired intervals include zero. The moved-cell comparisons were specified after the main allocation results were known and are exploratory rather than independent confirmation.

Table IA.16: The No-Move Identity and Moved-Cell Fixed-Share Detection

Panel A. The no-move identity and discrimination by subsample

Support	No-move cells	Large losses	No-move losses	Targeted AUC (full)	Targeted AUC (moved)	No-change AUC (full)	No-change AUC (moved)
Development (388 cells)	157 (40.5%)	98	4	0.712	0.636	0.700	0.523
Recursive (250 cells)	76 (30.4%)	65	4	0.709	0.686	0.777	0.691

Panel B. The 75%-coverage comparison on cells where the target moved

Support	Cells	Large losses	Targeted capture (%)	No-change capture (%)	Δ capture (pp)	Multiway 95% CI	Two-sided p
Development, moved cells	231	94	37.1	34.6	+2.5	[-12.6, 18.7]	0.723
Recursive, moved cells	174	61	43.6	50.8	-7.2	[-25.7, 12.0]	0.560

Notes: A no-move cell is one whose realized target midpoint equals the rate in force at the survey origin; on those cells the no-change distance equals the realized professional absolute error, so the within-subsample no-change AUC is 1 by construction and is not displayed. Panel B reruns the tie-aware fixed-share comparison at requested 75% coverage within the moved cells (58 and 44 expected flags); MAE among unflagged forecasts is 43.2 against 47.3 basis points on the development panel ($\Delta = -4.10$, CI $[-14.25, 4.47]$) and 33.6 against 31.9 on the recursive panel ($\Delta = +1.66$, CI $[-6.88, 8.32]$), and the moved-cell AUC differences are $+0.113$ $[-0.042, 0.268]$ and -0.005 $[-0.173, 0.150]$. Grid-native U.S. paths are converted to integer policy ticks before scores are formed, and intervals and p-values use 4,999 survey-by-target-FOMC pigeonhole-bootstrap draws that recompute allocations and cutoff ties within each draw. These comparisons were specified after the main allocation results were known and are exploratory rather than independent confirmation. Distant-horizon allocation is reported in Appendix Table IA.15.

Table IA.17 holds the Ridge estimator, tuning schedule, expanding folds, target grid, recursive support, flagged share, and tie rule fixed while changing only the public-information block. The 281 macro-and-market features in Table IA.2 map uniquely to 168 release-dated macro features and 113 strictly lagged market features; no feature is unclassified or assigned to both blocks. Adding the market block substantially lowers challenger MAE but produces less favorable allocation point estimates. The targeted Federal Reserve text specification has modestly higher ranking and allocation point estimates than the macro-and-market parent, but the paired allocation intervals include zero and the recursive proper scores do not improve. All parent and no-change AUC, PR-AUC, capture, and retained-MAE contrasts use the same 4,999 survey-by-target-FOMC draws.

Panel D reports a separate source diagnostic. The human-free model movement $D^{MN} = |M - N|$ has ROC-AUC 0.627 and assigns 44.3% of large losses to the flagged set. Relative to no change, its AUC difference is -0.150 ($[-0.275, -0.036]$), its PR-AUC difference is -0.253 ($[-0.372, -0.076]$), its capture difference is -8.6 percentage points ($[-30.1, 5.7]$), and its retained-MAE difference is 5.89 basis points ($[-0.58, 18.94]$). Relative to distant-horizon selection, the capture and retained-MAE differences are -7.8 points ($[-37.7, 14.5]$) and 7.03 basis points ($[-2.13, 21.85]$). The allocation intervals are wide. The exercise therefore supplies no affirmative evidence of model-specific allocation content, but it does not establish population inferiority or equivalence. Because the three absolute distances are not additively separable, these comparisons are not a structural decomposition.

Table IA.17: Public-Information Blocks and Fixed-Share Detection

Panel A. Forecast production, loss ranking, and fixed-share detection

Information set	Challenger MAE, 388 cells	AUC, 388 cells	AUC, 250 cells	PR-AUC, 250 cells	Capture (%)	MAE among unflagged
Release-dated macro data	65.34	0.687	0.739	0.562	51.1	20.87
Macro data + lagged markets	49.42	0.692	0.676	0.482	44.6	23.80
Macro data + lagged markets + targeted Fed text	49.61	0.712	0.709	0.516	47.8	23.52
No change	62.11	0.700	0.777	0.626	52.8	21.90

Panel B. Recursive probability performance

Risk model	ROC-AUC	PR-AUC	Brier score	Log loss
State and no-change baseline	0.600	0.413	0.2288	0.7873
Baseline + macro disagreement	0.598	0.423	0.2321	0.7921
Baseline + macro-and-market disagreement	0.611	0.425	0.2289	0.7730
Baseline + targeted disagreement	0.610	0.421	0.2294	0.7747

Panel C. Paired fixed-share contrasts

Comparison	Capture difference		Unflagged-MAE difference		Empirical tail probability	
	pp	95% interval	bp	95% interval	Capture	MAE
Lagged markets minus macro data	-6.4	[-24.5, 11.4]	2.93	[-1.66, 7.74]	0.477	0.211
Targeted Fed text minus macro and markets	3.2	[-2.6, 10.8]	-0.28	[-2.67, 1.38]	0.350	0.668
Macro data minus no change	-1.8	[-23.1, 12.9]	-1.03	[-5.21, 3.79]	0.729	0.829
Macro data and markets minus no change	-8.2	[-31.3, 11.0]	1.91	[-2.54, 8.02]	0.352	0.339
Targeted disagreement minus no change	-5.0	[-26.2, 11.8]	1.63	[-2.85, 6.99]	0.469	0.389

Panel D. Source-specific diagnostic scores on the 250-cell support

Diagnostic score	ROC-AUC	PR-AUC	Capture (%)	MAE among unflagged
$D^M = \lvert H - M \rvert$	0.709	0.516	47.8	23.52
$D^N = \lvert H - N \rvert$	0.777	0.626	52.8	21.90
$D^{MN} = \lvert M - N \rvert$	0.627	0.373	44.3	27.78
Distant-horizon selection	0.810	0.515	52.1	20.76

Notes: Panel A evaluates three nested Ridge challengers. The macro specification contains four context variables and 168 release-dated macro features (172 inputs). The second adds 113 market features observed before the survey cutoff (285 inputs). The third adds the 14 targeted Federal Reserve text measures used by the reference challenger (299 inputs). Challenger and unflagged MAE are in basis points; professional MAE is 32.86 on the 388-cell support. Panels A, C, and D use the same 250 cells, 65 large losses, and 62-cell flagged set. Panel B adds each disagreement score to the same recursive state-and-persistence risk specification; its policy-cycle indicators are retrospective calendar controls. Cutoff ties receive uniform fractional weight. Panel C differences are the richer information set minus its comparator, so higher capture and lower unflagged MAE favor the richer set. A table-specific bank of 4,999 paired survey-by-target-FOMC draws yields intervals and two-sided tail probabilities, conditions on stored paths, and recomputes cutoff ties; separate banks explain third-decimal p-value differences. Ridge re-estimation uncertainty is omitted. The information-block and human-free comparisons were designed after the main fixed-share results and are exploratory. They describe source dependence and do not identify causal input effects, post-flag responses, or model-specific information.

Table IA.18 compares the four-model median and its constituent models with no change at the same 24.8% flagged share. The median is the family summary; constituent-model rows are descriptive and are not used to select a new primary score. The median's capture point estimate is close to no change, while the paired interval permits economically meaningful advantages for either rule. No change has the highest AUC among the displayed rules. The two largest member capture point estimates, broad-text Ridge and the elastic net, differ by less than one percentage point. These comparisons use the same historical sample and are exploratory rather than independent evidence from a new setting. They do not alter the reference targeted-text specification.

Table IA.18: Public-Information Model Family and the No-Change Benchmark

Panel A. Fixed-share detection at a 24.8% flagged share

Signal	Capture (%)	Unreviewed MAE (bp)	ROC-AUC
Macro Ridge disagreement	44.6	23.8	0.676
Targeted disagreement	47.8	23.5	0.709
Broad-text Ridge disagreement	53.1	21.1	0.734
Elastic Net disagreement	53.7	21.0	0.728
Four-model median disagreement	53.4	21.4	0.734
No-change disagreement	52.8	21.9	0.777
Dealer–Participant disagreement	32.7	32.3	–

Panel B. Paired contrasts

Comparison	Difference	Multiway 95% interval	Two-sided p
Median minus no-change capture (pp)	0.6	[-20.3, 17.6]	0.979
Median minus no-change unreviewed MAE (bp)	-0.49	[-6.90, 4.47]	0.876
Targeted minus no-change capture (pp)	-5.0	[-26.2, 11.8]	0.469

Notes: The sample contains 250 cells across 40 survey paths and 40 target FOMCs, with 65 professional losses of at least 50 basis points. Requested 75% forecast coverage assigns 62 cells, or 24.8%, to the flagged set under a tie-aware allocation at the common capacity. Grid-native professional, model, and no-change paths are converted to integer policy ticks before scores are formed, matching Main Table 3. Individual model gaps therefore lie on integer ticks, while the median of four tick-valued paths may lie on half-ticks. Statistics in Panel A are capture rates, MAE among unflagged forecasts in basis points, and ROC-AUC. Panel B uses a family-specific bank of 4,999 paired survey-by-target-FOMC pigeonhole-bootstrap draws and recomputes both allocations within each draw; third-decimal p-value differences across repeated contrasts reflect separate banks. The four-model median is the family summary; constituent-model rows are descriptive. The Dealer–Participant row provides the cross-professional comparison; its ROC-AUC is not reported in this family panel. "Macro Ridge" denotes the macroeconomic-and-lagged-market benchmark without the targeted Federal Reserve text block; the macro-only ablation is reported separately in Table IA.17. These analyses use the same historical sample and are exploratory rather than independent evidence from a new setting. Post-flag behavior and welfare are not observed.

B7. Ranking, Fixed-Share Detection, and Post-Diagnostic Response

Table IA.19: Empirical Objects in Forecast Evaluation

Object	Empirical question	Reported measure	Interpretation
Forecast production	How closely does each reported path match the subsequently realized policy path?	MAE, RMSE, bias, benchmark-relative out-of-sample R^2	Point-prediction quality
Loss ranking	Does a larger pre-outcome score rank professional forecasts that later incur a loss of at least 50 basis points?	Linear-probability coefficient, ROC-AUC, PR-AUC	Ranking ability, not allocation value
Probability performance	Does disagreement improve a recursive risk forecast beyond a state-and-persistence specification?	ROC-AUC, PR-AUC, Brier loss, log loss	Incremental probability information
Common-share detection	Which score flags more large errors and leaves lower MAE among unflagged forecasts at the same share?	Capture, MAE among unflagged forecasts, risk–coverage path	Historical detection performance at a fixed share
Post-diagnostic response	Does a later action improve the final professional forecast?	Not observed	Not estimated; no team-performance claim

Notes: The table separates empirical objects that use related forecasts but answer different questions. Forecast-production and loss-ranking results use 388 cells across 60 survey paths. Recursive probability results and the primary matched common-capacity comparison use 250 cells across 40 paths; the fixed-score allocation sensitivity repeats the comparison on all 388 cells across 60 paths. The operational large-loss event is an absolute professional forecast error of at least 50 basis points. The common-capacity comparison assigns 62 of 250 cells, or 24.8%, using fractional treatment of cutoff ties. A response to additional scrutiny would require a post-warning revision or an intervention, neither of which is observed. The ordering of the rows is substantive: evidence that a score predicts errors does not by itself establish calibrated probability improvement, better case selection at a fixed capacity, or a beneficial response to review.

Global ranking and fixed-capacity capture. Let $Y \in \{0,1\}$ indicate a large professional forecast loss, let $\pi = \Pr(Y = 1)$, and let $A_j(q) \in [0,1]$ be the top-score assignment generated by rule j at capacity q, with uniform fractional assignment at cutoff ties. Define

$$K_j(q) = E[A_j(q) \mid Y = 1], \qquad F_j(q) = E[A_j(q) \mid Y = 0].$$

The capacity identity is

$$q = \pi K_j(q) + (1 - \pi)F_j(q).$$

Proposition 1 (AUC does not order fixed-capacity capture). For fixed prevalence and capacity, there exist scores S_1 and S_2 such that

$$\mathrm{AUC}(S_1) > \mathrm{AUC}(S_2) \quad \text{but} \quad K_1(q) < K_2(q).$$

To see this, consider three positive and five negative cases and review the top two. The ordering (P, N, P, P, N, N, N, N) has AUC 13/15 because its positive cases dominate respectively five, four, and four negative cases, but top-two capture is 1/3. The ordering (P, P, N, N, N, N, N, P) has lower AUC, 10/15, but top-two capture is 2/3. Thus a global pairwise ranking advantage need not improve selection at the operating capacity. This is an application of the established distinction between global discrimination and top-of-list or resource-constrained performance (Hand, 2009; El-Yaniv and Wiener, 2010; Kar et al., 2015; Shmueli, 2019).

Lemma 1 (a sufficient condition). Let $R_j(f)$ be rule j's randomized ROC curve at false-positive rate f. If $R_1(f) \geq R_2(f)$ for every $f \in [0,1]$, then $K_1(q) \geq K_2(q)$ for every $q \in [0,1]$. At false-positive rate f, selected mass is $g_j(f) = \pi R_j(f) + (1 - \pi)f$. Pointwise ROC dominance implies $g_1(f) \geq g_2(f)$, so the generalized inverse needed to attain capacity q satisfies $f_1(q) \leq f_2(q)$. The capacity identity then gives

$$K_j(q) = \frac{q - (1 - \pi)f_j(q)}{\pi},$$

which proves the result. AUC dominance is weaker because ROC curves can cross even when one has larger area.

Review-effectiveness break-even conditions. The common-capacity differences are imprecisely estimated. Table IA.20 reports the observed comparisons, multiway intervals, and normal-reference current-sample precision effect magnitudes implied by those interval widths. The calculation summarizes precision; it does not replace the reported pigeonhole-bootstrap inference or supply an equivalence bound.

Table IA.20: Uncertainty around Common-Capacity Allocation Differences

Metric	Targeted minus no change	Multiway 95% CI	Approx. detectable difference
Large-loss capture (percentage points)	-5.0	[-26.2, 11.8]	27.1
Retained MAE (bp)	1.63	[-2.85, 6.99]	7.03
Allocation loss, $\rho = 2$ (review-cost units)	0.026	[-0.059, 0.123]	0.130

Notes: Higher capture favors targeted disagreement; lower retained MAE and allocation loss favor targeted disagreement. Here $\rho = c_L/c_A$ is the cost of an unreviewed large loss relative to the cost of assigning one forecast to review; the displayed row uses the illustrative calibration $\rho = 2$. This calculation holds ρ constant across cells; unreviewed-loss costs may instead vary by forecast horizon and policy state. The first two confidence intervals reproduce Main Table 3. For each row, the normal-reference standard error is the interval width divided by $2z_{0.975}$, and the displayed detectable-difference magnitude is $(z_{0.975} + z_{0.80})\widehat{\mathrm{SE}}$ under conventional two-sided 5%-size and 80%-power conventions. Because the underlying intervals use a survey-by-target-FOMC pigeonhole bootstrap and need not be symmetric or Gaussian, these values are approximate summaries of current-sample precision, not design-stage power calculations. They are not used to select a sample, threshold, outcome, or specification. The detectable differences are large relative to the point estimates, so the table supports neither equivalence nor a precise population ranking of the rules.

The complete cost grid is omitted because it contains no additional ranking information. With common capacity, event prevalence m, a common review-effectiveness parameter, and

relative unreviewed-loss cost ρ, the difference reduces to

$$\Delta\widetilde{\ell} = -\rho m\, \Delta k,$$

where Δk is the capture difference. Rule-specific responses or implementation costs are not observed, so this identity is an accounting translation rather than a welfare calculation.

Theorem 1 (non-identification without review response). Let $L \geq 0$ denote the modeled loss that review could potentially reduce, let $A_j(q)$ indicate assignment to review by rule j at capacity q, and let $G_j = E[LA_j(q)]$ be the historical loss mass assigned to review. The historical panel provides sample analogs for G_M and G_N, but not the rule-specific response parameters $\eta_M, \eta_N \in [0, 1]$. For relative net value

$$\Delta V = \rho\{\eta_M G_M - \eta_N G_N\} - (\phi_M - \phi_N),$$

where ρ scales unreviewed loss and ϕ_j is implementation cost, the sharp identified set is

$$\Delta V \in \left[-\rho G_N - (\phi_M - \phi_N),\, \rho G_M - (\phi_M - \phi_N)\right].$$

With equal implementation cost and positive selected loss mass for both rules, this interval contains both signs. Hence neither AUC nor historical allocation identifies the relative value of the assisted-review systems without restrictions on review response. Designs that observe disclosure, effort, revision, and final decisions estimate a different object (Agarwal et al., 2023, 2025).

Proof. The lower endpoint of the stated identified set sets $(\eta_M, \eta_N) = (0, 1)$, while the upper endpoint sets $(\eta_M, \eta_N) = (1, 0)$, and continuity yields every intermediate value. With equal implementation cost and positive selected loss mass for both rules, the interval contains both signs, which proves the result.

B8. Information Available at Public Release

The main comparison treats the survey aggregate and model score as available to a hypothetical internal decision maker after responses close but before the relevant target outcomes. An outside user sees the published aggregate path later. The New York Fed states that results and data are released about three weeks after the corresponding FOMC meeting, following publication of the meeting minutes (Federal Reserve Bank of New York, 2026). The public-release sample therefore retains only target meetings strictly after the corresponding minutes-release proxy date; one origin without an exact official minutes-date match is excluded rather than assigned an imputed date.

Table IA.21 shows that, on the resulting 322-cell support from 58 survey origins, disagreement continues to rank future large professional losses, but its AUC advantage over no change is imprecise under survey-by-target dependence. On the 204-cell recursive support, adding disagreement lowers the Brier-loss point estimate, but the crossed-dependence interval includes zero and the survey-path equal-predictive-ability test does not reject. A recursive-prevalence benchmark has lower Brier and log loss than both displayed probability models, further limiting a public-release decision-value interpretation.

Table IA.21: Forecast-Loss Information Available at Public Release

Panel A. Raw ranking among targets unresolved at public release

Score	Cells	Origins	Targets	ROC-AUC	PR-AUC	ΔAUC [95% CI]
Targeted disagreement	322	58	59	0.735	0.568	0.055 [-0.064, 0.175]
No-change disagreement	322	58	59	0.680	0.561	—

Panel B. Recursive probability forecasts

Probability forecast	Cells	Origins	Targets	ROC-AUC	Brier	Log loss
Publication-time baseline	204	38	38	0.519	0.2896	1.3726
+ targeted disagreement	204	38	38	0.601	0.2668	1.0286
Recursive prevalence sensitivity	204	38	38	—	0.2320	0.6603

Notes: The information origin is the minutes-release proxy date, defined as the original release date of the corresponding FOMC minutes. The official minutes pages supply the release dates. For the September 2021 meeting, the October 13 press release supplies the original release date because the minutes page was subsequently updated on November 26. Targets on or before the resulting date and H1 are excluded. One March 2020 origin without an exact official date match is omitted rather than imputed. Panel A compares scores on identical support; the interval uses 4,999 survey-by-target-FOMC pigeonhole-bootstrap draws. The publication-time baseline includes horizon, policy-cycle state, lead time, professional-path magnitude, Dealer disagreement, no-change distance, the already realized same-path prefix, and prior same-horizon difficulty. The policy-cycle state is a retrospective calendar control, not a real-time state classifier. All recursive training labels precede the minutes-release proxy date. Negative Brier differences favor the augmented probability forecast. Its augmented-minus-baseline Brier differential is -0.0228, with crossed-dependence 95% interval $[-0.0972, 0.0102]$ and survey-path equal-predictive-ability $p = 0.577$. The recursive-prevalence benchmark has lower Brier and log loss than both displayed probability models. The results establish neither incremental probability value nor observed model-assisted decision performance for an outside user.

Part C. Professional Forecasts across Institutions

C1. ECB Survey of Monetary Analysts

The two New York Fed professional panels share one survey setting, so their symmetry does not provide an independent institutional sample. The ECB application uses the published median deposit-facility-rate path and an expanding-window Ridge based only on the official policy-rate history available at the survey date. It asks whether the distinction between forecast accuracy and loss ranking appears in another institution without transporting U.S. coefficients.

The professional median has lower realized-point MAE than the challenger, and disagreement has positive descriptive ranking. No change ranks at least as well, the professional–challenger AUC interval includes 0.5, and ranking weakens after removing unconditional horizon composition. The human-free challenger-versus-no-change score is null, leaving no dependence-adjusted evidence that model disagreement is more informative than persistence. Table IA.22 reports the application. Horizon- and path-rank transforms lower both AUCs and leave the no-change point estimate higher; the human-free challenger–no-change score is 0.495.

Table IA.22: ECB Survey of Monetary Analysts: Policy-Rate Forecast Application

Panel A. Forecast accuracy and large-error ranking			
	Professional SMA	Policy-history Ridge	No change
MAE (basis points)	34.30	61.42	78.17
Large-loss ROC-AUC	—	0.668	0.695
Large-loss PR-AUC	—	0.478	0.477
Recursive Brier loss	—	0.1496	0.1593
Panel B. Fixed-share large-error detection			
Large losses in flagged set	—	43.8%	49.1%
MAE among unflagged forecasts	—	23.91	25.84

Notes: The common support contains 308 cells across 42 origins and 46 targets, including 64 professional losses of at least 50 basis points. The challenger uses only lagged official policy-rate history. The professional–challenger AUC origin-by-target bootstrap interval is [0.461, 0.860], and the professional–challenger minus no-change AUC difference is -0.027 with a 95% interval of [-0.203, 0.112]. Uniform fractional assignment at cutoff ties preserves exactly 77 expected flags, or 25.0% of cells. Recursive Brier losses use the 228 cells for which both expanding-window probabilities are available. This scope application does not transport U.S. coefficients or establish machine-specific information.

C2. Bank of England Market Participants Survey

The Bank of England application uses the median modal Bank Rate path, explicitly dated MPC meetings, a public policy-history challenger that is not tuned to the evaluation outcomes, a 50-basis-point threshold, and a 25% attention capacity. The professional path has lower realized-point MAE than the challenger, and professional–challenger disagreement ranks large losses strongly. No change ranks slightly better and the paired two-way difference is imprecise, so the application does not establish challenger-specific information. Table IA.23 reports the application. Alternative thresholds and horizon cuts leave no change with the higher capture point estimate.

Table IA.23: Bank of England MaPS: Policy-Rate Forecast Application

Panel A. Forecast accuracy and large-error ranking			
	Professional	Policy-history challenger	No change
MAE (basis points)	32.97	56.05	78.45
Large-loss ROC-AUC	—	0.777	0.795
Large-loss PR-AUC	—	0.573	0.519
Panel B. Fixed-share large-error detection			
Large losses in flagged set	—	56.2%	57.9%
MAE among unflagged forecasts	—	17.27	16.95

Notes: The common support contains 261 cells across 37 origins and 36 targets, including 64 professional losses of at least 50 basis points. The challenger uses only lagged official policy-rate history. The professional–challenger AUC origin-by-target bootstrap interval is [0.619, 0.913], and the professional–challenger minus no-change AUC difference is -0.017 with a 95% interval of [-0.128, 0.092]. Uniform fractional assignment at cutoff ties preserves exactly 66 expected flags, or 25.3% of cells. This scope application does not transport U.S. coefficients or establish machine-specific information.

C3. Philadelphia Fed Survey of Professional Forecasters

The Philadelphia Fed application uses quarterly PGDP inflation, the official real-time direct-AR benchmark, no change, H2–H5, and a recursively determined large-loss threshold based only on previously available forecast errors. Professionals beat the challenger in average loss, but professional–DAR disagreement does not rank large professional losses and no change performs better. This adverse result limits transportability beyond policy-rate forecasts. Table IA.24 reports the application. The adverse ranking is not reversed by alternative loss cutoffs on the same support.

Table IA.24: Philadelphia Fed SPF: Different-Object Forecast Application

Panel A. Forecast accuracy			
	Professional	Direct-AR challenger	No change
MAE (annualized percentage points)	1.30	1.53	1.44
Panel B. Large-error ranking and fixed-share detection			
Ranking score	Professional–DAR gap	Professional–no-change gap	Difference
ROC-AUC	0.457	0.584	-0.127
Two-way 95% CI	[0.333, 0.590]	—	[-0.288, 0.037]
PR-AUC	0.093	0.191	—
Large-loss capture at 25% flagged share	0.247	0.455	—
MAE among unflagged forecasts	1.07	0.97	—

Notes: The professional forecast is the Philadelphia Fed SPF median for quarterly PGDP inflation. The challenger is the official real-time direct autoregression with AIC lag selection; no model is selected using these outcomes. The common H2–H5 panel excludes the current-quarter nowcast and contains 869 cells across 220 origins; the reported MAE rows use this support. AUC, capture, and retained-error rows use the 754 cells across 190 origins for which the recursive threshold is available. A large loss is defined recursively using the 75th percentile of professional absolute errors whose five-quarter-revised realization was already available. Intervals resample survey origins and target quarters. This is a different-object scope test, not policy-rate evidence or evidence of transportability.