When Does Model Disagreement Have Decision Value? Evidence from Professional Policy-Rate Forecasts

A-Young Park, Ph.D.

Senior Researcher, Supervisory Strategy Research Team

Financial Supervisory Service

ayoungp@fss.or.kr

August 2026

Abstract

When should a financial institution use a model signal to allocate professional attention? Using real-time policy-rate paths, I show that a reproducible public-information Ridge model has higher realized-point loss than professional forecasts, although disagreement ranks large professional forecast losses. Relative to a transparent no-change rule, however, targeted disagreement yields no reliable allocation gain: at a 24.8% review share, it captures 47.7% of large losses versus 52.1% for no change and leaves higher retained error. The estimates do not support a positive implementation-cost premium; adoption would require a cost or review-response advantage not identified in the historical data.

JEL classification: C53, E43, E47, E52, G12, G14.

Keywords: model use; attention allocation; professional forecasts; human–machine disagreement; monetary policy expectations.

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

Can a model that performs substantially worse than professional forecasters under realized-point loss still be useful to a financial institution? Professional–model disagreement offers an apparently favorable answer: large gaps identify forecasts that later incur large losses. Yet a flat no-change path produces a comparable ranking and, on the later recursive support, a more favorable allocation point estimate when both rules receive the same analytical capacity. The comparison is imprecise. The economically relevant question is therefore not whether model disagreement predicts forecast difficulty in isolation, but whether it improves a feasible decision relative to a transparent rule available to the same institution.

Professional policy-rate forecasts provide a setting in which these objects can be separated. I use Market Participant and Primary Dealer policy paths from the Federal Reserve Bank of New York's policy-expectations surveys, including the Survey of Market Expectations from 2025 onward, and compare them with an expanding-window Ridge model. The model combines real-time macroeconomic vintages, lagged market information, and targeted measures extracted from Federal Reserve documents available by each survey cutoff. Market observations strictly precede the cutoff, labels enter training only after the relevant FOMC meeting, and all preprocessing and tuning are repeated within the appropriate folds. The U.S. development panel contains 388 horizon cells across 60 survey paths and 60 target FOMCs. A later recursive sample contains 250 cells across 40 paths and 40 target FOMCs. The primary exercise asks how a hypothetical internal decision maker would allocate attention using the survey aggregate and model score; it does not reconstruct an observed New York Fed procedure or a public trading strategy.

Average realized-point loss strongly favors the professional path under the paper's reported MAE and RMSE criteria. Market Participant mean absolute error is 32.86 basis points, compared with 44.27 for federal-funds futures and 49.61 for the targeted-text Ridge. Nevertheless, targeted professional–model disagreement has an ROC-AUC of 0.733 for professional losses of at least 50 basis points, and professional MAE rises from 15.98 basis points

in the lowest disagreement quartile to 61.60 in the highest. A target-excluded cross-human score has an AUC of 0.722. A model with higher realized-point loss can therefore rank states in which professional forecast losses are large without providing a suitable replacement. The absolute disagreement score does not prescribe a revision direction.

The no-change benchmark substantially weakens a model-specific interpretation. Its AUC is 0.693 in the development panel, and the targeted-text difference is not statistically distinguishable from zero. On the recursive support, adding targeted disagreement to observable state and persistence variables raises AUC only from 0.600 to 0.610 and worsens Brier loss from 0.2288 to 0.2294. On the later recursive support and at a common 24.8% review share, the point estimates favor no change: targeted disagreement captures 47.7% of large losses and leaves retained MAE of 23.3 basis points, whereas no change captures 52.1% and leaves 22.0 basis points. The paired intervals are wide, so the evidence does not imply a precisely negative population effect or statistical equivalence. The paired intervals permit economically meaningful allocation advantages for either rule. They therefore establish neither superiority nor equivalence, while raw discrimination alone remains insufficient evidence of incremental decision value (Abadie, 2020).

The common-capacity comparison yields two sample-based break-even calculations. The first is the maximum implementation-cost premium supported by the observed allocation relative to no change. Its point estimate is nonpositive at every reported loss-cost ratio. The second is the minimum review-response advantage required when implementation costs are equal. At the sample point estimates, model-triggered review would need to be approximately 9.3% more effective in relative terms than no-change-triggered review merely to break even. This is an illustrative scenario based on the sample point estimates, not an estimated population hurdle, analyst response, or institutional welfare effect.

The economic object is benchmark-relative implementability. The analysis separates replacement accuracy, raw loss ranking, incremental probability information, and allocation at a fixed review capacity, and then maps the observed allocation gap into the cost or review-

response advantage required for adoption. It applies the implementation discipline used in empirical finance to model-assisted professional review, where scarce analytical capacity rather than trading cost is the binding friction (Gu et al., 2020; Jensen et al., 2026).

Applications to the ECB Survey of Monetary Analysts and the Bank of England Market Participants Survey reproduce the raw policy-rate ranking pattern but not an average advantage over no change. Across the U.S., ECB, and Bank of England applications, the equal-environment model-minus-no-change AUC difference is -0.001, with a 95% interval of [-0.080, 0.068]. A Philadelphia Fed inflation application yields weaker rankings than the benchmark. The evidence is therefore specific to the benchmark-relative evaluation of the reported policy-rate challengers rather than a general result about model disagreement.

2 Related Literature

2.1 Professional Forecasts and Forecast Disagreement

Professional judgment can dominate statistical alternatives in real time even when those alternatives use broad data sets. Using Greenbook vintages, Faust and Wright (2009) find that Federal Reserve staff forecasts outperform reduced-form methods for inflation, with a smaller GDP advantage after accounting for current-condition assessment. Average superiority does not imply uniform superiority across states. Giacomini and White (2006) distinguish unconditional forecast rankings from conditional predictive ability when forecasting models may be misspecified, and Giacomini and Rossi (2010) show how global comparisons can mask changes in relative performance. These results motivate separating average forecast accuracy from information about when professional performance deteriorates.

Forecast disagreement has usually been studied across human forecasters and across horizons. Its interpretation is not mechanical. Lahiri and Sheng (2010) show that forecast uncertainty combines disagreement with the perceived variability of future aggregate shocks, so disagreement is not itself a complete measure of uncertainty. Patton and Timmermann

(2010) relate the term structure and persistence of disagreement to heterogeneous priors and models, while Andrade et al. (2016) document an upward-sloping and time-varying term structure of disagreement for the federal funds rate. In a monetary-policy setting, Sastry (2026) separates disagreement arising from fundamentals, perceived policy rules, and confidence in public signals and finds only a small role for private central-bank information in the U.S. estimates.

The analysis below treats the professional and model paths as different mappings from a largely common public information set. Professional–model disagreement is an observable ranking score rather than structural uncertainty or tail risk. The question is whether that score identifies the professional forecast's own subsequent loss after comparison with a simple persistence-based alternative.

2.2 Statistical Forecast Performance and Economic Decision Value

Forecast evaluation depends on the objective being evaluated. Gneiting (2011) shows that the scoring rule must be aligned with the target functional and forecasting task, and Granger and Pesaran (2000) connect statistical forecast accuracy to the user's decision problem. Average point loss and the allocation of attention to economically large errors are therefore distinct tasks. This distinction does not favor a complex model by construction; each task still requires an informative benchmark under the same information set.

The finance literature distinguishes predictive fit from implementable economic performance. Gu et al. (2020) report both out-of-sample predictive fit and economic portfolio gains, and Jensen et al. (2026) evaluate machine-learning strategies after the trading costs that determine feasible returns enter the objective. I apply this implementability logic to professional decision support, where the relevant constraint is scarce analytical capacity rather than trading cost. The model signal is evaluated by the forecasts it sends to additional review, the cost of implementing and monitoring the signal, and the effectiveness of

that review. In monetary policy, Chernis et al. (2024) similarly synthesize model predictions and policy recommendations around explicit decision objectives instead of selecting a single forecasting winner.

This formulation does not presume that additional review is uniformly effective. It asks whether model-triggered review improves on transparent-rule-triggered review at the same capacity and then derives the cost or behavioral-response advantage required when it does not. Prediction, incremental information, and implementable decision performance are kept as separate empirical objects.

2.3 Human–Machine and Machine-Forecast Disagreement

The credibility of a machine signal depends first on point-in-time construction. Zhang et al. (2025) attribute previously reported analyst-forecast predictability and trading profits to look-ahead bias; removing the bias eliminates the alpha, and alternative machine models do not restore it. The expanding-window design here therefore restricts every feature, label, and transformation to the information available at the forecast origin.

Recent work also gives machine disagreement a distinct asset-pricing interpretation. Bali et al. (2026) measure dispersion across many return-prediction models that process common inputs and relate that dispersion to future returns. The object here is different: one professional policy path is compared with one primary public-information challenger, the outcome is the professional forecast's subsequent loss, and the feasible action is additional review under a capacity constraint. The relevant counterfactual is not whether machine disagreement predicts returns, but whether professional–model disagreement improves a review allocation relative to a transparent rule available to the forecast-producing institution.

Work comparing human decisions with machine predictions emphasizes evaluation under a common outcome and a feasible decision environment (Kleinberg et al., 2018). More recent human–AI collaboration research shows that disclosure, belief updating, and effort responses can determine whether assistance creates value (Agarwal et al., 2025). Together

with evidence that the effects of algorithmic risk information depend on how people use it (Stevenson and Doleac, 2024), these results motivate allowing review effectiveness to differ across warning rules. The historical forecast data do not identify that response; the analysis instead derives the response hurdle implied by benchmark-relative allocation.

A related machine-learning literature treats the assignment of cases between algorithms and human experts as a prediction-triage or learning-to-defer problem (Madras et al., 2018; Raghu et al., 2019; Mozannar and Sontag, 2020). That literature generally learns a deferral or triage policy using information about human and algorithm performance. The present setting does not learn a deferral policy or observe expert performance after review; it evaluates a fixed disagreement score against a transparent benchmark under a common review capacity. The contribution therefore concerns benchmark-relative evaluation of a candidate warning signal rather than optimal human–AI assignment.

3 Data and Empirical Design

3.1 Professional Forecasts and the Public-Information Challenger

The professional paths come from the Market Participant and Primary Dealer panels of the Federal Reserve Bank of New York'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. Source-level verification and archival checks are documented in the Internet Appendix. The resulting policy-path panel 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 each survey date t, professional panel j, and FOMC horizon $h \in \{1, \ldots, 7\}$, let $H^j_{t,h}$ denote the reported modal target-rate midpoint. The primary professional benchmark is the

Market Participant path, denoted $H_{t,h}$. Primary Dealers provide a second professional forecast and a within-profession disagreement control. Realized target midpoints are attached only after the relevant FOMC target date.

The primary challenger is a targeted-text Ridge that combines real-time macroeconomic and lagged market variables with targeted measures extracted from Federal Reserve documents. The model produces an expanding-window forecast of scheduled FOMC target-rate changes. A macro-only Ridge, broad-text Ridge, Elastic Net, and information-block ablations are secondary specifications. The Internet Appendix reports feature blocks, counts, preprocessing rules, and the penalty-selection procedure. The first out-of-sample test survey is in January 2019.

The targeted-text Ridge is not presented as the most accurate model that could be fit to these data. It is a reproducible, regularized challenger whose inputs can be reconstructed at each historical cutoff in a sample with 60 out-of-sample survey paths. The alternative linear models, information-block comparisons, and four-model consensus show whether the results depend on this particular text block or estimator. More flexible embeddings, large language models, or neural architectures would define different challengers and are not evaluated here. The reported evidence therefore does not establish how disagreement from those models would perform.

3.2 Real-Time Information Set and Recursive Estimation

At each survey deadline, the model information set contains release-dated macroeconomic vintages, Federal Reserve documents available by the cutoff, and financial-market observations dated strictly before the cutoff. Survey-day closes are excluded because daily closing timestamps are not uniformly earlier than survey responses. Macroeconomic vintages are reconstructed from ALFRED; delayed Tealbooks and transcripts are excluded. Imputation, scaling, feature transformations, and penalty selection are fitted separately inside each expanding training fold.

Every training row satisfies

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

The forecast model begins out-of-sample evaluation only after at least 30 prior survey dates are available for estimation. The recursive large-loss risk and allocation exercises begin later: they require 20 already realized out-of-sample survey paths in addition to the model's training history. This later burn-in accounts for the reduction from the 388-cell development panel to the 250-cell recursive sample. No target realized after a test survey enters the corresponding training sample, risk model, or recursive weight calculation.

The targeted-text path $M_{t,h}$ is the horizon-h prediction based on the information set available at survey t. Because policy-rate forecasts share persistent horizon and policy- cycle components, the analysis also reports within-horizon ranking, persistent-state residualization, and complete-path randomization within policy-cycle strata. The complete timing, estimator, and label-availability specification appears in the Internet Appendix.

3.3 Forecast Loss, Disagreement, and the Persistence Benchmark

Let $r_{t,h}$ be the subsequently realized target-rate midpoint and let $N_{t,h}$ equal the target-rate midpoint in force at survey date t. The no-change path is therefore a flat persistence forecast requiring no information after the forecast origin. Professional absolute loss and the two primary ranking scores are

$$L^H_{t,h} = |H_{t,h} - r_{t,h}|,$$

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

Raw ranking asks whether larger values of targeted disagreement tend to receive higher ranks among subsequent large-loss observations. This property is summarized using ROC-AUC and precision–recall AUC; loss gradients across disagreement quartiles are reported

separately. An AUC above 0.5 can arise even when the model has higher realized-point loss than the professional path under the reported criteria and when both disagreement and loss share predictable horizon or policy-state components. Incremental model value therefore requires favorable performance relative to D^N on identical support and improvement in a recursive risk model containing only information available by the survey deadline.

No change is used as a low-cost economic reference, not as a reconstruction of an institution's review procedure. The richer recursive comparison includes horizon, policy-cycle state, lead time, professional-path magnitude, within-profession disagreement, no-change distance, and prior same-horizon difficulty. Federal-funds futures provide a separate market forecast in the average-accuracy comparison. Because the historical data do not reveal the rule an institution actually used to allocate review, the paper compares feasible information sets rather than claiming that any one benchmark describes current practice.

The primary large-loss event is an absolute professional forecast error of at least 50 basis points, corresponding to two standard 25-basis-point policy moves. This is an operational loss definition rather than an estimated classification boundary. Continuous absolute loss remains visible throughout the analysis.

3.4 Attention Allocation and Model Adoption

Let $R_{t,h}^{j} \in \{0,1\}$ indicate that a forecast receives additional scrutiny under ranking rule $j \in \{M, N\}$. For review cost c_A and the cost c_L of leaving a large loss unattended, define

$$C(R, L) = c_A R + c_L(1 - R)\mathbf{1}\{L \geq c\}.$$

At attention share s, large-loss prevalence m, captured share k_j, and $\rho = c_L/c_A$, normalized allocation loss is

$$\ell_j(\rho) = s + \rho m(1 - k_j).$$

For the targeted-text rule M relative to no change N, the benchmark-relative adoption margin is

$$A(\rho) \equiv \ell_N(\rho) - \ell_M(\rho).$$

If ϕ_j denotes implementation and monitoring cost in review-cost units per evaluated decision, the model is preferred only when $\phi_M - \phi_N < A(\rho)$. A positive margin is the largest implementation-cost premium supported by the observed allocation; a negative margin requires a cost advantage of at least $|A(\rho)|$.

Let $\eta_j \in [0, 1]$ denote a counterfactual review-effectiveness parameter: the fraction of the modeled large-loss cost that would be prevented when a case flagged by rule j receives additional review. I do not estimate η_j; it parameterizes the behavioral response required for adoption. Total normalized institutional cost is

$$\widetilde{\ell}_j(\rho, \eta_j) = s + \rho m(1 - \eta_j k_j) + \phi_j.$$

The targeted-text rule is preferred only if

$$\rho m(\eta_M k_M - \eta_N k_N) > \phi_M - \phi_N, \tag{1}$$

$$\eta_M > \frac{\eta_N k_N}{k_M} + \frac{\phi_M - \phi_N}{\rho m k_M}. \tag{2}$$

This review-effectiveness break-even condition reports the response required for adoption; it is not a welfare estimate or a structural parameter. The main operating point requests 75% forecast coverage and yields a realized 24.8% review share.

3.5 Information Available at Public Release

The primary exercise evaluates a hypothetical internal review rule at the survey origin. The survey aggregate supplies the professional forecast object, and the design treats the aggregate path and model score as available to a decision maker after responses close but before the

relevant target outcomes. The paper does not observe the New York Fed's aggregation or review process, and individual survey respondents do not observe the cross-sectional aggregate at the survey deadline.

This estimand differs from the decision problem facing an outside user of the public archive. The New York Fed releases aggregate survey results about three weeks after the corresponding FOMC meeting, following publication of the meeting minutes (Federal Reserve Bank of New York, 2026). By then, some path targets have already been resolved.

A separate public-release analysis therefore retains only targets that remain unresolved when the aggregate path becomes observable to an outside user. The same timing discipline applies: risk models use only outcomes realized before the public-results date. The Internet Appendix reports the exact release dates, exclusions, recursive support, and crossed- dependence inference. This distinction changes the feasible information set, not the historical professional or model forecast.

3.6 Inference and Pre-Specified Comparisons

Horizon cells are dependent along both survey-path and target-FOMC dimensions. Primary benchmark-relative intervals therefore use 4,999 survey-by-target-FOMC pigeonhole-bootstrap draws (Owen, 2007; Cameron et al., 2011), with the relevant ranking scores and cutoff-tie allocations recomputed within each draw. Survey-path proper-score tests and complete-path randomization provide complementary inference for the incremental-information and ranking exercises. The 50-basis-point large-loss definition, the reported horizon and policy-state groups, and the 75% requested forecast-coverage operating point are fixed rather than selected from the reported outcomes. The Internet Appendix reports survey-only, target-only, multiway, leave-one-cluster, and randomization sensitivities where applicable. Figure 2 and the recursive risk-model sensitivities retain their stated survey-cluster procedures rather than applying the pigeonhole bootstrap universally.

4 Forecast Accuracy and Model Disagreement

4.1 Average Forecast Accuracy

Table 1 compares professional, futures, and public-information model forecasts on common support. Market Participant MAE is 32.86 basis points and Primary Dealer MAE is 33.51. Federal-funds futures have MAE of 44.27, while the targeted-text and macro Ridge models have MAEs of 49.61 and 49.42. Average forecast loss therefore favors professional judgment over the public-information challenger under the reported criteria.

The ordering is not driven by a single loss measure. Professional paths also compare favorably under RMSE. Because the surveys report modal paths while the Ridge predictions need not target the same statistical functional, this horse race establishes a realized-loss ordering under the reported criteria rather than universal dominance of one forecast functional. Bias is reported in Table 1; rounding, grid alignment, and related numerical checks appear in the Internet Appendix. These comparisons concern production of the point forecast; they do not determine whether disagreement has value as a ranking signal.

4.2 Disagreement and Large Professional Forecast Losses

Panel A of Table 2 reports raw ranking on 388 cells across 60 survey paths. Targeted professional–model disagreement has ROC-AUC 0.733 and PR-AUC 0.528 for the 98 professional losses of at least 50 basis points. Market Participant MAE increases monotonically from 15.98 basis points in the lowest disagreement quartile to 61.60 in the highest. A target-excluded cross-human score, built without the evaluated professional path, has AUC 0.722. Human-free model–futures disagreement and leave-one-professional-out comparisons further show that exact mechanical reuse of the evaluated target is not required for raw ranking.

The absolute disagreement score supplies a descriptive ranking but does not itself prescribe a corrective action. The realized-loss evidence also does not support replacing the professional path with the model path.

4.3 Common Forecast Difficulty and Incremental Information

Panel B of Table 2 and Panel A of Table 3 compare model disagreement with common forecast-difficulty measures. No-change disagreement has development-sample AUC 0.693. The targeted-minus-no-change difference is 0.040 with a paired two-way 95% interval of [-0.079, 0.157]. Whole-path randomization confirms that the observed score contains raw ranking information, but it does not demonstrate an advantage over persistence.

On the 250-cell recursive sample, a probability model using horizon, policy-cycle state, lead time, professional-path magnitude, within-profession disagreement, no-change distance, and prior same-horizon difficulty has AUC 0.600 and Brier loss 0.2288. Adding targeted disagreement raises AUC to 0.610 but increases Brier loss to 0.2294. Proper-score, residualization, and orthogonal incremental comparisons reach the same substantive result. Thus the benchmark result is not solely a horse race against a flat path: adding the model score to a richer transparent information set does not improve the reported probability loss. The evidence substantially weakens the interpretation of raw disagreement as model-specific information.

The public-release analysis produces a related distinction for outside users. Among targets still unresolved when aggregate results become public, targeted disagreement has AUC 0.735, compared with 0.680 for no change. The AUC difference is imprecise, and the recursive Brier improvement does not survive crossed-dependence inference. Public availability alone does not turn the raw ranking into established incremental probability value.

4.4 Horizon and Policy-State Variation

Panel B of Table 3 and Figure 2 report all horizon and policy-state groups. Large-loss prevalence ranges from 9.4% in H2–H3 to 40.8% in H4–H7 and changes sharply across policy regimes. Raw loss gradients vary with state, but model-minus-no-change AUC differences change sign and remain imprecise. The joint horizon interaction has $p = 0.593$. The policy-cycle interaction is statistically nonzero but does not yield stable benchmark dominance.

These patterns describe where forecasting is difficult; they do not isolate a separate model-specific mechanism.

5 Attention Allocation and Model Adoption

5.1 Allocation at a Common Review Capacity

Table 4 and Figure 1 translate ranking into a common-capacity allocation on 250 cells across 40 survey paths and 40 target FOMCs. Requested 75% forecast coverage retains 188 cells and flags 62, or 24.8%, for additional scrutiny. The realized share follows from the integer sample size; neither the share nor a score cutoff is selected from the outcomes.

On the recursive allocation sample, the point estimates favor no change at the common review capacity. The primary targeted-text rule has recursive AUC 0.723, captures 47.7% of 65 large losses, and leaves retained MAE of 23.3 basis points. No change has AUC 0.763, captures 52.1%, and leaves retained MAE of 22.0 basis points. The targeted-minus-no-change capture difference is -4.4 percentage points, with a paired two-way 95% interval of [-24.8, 18.8]. The retained-MAE difference is 1.24 basis points, with an interval of [-4.42, 7.35]. The intervals are too wide to establish a precisely negative population effect or a practically small difference. The paired intervals do not establish equivalence: they permit economically meaningful allocation advantages for either rule. The point estimates provide no benchmark-relative allocation improvement.

Internet Appendix Table IA.10 applies the same 24.8% capacity to the recursive state-and-persistence risk score. At the point estimates, it captures 33.8% of large losses and leaves retained MAE of 34.0 basis points, compared with 47.7% and 23.3 basis points for targeted disagreement and 52.1% and 22.0 basis points for no change. The paired intervals remain too wide to establish population rankings.

The four-model consensus remains a secondary comparison. It captures 53.8% of large losses and leaves retained MAE of 21.5 basis points. These values are slightly better than

no change, but the multiway comparisons remain imprecise. The risk–coverage curves show no uniform dominance across alternative capacity levels.

5.2 Implementation Costs and Benchmark-Relative Allocation

Relative to no additional review, the point estimates break even at $c_L/c_A = 2.00$ for targeted disagreement and 1.83 for no change. The transparent rule therefore reaches the review-versus-no-review threshold at a lower unattended-loss cost.

Panel C of Table 4 reports targeted minus no change, so the adoption margin is the negative of the displayed allocation-loss difference. At $\rho = 2$, targeted disagreement has allocation loss 0.520, compared with 0.497 for no change, yielding $A(2) = -0.023$ review-cost units per evaluated decision. The point estimate does not support a positive implementation-cost premium for the targeted rule. It instead requires a cost advantage of 0.023 units to break even. Across 100 evaluated forecasts, this is an implementation-and-monitoring cost advantage equal to the cost of 2.3 individual reviews relative to no change. This is only a rescaling of the normalized point estimate, not an observed dollar cost.

The paired 95% interval for the margin is [-0.120, 0.085]. The estimate is therefore a sample-based break-even calculation rather than evidence that every model deployment has negative value. The point margin is nonpositive at each reported value of ρ, even before model-building, monitoring, and explanation costs are included.

5.3 Review Effectiveness and the Break-Even Condition

When review effectiveness differs by warning rule, the adoption condition makes the missing behavioral channel explicit. For $\eta_N > 0$ and equal implementation cost, the sample point estimates imply

$$\frac{\eta_M}{\eta_N} > \frac{k_N}{k_M} = \frac{0.521231}{0.476923} = 1.093.$$

Model-triggered review must therefore be approximately 9.3% more effective in relative terms merely to offset its lower capture rate. The capture difference is imprecisely estimated, so this ratio is an illustrative mapping of the sample point estimates rather than a population threshold. The data admit capture differences of either sign and therefore do not identify a population review-effectiveness hurdle.

A positive implementation-cost premium raises the required response advantage further.

5.4 Implications for Model-Assisted Professional Decisions

The evidence separates three tests for model use. Forecast accuracy asks whether the model improves the prediction. Incremental-information performance asks whether its warning adds to a transparent benchmark. Decision performance asks whether that increment changes a feasible allocation under the institution's capacity and loss function. The model has higher realized-point loss under the reported criteria, its disagreement passes a raw ranking comparison but not the incremental probability comparison, and its primary allocation does not improve on no change at the displayed-capacity point estimates. The uncertainty around that comparison is too wide to establish equivalence or a small effect.

If model-based allocation does not improve at a common capacity, adoption can be justified only by a sufficiently lower implementation cost or a sufficiently stronger review response. The relevant counterfactual is model-triggered review versus transparent-rule-triggered review, not model use versus no review. Recent supervisory guidance similarly ties model validation and monitoring to intended use and materiality (Board of Governors of the Federal Reserve System, 2026).

That guidance emphasizes a risk-based approach tailored to a model's use and calls for ongoing monitoring and outcome analysis of whether a model remains fit for purpose. This institutional principle motivates evaluating the allocation generated by a warning rule rather than its AUC alone. It does not establish that the hypothetical review protocol studied here describes the practice of the New York Fed or another forecast-producing institution.

The conclusion is conditional on the reported challengers. A more accurate model could alter both the forecasts it flags and the response it induces. The evidence therefore does not imply that richer text representations or other model classes lack decision value; it shows that raw discrimination from the reproducible challengers studied here is not sufficient to establish it.

Credit underwriting, risk-limit exceptions, portfolio-manager overrides, trade surveillance, and supervisory review can share this allocation structure. Across these settings, the portable object is the adoption inequality, not the 9.3% calibration. Each institution must evaluate its own benchmark allocation, implementation-cost differential, and review-response advantage. Actual analyst responses and implementation costs are not observed here, so the break-even exercise maps preference regions rather than institutional welfare.

6 Evidence across Forecasting Institutions

6.1 Policy-Rate Forecasts at the ECB and Bank of England

The ECB application uses the published cross-sectional median of Survey of Monetary Analysts policy-rate paths and a public-history challenger estimated with information available at each forecast origin. The professional forecast has lower realized-point MAE than the challenger, while professional–model disagreement has AUC 0.668 for 64 large losses among 308 cells. No-change disagreement has AUC 0.695. The raw ranking recurs, but the point comparison does not favor the model score relative to persistence.

The Bank of England application uses modal Bank Rate paths from 37 Market Participants Surveys. Professional–model disagreement has AUC 0.777 for 64 large losses among 261 cells, compared with 0.795 for no change. The model-to-professional MAE ratio is 1.70. As in the ECB application, a challenger with higher realized-point MAE ranks policy-rate losses without delivering a benchmark-relative improvement.

6.2 Inflation Forecasts in the Philadelphia Fed SPF

The Philadelphia Fed application deliberately changes the forecast object. For GDP price inflation, 77 of 754 recursively evaluated cells meet the prior-error threshold. Professional forecasts have lower realized-point MAE than the direct autoregressive challenger, but professional–model disagreement has AUC 0.457, compared with 0.584 for no change. The inflation application shows that the raw-ranking pattern is not a generic property of comparing a professional forecast with a challenger that has higher realized-point MAE.

Table 5 summarizes all four settings on their respective supports. Across the three policy-rate environments, the equal-environment average professional–model AUC is 0.726 and the no-change average is 0.728; the mean difference is -0.001 with a 95% interval of [-0.080, 0.068]. The surveys, horizons, models, and policy regimes differ, so these comparisons describe scope rather than a pooled structural effect.

7 Conclusion

A reproducible public-information model can rank large professional forecast losses even when its realized-point loss exceeds that of professional forecasts under the reported criteria. In the U.S. panel, professional–model disagreement produces a steep loss gradient but does not prescribe a revision direction, while no-change disagreement reproduces much of the ranking. Similar patterns appear in the ECB and Bank of England applications; the Philadelphia Fed inflation application is unfavorable.

At a common 24.8% review share, targeted disagreement captures 47.7% of large losses versus 52.1% for no change and leaves a higher retained MAE. The wide paired intervals establish neither a precisely negative population effect nor equivalence and permit economically meaningful allocation advantages for either rule. The point estimate nevertheless yields a nonpositive adoption margin at every reported loss-cost ratio.

With equal implementation costs, the sample capture rates imply that model-triggered review must be approximately 9.3% more effective in relative terms to break even. This

calibration parameterizes an unobserved response; it is not a population hurdle or an estimate of analyst behavior or welfare.

Predictive discrimination alone does not justify adoption. Relative to the transparent rule, a model warning must improve allocation or offer lower implementation cost or a stronger review response.

References

Abadie, A. (2020). Statistical nonsignificance in empirical economics. *American Economic Review: Insights 2*(2), 193–208.

Acosta, M., A. Ajello, M. D. Bauer, F. Loria, and S. Miranda-Agrippino (2025). Financial market effects of FOMC communication: Evidence from a new event-study database. Working Paper 2025-30, Federal Reserve Bank of San Francisco.

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.

Andrade, P., R. K. Crump, S. Eusepi, and E. Moench (2016). Fundamental disagreement. *Journal of Monetary Economics 83*, 106–128.

Bali, T. G., B. T. Kelly, M. Mörke, and J. Rahman (2026). Machine forecast disagreement. The Review of Financial Studies, corrected proof, article hhag042. Published April 21, 2026; corrected June 22, 2026.

Bauer, M. D. and E. T. Swanson (2023). An alternative explanation for the Fed information effect. *American Economic Review 113*(3), 664–700.

Board of Governors of the Federal Reserve System (2026). Revised guidance on model risk management. Supervisory Letter SR 26-2. Issued April 17, 2026.

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.

Candès, E., Y. Fan, L. Janson, and J. Lv (2018). Panning for gold: Model-x knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B 80*(3), 551–577.

Chernis, T., G. Koop, E. Tallman, and M. West (2024). Decision synthesis in monetary policy. Staff Working Paper 2024-30, Bank of Canada.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Cotton, C. D. (2025). The predictability of global monetary policy surprises. Research Department Working Paper 25-14, Federal Reserve Bank of Boston.

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. Accessed August 13, 2026.

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

Giacomini, R. and H. White (2006). Tests of conditional predictive ability. *Econometrica 74*(6), 1545–1578.

Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association 106*(494), 746–762.

Granger, C. W. J. and M. H. Pesaran (2000). Economic and statistical measures of forecast accuracy. *Journal of Forecasting 19*(7), 537–560.

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

Jensen, T. I., B. Kelly, S. Malamud, and L. H. Pedersen (2026). Machine learning and the implementable efficient frontier. The Review of Financial Studies, corrected proof, article hhag022. Published March 15, 2026; corrected and typeset April 6, 2026.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018). Human decisions and machine predictions. *Quarterly Journal of Economics 133*(1), 237–293.

Lahiri, K. and X. Sheng (2010). Measuring forecast uncertainty by disagreement: The missing link. *Journal of Applied Econometrics 25*(4), 514–538.

Madras, D., T. Pitassi, and R. S. Zemel (2018). Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, Volume 31, pp. 6150–6160.

Mozannar, H. and D. Sontag (2020). Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, Volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087. PMLR.

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

Patton, A. J. and A. Timmermann (2010). Why do forecasters disagree? lessons from the term structure of cross-sectional dispersion. *Journal of Monetary Economics 57*(7), 803–820.

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

Raghu, M., K. Blumer, G. Corrado, J. Kleinberg, Z. Obermeyer, and S. Mullainathan (2019). The algorithmic automation problem: Prediction, triage, and human effort. arXiv preprint arXiv:1903.12220.

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

Stevenson, M. T. and J. L. Doleac (2024). Algorithmic risk assessment in the hands of humans. *American Economic Journal: Economic Policy 16*(4), 382–414.

Zhang, Y., Y. Zhu, and J. T. Linnainmaa (2025). Man versus machine learning revisited. *Review of Financial Studies 38*(12), 3768–3790.

Tables

Table 1: Average Forecast Accuracy

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

Notes: The common out-of-sample panel contains 388 horizon cells across 60 survey paths. Errors are forecast minus realized policy-path change in basis points. Professional modal paths have lower realized-point MAE and RMSE under these loss criteria. This is not a comparison of forecasts elicited for a common statistical functional.

Table 2: Model Disagreement and Large Professional Forecast Losses

Panel A. Loss ranking relative to no change

Evaluated loss	Ranking score	Cells	ΔAUC	Multiway 95% CI	H–M AUC	PR-AUC	NC AUC
Pooled professional	Cross-human H–M	776	0.050	[-0.057, 0.166]	0.738	0.540	0.688
Market Participant	Cross-human H–M	388	0.033	[-0.092, 0.148]	0.722	0.523	0.689
Market Participant	Own-human H–M	388	0.040	[-0.079, 0.157]	0.733	0.528	0.693
Primary Dealer	Cross-human H–M	388	0.068	[-0.044, 0.188]	0.755	0.559	0.687

Panel B. Incremental probability forecasts and proper scores

Comparison	Cells	ROC-AUC	PR-AUC	Brier	ΔBrier	Path-CPA p
Human + no-change + state	250	0.600	0.413	0.2288	—	—
+ targeted H–M gap	250	0.610	0.421	0.2294	0.0007	0.747

Panel C. Professional forecast losses across disagreement quartiles

Quartile	Cells	Mean H–M gap	Professional MAE	Median AE	Large-loss rate
Q1	97	0.00	15.98	0.00	12.4%
Q2	97	12.11	21.65	0.00	13.4%
Q3	97	25.00	32.22	0.00	25.8%
Q4	97	63.40	61.60	25.00	49.5%

Notes: Panel A uses the common 388-cell, 60-survey development panel; the pooled row stacks Market Participant and Primary Dealer losses on the same cells and does not treat 776 rows as independent. ΔAUC is H–M minus the geometry-matched no-change score. Multiway intervals preserve survey paths and target-FOMC dependence. Panel B is restricted throughout to the 250-cell recursive sample. Its ΔBrier is the first model minus the comparator, so negative values are favorable; CPA is a survey-path equal-predictive-ability test. ΔBrier is computed from unrounded probability forecasts and need not equal the difference between displayed rounded Brier values. Panels A–C use targeted-text Ridge disagreement. Panels A and C summarize raw ranking and loss gradients; Panel B evaluates incremental probability performance. The 50-basis-point event is an operational loss threshold rather than an estimated cutoff. Because the ranking scores are absolute distances, they do not encode a revision direction. The zero mean H–M gap in Q1 reflects tied exact agreements on the policy grid.

Table 3: Disagreement, Shared Components, and Forecast Heterogeneity

Panel A. Removing exact evaluated-forecast overlap

Evaluated loss	Ranking score	AUC	Multiway 95% CI	ΔAUC vs. NC	Δ 95% CI
Market Participant	Own-human H–M	0.733	[0.601, 0.847]	0.040	[-0.079, 0.157]
Market Participant	Cross-human H–M	0.722	[0.591, 0.842]	0.033	[-0.092, 0.148]
Market Participant	Human-free M–futures	0.775	[0.662, 0.874]	—	—
Primary Dealer	Cross-human H–M	0.755	[0.619, 0.873]	0.068	[-0.044, 0.188]
Pooled	Cross-human H–M	0.738	[0.617, 0.856]	0.050	[-0.057, 0.166]

Panel B. Forecast-horizon and policy-cycle heterogeneity

Group	Cells	Losses	H–M AUC	No-change AUC	ΔAUC	Multiway 95% CI
H2–H3	117	11	0.745	0.720	0.025	[-0.156, 0.219]
H4–H7	211	86	0.709	0.628	0.081	[-0.018, 0.183]
Pre-pandemic	63	32	0.546	0.454	0.092	[-0.080, 0.266]
Pandemic / ELB	110	16	0.778	0.728	0.049	[-0.124, 0.250]
Rapid tightening	81	33	0.841	0.890	-0.049	[-0.222, 0.098]
Plateau / easing	134	17	0.716	0.733	-0.017	[-0.142, 0.088]

Panel C. Formal interaction tests

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

Notes: The common development panel contains 388 cells across 60 survey paths and 60 target FOMCs. Cross-human rows remove the evaluated professional forecast from the ranking score and use the corresponding other-human no-change comparator. Multiway intervals reweight complete paths along survey and target-FOMC dimensions. The human-free score has no geometry-matched human no-change comparator. Horizon groups and the four policy-cycle states were fixed before this analysis; no null group is suppressed. H1 contains only one large forecast loss and is omitted from Panel B's inferential comparison. The heterogeneity estimates are descriptive rather than confirmatory.

Table 4: Attention Allocation under a Common Capacity Constraint

Panel A. Attention allocation at 75% forecast coverage

Signal	Cells	Flagged	Loss capture	Retained MAE	Full MAE	ROC-AUC
Targeted-text disagreement	250	24.8%	47.7%	23.3	33.7	0.723
Four-model consensus disagreement	250	24.8%	53.8%	21.5	33.7	0.734
No-change disagreement	250	24.8%	52.1%	22.0	33.7	0.763
Primary Dealer–Market Participant disagreement	250	24.8%	32.7%	32.3	33.7	0.562

Panel B. Break-even loss-to-review cost ratios

Ranking signal	Flagged share s	Loss capture k	Break-even c_L/c_A
Targeted-text disagreement	24.8%	47.7%	2.00
Four-model consensus disagreement	24.8%	53.8%	1.77
No-change disagreement	24.8%	52.1%	1.83
Primary Dealer–Market Participant disagreement	24.8%	32.7%	2.92

Panel C. Benchmark-Relative Allocation Performance

Metric	Targeted	No change	Difference	Multiway 95% CI
Large-loss capture (%)	47.7	52.1	-4.4	[-24.8, 18.8]
Retained MAE (bp)	23.3	22.0	1.24	[-4.42, 7.35]
Allocation loss, $\rho = 1$	0.384	0.372	0.012	[-0.042, 0.060]
Allocation loss, $\rho = 1.5$	0.452	0.435	0.017	[-0.064, 0.090]
Allocation loss, $\rho = 2$	0.520	0.497	0.023	[-0.085, 0.120]
Allocation loss, $\rho = 3$	0.656	0.621	0.035	[-0.127, 0.180]
Allocation loss, $\rho = 5$	0.928	0.870	0.058	[-0.212, 0.300]

Notes: The recursive sample contains 250 cells across 40 survey paths and 40 target FOMCs; requested 75% forecast coverage retains 188 and flags 62 cells. Targeted-text disagreement is the primary signal, and the consensus is a secondary comparison. Panel B evaluates $\ell(\rho) = s + \rho m(1-k)$ against $\ell_0(\rho) = \rho m$, where $\rho = c_L/c_A$ and $m = 65/250$, assuming that additional scrutiny prevents a captured large-loss cost. Panel C reports targeted minus no change; positive differences in capture and negative differences in retained MAE and allocation loss favor the targeted rule. Panel C differences are computed before rounding and may not equal differences between displayed rounded levels. At equal implementation cost and $\eta_N > 0$, the unrounded tie-aware capture shares $k_N = 0.521231$ and $k_M = 0.476923$ imply the point review-effectiveness ratio $k_N/k_M = 1.093$. This ratio maps the sample point estimates and is not a population threshold. Paired 95% intervals use 4,999 draws from a pigeonhole bootstrap over survey and target-FOMC dimensions that recomputes both allocations, including cutoff ties. The calculations do not estimate institutional welfare or causal effects.

Table 5: Professional and Model Forecasts Across Institutions

Panel A. Base rates, average accuracy, and raw ranking

Environment	Cells / origins	Large losses N (share)	Model / professional MAE	Professional–model AUC [95% CI]	Within-horizon AUC
United States	388 / 60	98 (25.3%)	1.51	0.733 [0.601, 0.847]	0.664
ECB SMA	308 / 42	64 (20.8%)	1.79	0.668 [0.461, 0.860]	0.605
BoE MaPS	261 / 37	64 (24.5%)	1.70	0.777 [0.621, 0.916]	0.721
Policy-rate environments: equal-weight mean	957 / 139 pooled	226 (23.6%)	1.67	0.726 [0.633, 0.818]	0.663
Philadelphia SPF (GDP price inflation)	754 / 190	77 (10.2%)	1.20	0.457 [0.336, 0.588]	0.454

Panel B. Increment relative to no change

Environment	No-change AUC	Professional–model minus no-change AUC [95% CI]
United States	0.693	0.040 [-0.079, 0.157]
ECB SMA	0.695	-0.027 [-0.203, 0.112]
BoE MaPS	0.795	-0.017 [-0.129, 0.087]
Policy-rate environments: equal-weight mean	0.728	-0.001 [-0.080, 0.068]
Philadelphia SPF (inflation)	0.584	-0.127 [-0.285, 0.025]

Notes: A large policy-rate loss is an error of at least 50 basis points. The Philadelphia SPF threshold is the recursively estimated 75th percentile of prior forecast errors because the outcome has a different scale. The MAE ratio is model MAE divided by professional MAE. AUC uses absolute professional–model disagreement to rank subsequent large professional losses. No change uses the professional forecast's distance from an unchanged policy or outcome. Confidence intervals use 4,999 draws from a pigeonhole bootstrap over origin and target dimensions. Cells, origins, and loss counts in the policy-rate summary are pooled totals; MAE ratios and AUCs give equal weight to the three environments. The summary is descriptive rather than a formal meta-analysis, and all within-environment comparisons use identical support. Equal-weight means and model-minus-no-change differences are computed from unrounded environment-level estimates; displayed components are rounded.

Figures

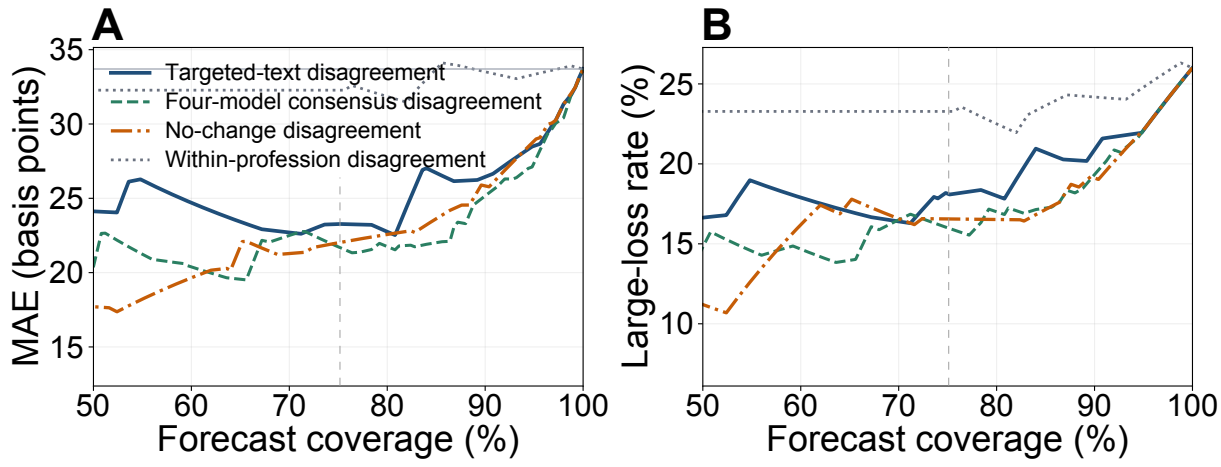


Figure 1: Risk–Coverage Performance at a Common Attention Capacity

Notes: Panel A reports retained professional-forecast MAE and Panel B reports the retained 50-basis-point-miss rate as forecast coverage increases. The sample contains 250 recursively scored cells across 40 surveys; all signals and labels obey the survey-date cutoff. The targeted-text line is the primary single-model score, while the four-model consensus is a separately reported comparison. The horizontal line in Panel A is unconditional professional MAE. The thin vertical dashed line marks the descriptive operating point: 75% requested coverage, corresponding to 75.2% realized coverage in the 250-cell sample. No score dominates across all attention capacities, and the curves compare allocation rules rather than causal welfare effects.

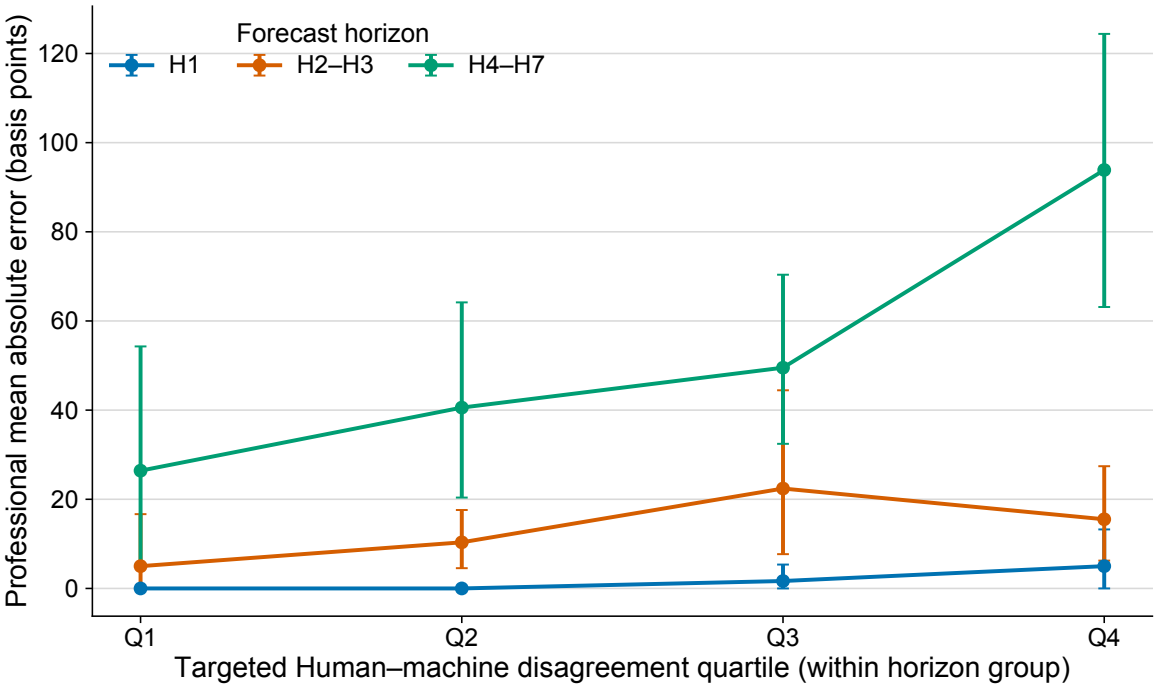


Figure 2: Professional Loss across Disagreement States and Forecast Horizons

Notes: Each line reports professional MAE across within-horizon-group quartiles of the primary targeted-text gap. Horizon groups and quartiles were fixed without outcomes; vertical bars are percentile intervals from 4,999 survey-cluster bootstrap draws that resample complete professional paths. The point-in-time development sample has 388 cells across 60 surveys. H1 has only one 50-basis-point miss, and Panel B of Table 3 therefore withholds its AUC interval. The figure is a descriptive heterogeneity summary, not a causal effect.

Internet Appendix to

"When Does Model Disagreement Have Decision Value? Evidence from Professional Policy-Rate Forecasts"

A-Young Park, Ph.D.

Financial Supervisory Service

August 2026

Part A. Real-Time Data and Forecast Construction

A1. Sample Construction

The U.S. analysis uses three samples. Table IA.1 records their calendar support and inclusion rules. The recursive probability and allocation sample begins only after 20 realized out-of-sample survey paths are available for recursive risk estimation, accounting for the reduction from 388 to 250 horizon cells.

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

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 and allocation sample	250	40	40	2021-07-19–2026-06-08	Twenty realized out-of-sample survey paths provide the minimum recursive risk/allocation history; all labels must be realized before each test survey. Target dates 2021-07-28–2026-06-17.
Secondary MP2 sample	43	43	43	2019-01-22–2025-09-08	Complete H2–H7 paths and observed MP2; target dates 2019-01-30–2025-09-17.

Notes: The reduction from 388 to 250 cells reflects the requirement of 20 realized out-of-sample survey paths for recursive risk and allocation estimation, in addition to the challenger model's own training history. It is not an outcome-selected subsample. The MP2 analysis is event-level and uses one observation per survey–FOMC pair.

A2. Public-Information Challenger Specification

Table IA.2 collects the information blocks, estimator, tuning rule, and timing restrictions for the primary targeted-text Ridge model. It also distinguishes that single-model score from the four-model consensus, which is reported only as a secondary comparison in the attention-allocation analysis.

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 and market block	281 release-dated real-time macroeconomic and lagged financial-market features	ALFRED vintage by survey date; market observation date strictly precedes the survey cutoff
Federal Reserve text block	14 targeted measures: document-level hawkishness and topics for inflation, labor, growth, financial conditions, uncertainty, guidance, tightening, and easing	Only documents available by the survey cutoff; delayed Tealbooks and transcripts excluded
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 regression with 299 total input features; median imputation with missingness indicators and standardization	All preprocessing is fitted within the relevant training fold
Penalty selection	$\alpha \in \{1, 10, 100, 1000\}$ chosen by mean absolute validation loss	Three inner time-series splits; re-estimated 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
Multi-model comparison	Cell-wise median of macro Ridge, targeted-text Ridge, broad-text Ridge, and Elastic Net macro-plus-text paths	Reported separately from the primary targeted-text score

Notes: Feature counts are read from the generated prediction and feature-dictionary files. The primary targeted-text Ridge uses 281 macro and market features, 14 targeted text features, and four horizon or context features. The specification is evaluated as a reproducible public-information challenger, not selected as the most accurate model in the reported sample.

A3. Integer Policy-Tick Outcome Construction

Table IA.3 converts stored basis-point forecasts and outcomes to the integer 25-basis-point policy grid before forming errors. It reproduces 98 Participant two-tick losses and the approximately 0.733 targeted AUC; tiny numerical perturbations leave every label unchanged.

Table IA.3: Integer Policy-Tick Outcome Check

check	Observed
Participant two-tick loss events	98
Targeted disagreement AUC	0.732935
Maximum deviation from the 25bp grid	$< 10^{-12}$
Integer labels vs. grid-tolerant operational labels	0
Literal-float labels vs. integer operational labels	21
Labels changed by tiny numerical perturbation	0

Notes: Forecast and realized path changes are converted from stored basis points to integer 25-basis-point ticks with round-half-away-from-zero before errors are formed. The operational event is $|q| \geq 2$. The integer construction exactly reproduces the grid-tolerant operational labels. The 21 literal-float differences are theoretically exact 50-basis-point losses stored infinitesimally below 50; this numerical sensitivity is retained as a record and does not alter the 50-basis-point economic threshold.

Part B. Benchmark-Relative Forecast Evaluation

B1. Threshold and Numerical Sensitivity

Table IA.4 reports the fixed operational definition together with literal-comparison and rounding sensitivities. These alternatives change numerical ties but do not alter the information set or the primary targeted-text model.

Table IA.4: Numerical sensitivity of the operational 50-basis-point threshold

Definition	Misses	Targeted AUC	Consensus AUC	No-change AUC	Dealer gap AUC
Grid-tolerant operational $\geq$ 50bp	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 first row applies the fixed grid-tolerance rule; the second preserves the previously reported literal-comparison sensitivity. The 0.733/0.695 difference is numerical, not a different information set or model selection.

B2. Pre-Survey Whole-Path Randomization

Table IA.5 reports the cutoff-feasible randomization, which permutes complete pre-survey targeted-machine forecast-change paths and never uses the recipient realized outcome. The actual score beats both pre-survey reference distributions, but Main Table 2 does not show a reliable advantage over no change. The exercise therefore supports raw loss ranking rather than machine-specific incremental value.

Table IA.5: Pre-Survey Whole-Path Randomization

Null / statistic	Actual	Null mean	Actual-minus-null	Favorable p	Two-sided p
Cycle/support path permutation / ROC-AUC	0.733	0.675	0.058	0.017	0.031
Cycle/support path permutation / PR-AUC	0.528	0.432	0.096	0.007	0.008
Support-only path permutation / ROC-AUC	0.733	0.638	0.094	0.005	0.010
Support-only path permutation / PR-AUC	0.528	0.434	0.093	0.017	0.029

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. 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 Table 2.

B3. Crossed-Dependence Inference

Table IA.6 reports survey-path, target-FOMC, and pigeonhole- bootstrap intervals (Cameron et al., 2011; Owen, 2007). Each draw recomputes the ranking score and matched baseline on common support. Leave-one-survey and leave-one-target ranges are reported without selecting the most favorable dependence rule.

Table IA.6: AUC Contrasts under Crossed Dependence

Loss	Score	Survey CI	Target CI	Multiway CI	LOO survey	LOO target
participant	Own-human H–M	[-0.041, 0.120]	[-0.029, 0.100]	[-0.079, 0.157]	[0.017, 0.057]	[0.024, 0.052]
participant	Cross-human H–M	[-0.043, 0.112]	[-0.038, 0.097]	[-0.092, 0.148]	[0.011, 0.050]	[0.016, 0.042]
participant	Human-free M–futures	[-0.009, 0.170]	[0.012, 0.151]	[-0.049, 0.204]	[0.062, 0.107]	[0.070, 0.093]
dealer	Own-human H–M	[0.002, 0.200]	[0.027, 0.163]	[-0.022, 0.236]	[0.070, 0.116]	[0.081, 0.104]
dealer	Cross-human H–M	[-0.018, 0.157]	[0.008, 0.128]	[-0.044, 0.188]	[0.046, 0.086]	[0.053, 0.073]
dealer	Human-free M–futures	[-0.007, 0.225]	[0.034, 0.175]	[-0.031, 0.252]	[0.077, 0.134]	[0.092, 0.114]
pooled cross human	Cross-human H–M	[-0.027, 0.127]	[-0.009, 0.109]	[-0.057, 0.166]	[0.028, 0.068]	[0.034, 0.055]

B4. Geometry-Conditioned Complete-Path Randomization

Table IA.7 reports the estimated path randomization, which follows conditional-randomization logic (Candès et al., 2018). Complete machine paths are reassigned within identical horizon-support and policy-cycle strata using only pre-outcome geometry for donor matching. Four singleton paths are excluded rather than relaxing the conditioning rule. Because the path distribution is estimated, this is not an exact finite-sample test.

Table IA.7: Geometry-Conditioned Complete-Path Randomization

Loss	Score	Statistic	Actual	Null mean	Null 95% interval	Favorable p
participant	Own-human H–M	auc	0.731	0.683	[0.624, 0.740]	0.053
participant	Own-human H–M	pr auc	0.531	0.434	[0.355, 0.513]	0.007
participant	Own-human H–M	delta auc no change	0.044	-0.005	[-0.063, 0.053]	0.053
participant	Own-human H–M	quartile mae gradient	46.579	42.301	[26.579, 57.895]	0.301
participant	Cross-human H–M	auc	0.722	0.675	[0.615, 0.734]	0.058
participant	Cross-human H–M	pr auc	0.528	0.431	[0.348, 0.516]	0.012
participant	Cross-human H–M	delta auc no change	0.037	-0.010	[-0.070, 0.049]	0.058
participant	Cross-human H–M	quartile mae gradient	49.737	40.002	[23.947, 56.316]	0.120
dealer	Own-human H–M	auc	0.764	0.672	[0.607, 0.737]	0.002
dealer	Own-human H–M	pr auc	0.569	0.425	[0.339, 0.517]	0.002
dealer	Own-human H–M	delta auc no change	0.100	0.009	[-0.057, 0.073]	0.002
dealer	Own-human H–M	quartile mae gradient	53.158	42.513	[26.053, 59.474]	0.105
dealer	Cross-human H–M	auc	0.753	0.688	[0.625, 0.748]	0.018
dealer	Cross-human H–M	pr auc	0.563	0.429	[0.344, 0.517]	0.001
dealer	Cross-human H–M	delta auc no change	0.072	0.007	[-0.055, 0.067]	0.018
dealer	Cross-human H–M	quartile mae gradient	49.474	45.640	[29.737, 61.316]	0.322

Notes: Each of 4,999 draws reassigns complete machine paths from nonself donors with identical horizon support and policy-cycle state. Donors also match coarse, outcome-free bins of professional no-change distance, path level, path slope, and lead time when at least three are available; otherwise the five nearest donors on those pre-survey features are used within the same support/cycle. Singleton support–cycle strata are disclosed and excluded because the declared randomization has no admissible nonself donor; conditions are not relaxed. Recipient realized outcomes never enter donor construction. This estimated-conditional randomization is calibration-sensitive and is not an exact finite-sample CRT.

B5. Incremental Information beyond Forecast Difficulty

Table IA.8 reports the expanding-window risk models, discrimination and proper scores, risk-coverage loss, retained error, miss capture, and survey-path comparisons summarized in main Table 2. The survey-cluster intervals are reported in the same table family rather than repeated in a separate table.

Table IA.9 reports a specification in which both the loss indicator and ranking score are residualized through expanding-window nuisance fits using only labels available before each test survey. The construction is motivated by orthogonal-score logic (Chernozhukov et al., 2018), but remains a predictive partialling-out sensitivity rather than a causal DML estimand.

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 H–M gap	250	0.610	0.421	0.2294	31.65	33.78
+ residual H–M gap	250	0.579	0.372	0.2625	33.93	33.38

Panel B. Targeted gap beyond professional disagreement, no change, and policy state

Metric	Favorable difference	95% CI low	95% CI high	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	95% CI low	95% CI high	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-human H–M	brier	0.0007	[-0.0033, 0.0047]	0.747	0.233
Own-human H–M	log loss	-0.0114	[-0.0306, 0.0078]	0.276	0.191
Cross-human H–M	brier	0.0033	[-0.0026, 0.0091]	0.323	0.374
Cross-human H–M	log loss	0.0500	[0.0043, 0.0961]	0.026	0.224
Human-free M–futures	brier	0.0086	[0.0008, 0.0163]	0.028	0.116
Human-free M–futures	log loss	0.0397	[0.0124, 0.0673]	0.003	0.011

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. Positive favorable differences always favor the augmented model; signs are reversed for losses. Intervals use 4,999 survey-cluster bootstrap draws. 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.

Table IA.9: Forward Orthogonal Incremental Sensitivity

Loss	Score	Outcome nuisance	θ/residual SD	Multiway 95% CI	p
participant	Own-human H–M	logistic ridge	0.101	[-0.022, 0.228]	0.112
participant	Own-human H–M	linear probability ridge	0.104*	[-0.005, 0.224]	0.064
participant	Cross-human H–M	logistic ridge	0.077	[-0.044, 0.184]	0.191
participant	Cross-human H–M	linear probability ridge	0.072	[-0.030, 0.167]	0.154
dealer	Own-human H–M	logistic ridge	0.108	[-0.044, 0.232]	0.160
dealer	Own-human H–M	linear probability ridge	0.114*	[-0.018, 0.217]	0.088
dealer	Cross-human H–M	logistic ridge	0.147*	[-0.021, 0.289]	0.080
dealer	Cross-human H–M	linear probability ridge	0.158**	[0.008, 0.285]	0.041

Notes: Both the integer-tick loss indicator and ranking score are residualized with forward nuisance fits using only labels available before each survey. Logistic Ridge and linear-probability Ridge outcome nuisances are both shown; the ranking-score nuisance is fixed Ridge. This is a predictive partialling-out sensitivity, not a causal DML estimand. *, **, and *** denote statistical significance at the 10, 5, and 1 percent levels, respectively; p-values are based on the two-sided multiway procedure.

The same recursively estimated probabilities can also be used to allocate the fixed review capacity. Table IA.10 compares the observable-state probability, its targeted-gap augmentation, targeted disagreement, and no-change disagreement on identical support. This exploratory comparison was not part of the primary allocation design and should not be interpreted as independent validation.

Table IA.10: Recursive Observable-State Risk Models as Attention-Allocation Rules

Panel A. Same-capacity point estimates

Rule or score	AUC	PR-AUC	Capture (%)	Retained MAE	Loss, $\rho = 2$	RC-MAE area
Observable state and persistence	0.600	0.413	33.8	34.04	0.592	31.89
+ targeted H–M gap	0.610	0.421	33.8	33.78	0.592	31.65
Targeted H–M disagreement	0.723	0.565	47.7	23.27	0.520	25.30
No-change disagreement	0.763	0.638	52.1	22.03	0.497	22.81

Panel B. Paired differences with multiway 95% intervals

Candidate minus comparator	ΔAUC	ΔCapture (pp)	ΔRetained MAE (bp)
Observable state minus targeted H–M	-0.122 [-0.378, 0.137]	-13.8 [-42.0, 15.7]	10.77 [-2.96, 25.32]
Observable state minus no change	-0.163 [-0.404, 0.067]	-18.3 [-44.2, 9.7]	12.01 [-0.04, 25.54]
+ targeted gap minus observable state	0.010 [-0.001, 0.023]	0.0 [-2.5, 7.4]	-0.27 [-3.81, 0.60]

Notes: The recursive support contains 250 cells across 40 survey paths and 40 target FOMCs, including 65 professional losses of at least 50 basis points. Requested 75% forecast coverage retains 188 cells and flags 62, or 24.8%, under every rule. The observable-state probability 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 human–model disagreement. The existing probabilities are evaluated without refitting them for the allocation exercise. Higher AUC, PR-AUC, and capture are favorable; lower retained MAE, allocation loss, and RC-MAE area are favorable. Panel B reports candidate minus comparator. Its intervals use 4,999 paired survey-by-target-FOMC pigeonhole-bootstrap draws and recompute each allocation, including cutoff ties. This exploratory comparison was not part of the primary allocation design and should not be interpreted as independent validation. The AUC interval differs from Table IA.8 because this table uses paired survey-by-target-FOMC pigeonhole inference, whereas Table IA.8 reports survey-cluster inference for the recursive risk-model comparison. The wide intervals do not establish equivalence, population rankings, or a causal effect of review.

B6. Symmetry across Professional Forecast Units

The design can be evaluated symmetrically without allowing the evaluated professional path to enter its own ranking score. For Market Participant loss, the score is the absolute Dealer–machine gap; for Primary Dealer loss, it is the absolute Participant–machine gap. Table IA.11 shows positive raw ranking in both directions. This removes exact own-human arithmetic but not common survey-date, policy-state, or realized-outcome dependence, and is therefore a within-survey symmetry check rather than independent evidence from another setting.

Table IA.11: Symmetry Across Professional Forecast Panels

Evaluated loss	Ranking score (target excluded)	Cells	Losses	Raw AUC [95% CI]	ΔAUC vs. NC [95% CI]
Market Participant	|Dealer − Machine|	388	98	0.722 [0.591, 0.842]	0.033 [-0.092, 0.148]
Primary Dealer	|Participant − Machine|	388	96	0.755 [0.619, 0.873]	0.068 [-0.044, 0.188]
Pooled professional	Leave-one-human-out score	776	194	0.738 [0.617, 0.856]	0.050 [-0.057, 0.166]

Notes: Each target-unit row uses the other professional path when constructing the human–machine ranking score, so the evaluated forecast is absent from its own score. Brackets are 4,999-draw survey-by-target-FOMC pigeonhole-bootstrap intervals. The pooled row stacks the two professional outcomes on the same 388 cells and does not treat 776 rows as independent. Both raw AUC intervals exclude 0.5, but both matched no-change differences include zero. This is a symmetric comparison across forecast-producing units within the NY Fed surveys, not an independent cross-country, cross-survey, or temporal evidence; the two units share survey dates, policy regimes, and realized FOMC outcomes.

B7. Forecast-Horizon and Policy-State Heterogeneity

Table IA.12 shows that 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: When Does Human–Machine Disagreement Rank Large Losses?

Panel A. Forecast-horizon groups

Horizon group	Cells	Misses (rate)	H–M AUC	No-change AUC	Δ AUC	95% CI
H1	60	1 (1.7%)	0.932	0.915	+0.017	Not reported
H2–H3	117	11 (9.4%)	0.745	0.720	+0.025	[-0.156, 0.219]
H4–H7	211	86 (40.8%)	0.709	0.628	+0.081	[-0.018, 0.183]

Panel B. Existing policy-cycle states

Policy-cycle state	Cells	Misses (rate)	H–M AUC	No-change AUC	Δ AUC	95% CI
Pre-pandemic	63	32 (50.8%)	0.546	0.454	+0.092	[-0.080, 0.266]
Pandemic / ELB	110	16 (14.5%)	0.778	0.728	+0.049	[-0.124, 0.250]
Rapid tightening	81	33 (40.7%)	0.841	0.890	-0.049	[-0.222, 0.098]
Plateau / easing	134	17 (12.7%)	0.716	0.733	-0.017	[-0.142, 0.088]

Panel C. Formal interaction tests

Interaction family	Reference group	Joint cluster-robust p
H–M gap × horizon group	H2–H3	0.593
H–M gap × 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. The H–M score is the targeted public-model gap. 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; 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.

B8. Allocation Overlap and Alternative Signals

At the unchanged 62-cell capacity, the secondary four-model consensus and no-change rules have nearly identical marginal capture rates. Table IA.13 asks whether they nevertheless flag different realized large losses. Expected overlap uses the reported tie-aware marginal allocation, and sharp ranges allow every admissible cutoff-tie allocation. The unique-capture difference is small and imprecise. This supports alternative challenger views, not consensus-specific dominance. Main Table 4 Panel B already reports the corresponding break-even loss-to-review cost ratios, so no separate consensus decision-cost figure is repeated here.

Table IA.13: Overlap of Large Forecast Losses Captured by Four-Model Consensus and No Change

Category	Expected losses	Share of 65	Sharp tie range	Multiway 95% CI
Four-model consensus only	8.12	12.5%	7–11	[1.8, 29.3]%
No-change only	7.00	10.8%	7–7	[0.0, 32.0]%
Both rules	26.88	41.4%	24–28	[16.2, 62.6]%
Neither rule	23.00	35.4%	23–23	[12.0, 60.0]%

Notes: The sample contains 250 recursively scored cells and 65 professional losses of at least 50 basis points. The four-model consensus and no-change rules each flag 62 cells (24.8%) under the 75% forecast-coverage rule. Expected overlap applies the reported tie-aware marginal allocation and independent uniform tie resolution across rules. Sharp ranges allow every allocation within each rule's fixed cutoff tie while preserving exactly 62 flags. Multiway intervals use 4,999 paired survey-by-target-FOMC pigeonhole-bootstrap draws and recompute both rules. This overlap decomposition does not establish consensus superiority, a causal effect of additional scrutiny, or external validity.

B9. Review-Effectiveness Requirements for Model Adoption

The common-capacity differences are imprecisely estimated. Table IA.14 reports the observed comparisons, multiway intervals, and normal-reference minimum detectable effect magnitudes implied by those interval widths. The calculation summarizes precision; it does not replace the reported pigeonhole-bootstrap inference or establish equivalence.

Table IA.14: Precision of Benchmark-Relative Allocation Comparisons

Metric	Targeted minus no change	Multiway 95% CI	Approx. 80% MDE
Large-loss capture (percentage points)	-4.4	[-24.8, 18.8]	31.2
Retained MAE (bp)	1.24	[-4.42, 7.35]	8.41
Allocation loss, $\rho = 2$ (review-cost units)	0.023	[-0.085, 0.120]	0.147

Notes: Higher capture favors targeted disagreement; lower retained MAE and allocation loss favor targeted disagreement. The confidence intervals reproduce Main Table 4. For each row, the normal-reference standard error is the interval width divided by $2z_{0.975}$, and the two-sided 5%, 80%-power minimum detectable effect magnitude is $(z_{0.975} + z_{0.80})\widehat{\text{SE}}$. Because the underlying intervals use a survey-by-target-FOMC pigeonhole bootstrap and need not be symmetric or Gaussian, these MDEs are approximate summaries of current precision, not exact power calculations. They are not used to select a sample, threshold, outcome, or specification.

Let $\eta_j \in [0, 1]$ denote a counterfactual review-effectiveness parameter: the fraction of the modeled large-loss cost that would be prevented when a case flagged by rule j receives additional review. Neither η_M nor η_N is estimated in the historical forecast data. Let ϕ_j be normalized implementation and monitoring cost. Total institutional cost is

$$\widetilde{\ell}_j(\rho, \eta_j) = s + \rho m(1 - \eta_j k_j) + \phi_j.$$

The targeted-text rule is preferred to no change only if

$$\eta_M > \frac{\eta_N k_N}{k_M} + \frac{\phi_M - \phi_N}{\rho m k_M}.$$

This condition is the review-effectiveness break-even condition. For $\eta_N > 0$ and equal implementation cost, the sample capture rates imply the sample-implied point hurdle $\eta_M/\eta_N > 0.521231/0.476923 = 1.093$: review prompted by the targeted-text warning must be approxi-

mately 9.3% more effective in relative terms merely to offset its lower capture rate. Because the capture interval includes effects of either sign, this ratio is an illustrative mapping of the sample point estimates rather than an estimated population hurdle. If no-change-triggered review prevents 50% of the modeled cost, targeted-model review must prevent more than approximately 54.6%; if the former prevents 75%, the latter must prevent more than approximately 82.0%. These examples illustrate the adoption condition rather than estimate analyst behavior. The capture difference is imprecisely estimated, so the data do not precisely estimate the corresponding population hurdle. A positive implementation-cost premium shifts the hurdle upward. Because $\eta_M \leq 1$, the equal-cost hurdle is feasible at the sample point estimates only when $\eta_N < k_M/k_N = 0.915$. Above that no-change effectiveness, an equally costly targeted-text rule cannot offset its lower capture rate through effectiveness alone. This feasibility boundary is sample-implied and imprecise. The calculation does not estimate either warning rule's review effectiveness, the causal effect of review on professional forecast losses, or an institutional welfare effect.

B10. Information Available at Public Release

The main allocation exercise 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 public-results date; one origin without an exact official minutes-date match is excluded rather than assigned an imputed date.

Table IA.15 shows that, on the resulting 322-cell support, 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.15: 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 H–M 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 H–M 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 public-results date, proxied by 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. All recursive training labels precede the public-results 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 H–M AUC interval includes 0.5, and ranking weakens after removing unconditional horizon composition. The human-free challenger-versus-no-change score is null. The evidence therefore does not establish that model disagreement is more informative than persistence. Tables IA.16 and IA.17 report the main and robustness results.

Table IA.16: ECB Survey of Monetary Analysts: Policy-Rate Forecast Application

	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 score	—	0.1496	0.1593
Fixed 25% attention-allocation exercise			
Large losses captured	—	43.8%	50.0%
Retained professional MAE	—	23.91	24.52

Notes: The sample contains 308 realized DFR-path cells from 42 independent ECB SMA rounds released between June 2021 and July 2026. The professional forecast is the published cross-sectional median. The challenger is a fixed expanding-window Ridge using only lagged official DFR history; every training label is realized by the relevant survey deadline. Large loss means at least 50 basis points and 25% of cells are assigned additional scrutiny; neither is optimized. The H–M AUC survey-by-target-month bootstrap interval is [0.462, 0.860]. The H–M minus no-change AUC difference is -0.027, with interval [-0.206, 0.109]. This independent-environment exercise is a descriptive scope test and does not establish machine-specific information or validate the U.S. model.

Table IA.17: Robustness Checks for the ECB Application

Ranking score	ROC-AUC	PR-AUC
Raw professional–challenger gap	0.668	0.478
Raw professional–no-change gap	0.695	0.477
Human-free challenger–no-change gap	0.495	0.214
H–M rank within horizon	0.605	0.376
No-change rank within horizon	0.619	0.408
H–M rank within survey path	0.653	0.286

Notes: These falsification checks apply predefined transformations of the reported ECB sample, not alternative model searches. Horizon ranks remove unconditional horizon levels; survey-path ranks remove between-survey score levels; the human-free score asks whether challenger movement away from persistence alone predicts professional losses. The raw H–M leave-one-survey AUC range is [0.636, 0.699]. The human-free null and no-change comparison do not support unique challenger information. The two-way bootstrap interval around the raw H–M AUC still includes 0.5, so the independent-unit result remains limited.

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 H–M 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. Tables IA.18 and IA.19 report the main and robustness results.

Table IA.18: Bank of England MaPS: Policy-Rate Forecast Application

	Professional	Policy-history challenger	No change
MAE (basis points)	32.97	56.05	78.45
Ranking score	H–M gap	H–no-change gap	Difference
ROC-AUC	0.777	0.795	-0.017
Two-way 95% interval	[0.621, 0.916]	—	[-0.129, 0.087]
PR-AUC	0.573	0.519	—
Large-loss capture at 25% attention allocation	0.562	0.578	—
Retained professional MAE (bp)	17.27	17.06	—

Notes: The professional object is the MaPS median modal Bank Rate after explicitly dated MPC meetings. The challenger is a no-tuning Ridge using only lagged official Bank Rate history. The realized sample contains 261 cells across 37 surveys and 36 target meetings. A large loss is at least 50bp. Intervals resample survey paths and target meetings. The result is descriptive and does not establish machine-specific information or causal identification.

Table IA.19: Robustness Checks for the BoE MaPS Application

Ranking score	ROC-AUC	PR-AUC
Raw professional–challenger gap	0.777	0.573
Raw professional–no-change gap	0.795	0.519
Human-free challenger–no-change gap	0.560	0.261
H–M rank within horizon	0.721	0.434
No-change rank within horizon	0.731	0.454

Notes: These falsification checks apply the reported MaPS design without modification. Within-horizon ranks remove unconditional horizon levels; the human-free score removes the professional forecast from the score. Leave-one-survey AUCs range from 0.759 to 0.833. Adoption requires a challenger with higher realized-point MAE, a two-way H–M interval above 0.5, and a positive two-way H–M-minus-no-change interval.

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 H–DAR disagreement does not rank large professional losses and no change performs better. This adverse result limits transportability beyond policy-rate forecasts. Tables IA.20 and IA.21 report the main and robustness results.

Table IA.20: Philadelphia Fed SPF: Different-Object Forecast Application

	Professional	Direct-AR challenger	No change
MAE (annualized percentage points)	1.30	1.53	1.44
Ranking score	H–DAR gap	H–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% attention allocation	0.247	0.455	—
Retained professional MAE	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.

Table IA.21: Robustness Checks for the Philadelphia SPF Application

Ranking score	ROC-AUC	PR-AUC
Raw professional–DAR gap	0.457	0.093
Raw professional–no-change gap	0.584	0.191
Human-free DAR–no-change gap	0.540	0.152
H–DAR rank within horizon	0.454	0.091
No-change rank within horizon	0.580	0.186
H–DAR excluding 2020Q2–Q3	0.468	0.087

Notes: These falsification checks apply the reported PGDP design without modification. Within-horizon ranks remove unconditional horizon levels. The human-free row asks whether movement of the benchmark away from persistence predicts professional loss without using the professional forecast in the score. The pandemic exclusion was declared before results. Leave-one-origin AUCs range from 0.442 to 0.468. The pattern is treated as recurring only if the challenger has higher realized-point MAE, the two-way H–DAR AUC interval excludes 0.5, and H–DAR improves on no-change under the same two-way inference.

Part D. Market Expectations and FOMC Repricing

The analyses in this part describe market-path geometry and subsequent repricing. They do not identify the causal effect of additional professional scrutiny or an institutional welfare gain.

D1. Market-Path Construction and Intervening Information

Federal-funds-futures generic ranks are mapped to delivery-month contracts. Every pre-survey path uses the latest available settlement strictly before the survey deadline. The full human–market panel contains 547 H1–H7 cells over 83 surveys; intersection with the machine panel produces 388 cells over 60 surveys. Futures prices are risk-neutral and may contain risk and liquidity premia, so they are not converted into physical probability distributions.

The reported sensitivities include the raw path, four linear maturity corrections, a recursive horizon-mean forecast-error correction, an asymmetric NASDAQ-state correction, and two rolling-endpoint corrections using real-time payroll growth or financial spreads. Every estimated correction uses only errors whose target FOMC precedes the test survey. All members are reported rather than choosing a premium sign or maturity pattern from MP2.

MP2 is the announcement-window federal-funds-futures revision in the target rate expected at the next scheduled FOMC meeting. Controls include policy-cycle indicators, lead time, Primary Dealer disagreement, professional distance from no change, and pre-survey two-year-rate volatility. Release-window measures use official CPI, employment, PCE, GDP, and GDPNow information. Public-provider innovations are standardized with forecast errors available before each release; missing innovations are not replaced by zero. Additional market-absorption measures summarize H2–H7 futures-path movements on release dates and from the survey cutoff to the last pre-FOMC settlement. These downstream measures are sensitivity controls rather than baseline causal variables.

D2. Futures Displacement and FOMC Repricing

Preannouncement information can predict monetary-policy surprises (Cotton, 2025), while federal-funds futures embed risk premia as well as expected policy (Piazzesi and Swanson, 2008). Announcement-window price changes can also combine policy news with information about the economic outlook (Bauer and Swanson, 2023). These features make futures displacement economically interesting but prevent it from serving as a structural measure of professional-forecast quality. Table IA.22 reports the association between pre-survey midpoint displacement and the magnitude of the next-meeting MP2 revision constructed from the event study of Acosta et al. (2025). The event-bootstrap and whole-path randomization results are shown alongside the related-instrument, policy-cycle, and independent-outcome evidence.

Table IA.22: Futures Displacement and FOMC Repricing

Measure	Events	Coefficient (bp/SD)	95% bootstrap CI	Boot. p	Path-perm. p	Adj. R^2
Midpoint displacement	43	2.36^{***}	[1.14, 3.25]	0.004	< 0.001	0.455

Notes: The event sample contains 43 complete H2–H7 paths; the outcome is absolute MP2 in basis points. Pre-survey paths use only information available by the survey deadline. The semi-standardized coefficient standardizes the predictor while leaving MP2 in basis points; the fully standardized beta is 0.787, and the raw slope is 0.054 basis points of absolute MP2 per one-basis-point displacement (0.54 per 10 basis points). The event bootstrap uses 49,999 paired resamples and the 99,999-draw randomization reassigns complete pre-survey paths within policy cycle. Survey clustering is not the relevant unit for this event regression. Predictor and outcome use related federal-funds-futures contracts, only four broad policy cycles contribute, and the independent market outcomes do not support a common mechanism. The estimate is a descriptive predictive association rather than a causal effect. *, **, and *** denote statistical significance at the 10, 5, and 1 percent levels, respectively; p-values are based on the two-sided event bootstrap.

D3. Professional-Loss Risk and FOMC Repricing

Table IA.23 reports the bridge analysis, which asks whether recursive ex ante professional-loss risk predicts subsequent MP2. It is a sequential predictive test, not causal mediation. The coefficients are not distinguishable from zero across the specified family, so professional-loss ranking and market displacement remain separate findings.

Table IA.23: Professional-Loss Risk and FOMC Repricing

Bridge test	N	Coefficient (bp/SD)	95% bootstrap CI	Holm p	Path-perm. p
risk50 full	26	-0.95	[-2.84, 1.99]	1.000	0.365
risk50 full conditional on human only	26	2.57	[-1.21, 10.92]	0.635	0.013
expected loss full	26	0.11	[-1.58, 2.73]	1.000	0.881
expected loss full conditional on human only	26	-1.73	[-7.95, 10.38]	1.000	0.022

Notes: Recursive risk and expected-loss scores use only labels available before each survey and aggregate complete H2–H7 paths. The family contains four fixed tests on 26 events. None is supported jointly by the bootstrap, multiplicity, permutation, and leave-one-cycle comparisons; the estimates do not support a disagreement-to-fragility-to-repricing chain.

D4. Options, Treasury, and Equity Outcomes

Twelve months is the primary options maturity. The complete uncertainty table reports the six 12-month tests with both the primary-maturity and all-maturity Holm adjustments, followed by the additional 6-, 18-, and 24-month results. The all-maturity family crosses two outcomes, four maturities, and three predictors. The Treasury, equity, and options comparisons provide independent-outcome falsifications; no row survives multiplicity adjustment. Tables IA.24 and IA.25 report the estimates, and Figure IA.1 summarizes the options results.

Table IA.24: Options-Based Monetary-Policy Uncertainty Falsifications

Panel A. Primary 12-month maturity

Predictor	N	Coefficient	95% bootstrap CI	Boot. p	Primary-6 Holm p	All-24 Holm p
Pre-survey uncertainty						
TailDepth	22	-0.021	[-0.164, 0.079]	0.6361	1.000	1.0000
Raw nonconformity	22	0.055	[-0.040, 0.136]	0.2316	0.903	1.0000
Midpoint displacement	22	0.027	[-0.085, 0.090]	0.5113	1.000	1.0000
FOMC uncertainty resolution						
TailDepth	22	0.010	[-0.007, 0.023]	0.2233	0.903	1.0000
Raw nonconformity	22	0.014	[-0.005, 0.031]	0.1124	0.674	1.0000
Midpoint displacement	22	0.015	[-0.009, 0.033]	0.1807	0.903	1.0000

Panel B. Additional maturities

Outcome / maturity / predictor	N	Coefficient	95% bootstrap CI	Boot. p	All-24 Holm p
Pre-survey uncertainty / 6m / TailDepth	22	0.091	[-0.054, 0.176]	0.1882	1.0000
Pre-survey uncertainty / 6m / Raw nonconformity	22	0.094	[-0.021, 0.170]	0.0912	1.0000
Pre-survey uncertainty / 6m / Midpoint displacement	22	0.099	[-0.028, 0.177]	0.1097	1.0000
FOMC resolution / 6m / TailDepth	22	0.014	[-0.005, 0.028]	0.1416	1.0000
FOMC resolution / 6m / Raw nonconformity	22	0.007	[-0.009, 0.022]	0.3179	1.0000
FOMC resolution / 6m / Midpoint displacement	22	0.009	[-0.013, 0.025]	0.3275	1.0000
Pre-survey uncertainty / 18m / TailDepth	22	-0.064	[-0.203, 0.039]	0.2124	1.0000
Pre-survey uncertainty / 18m / Raw nonconformity	22	0.032	[-0.071, 0.115]	0.4889	1.0000
Pre-survey uncertainty / 18m / Midpoint displacement	22	-0.003	[-0.133, 0.081]	0.8918	1.0000
FOMC resolution / 18m / TailDepth	22	0.004	[-0.012, 0.014]	0.6234	1.0000
FOMC resolution / 18m / Raw nonconformity	22	0.009	[-0.009, 0.025]	0.1915	1.0000
FOMC resolution / 18m / Midpoint displacement	22	0.009	[-0.010, 0.025]	0.2668	1.0000
Pre-survey uncertainty / 24m / TailDepth	22	-0.083	[-0.226, 0.020]	0.1115	1.0000
Pre-survey uncertainty / 24m / Raw nonconformity	22	0.019	[-0.094, 0.111]	0.7134	1.0000
Pre-survey uncertainty / 24m / Midpoint displacement	22	-0.018	[-0.162, 0.084]	0.6355	1.0000
FOMC resolution / 24m / TailDepth	22	0.003	[-0.009, 0.014]	0.5769	1.0000
FOMC resolution / 24m / Raw nonconformity	22	0.013	[-0.003, 0.028]	0.1049	1.0000
FOMC resolution / 24m / Midpoint displacement	22	0.009	[-0.004, 0.024]	0.2021	1.0000

Notes: The official SF Fed daily series is aligned without imputation to all 22 events. Survey-date uncertainty uses the last observation strictly before the survey cutoff. FOMC resolution equals the FOMC-eve value minus the FOMC-day close. Coefficients are MPMU index units per predictor standard deviation. Panel A reports the primary 12-month maturity; Primary-6 Holm p-values adjust across its six tests, and All-24 Holm p-values adjust across the full family of two outcomes, four maturities, and three predictors. Panel B reports the remaining maturities. No row survives either applicable multiplicity adjustment.

Table IA.25: Independent Treasury and equity outcome falsifications

Predictor	N	Coefficient	95% bootstrap CI	Boot. p	9-test Holm p
Panel A. Absolute 2-year Treasury event change (bp)					
TailDepth	22	-0.22	[-2.21, 2.38]	0.733	1.000
Raw nonconformity	22	1.05	[-1.36, 3.59]	0.464	1.000
Midpoint displacement	22	1.20	[-2.32, 4.04]	0.562	1.000
Panel B. Absolute 5-year Treasury event change (bp)					
TailDepth	22	-0.74	[-2.95, 2.22]	0.472	1.000
Raw nonconformity	22	0.23	[-2.28, 2.87]	0.944	1.000
Midpoint displacement	22	0.68	[-3.09, 3.57]	0.773	1.000
Panel C. Absolute S&P 500 event return (percentage points)					
TailDepth	22	-0.04	[-0.25, 0.29]	0.651	1.000
Raw nonconformity	22	0.08	[-0.34, 0.43]	0.692	1.000
Midpoint displacement	22	0.00	[-0.58, 0.40]	0.929	1.000

Notes: The fixed family crosses TailDepth, raw nonconformity, and midpoint displacement with absolute two- and five-year Treasury changes and the absolute S&P 500 event return. The final column is Holm-adjusted across exactly these nine tests. No row survives.

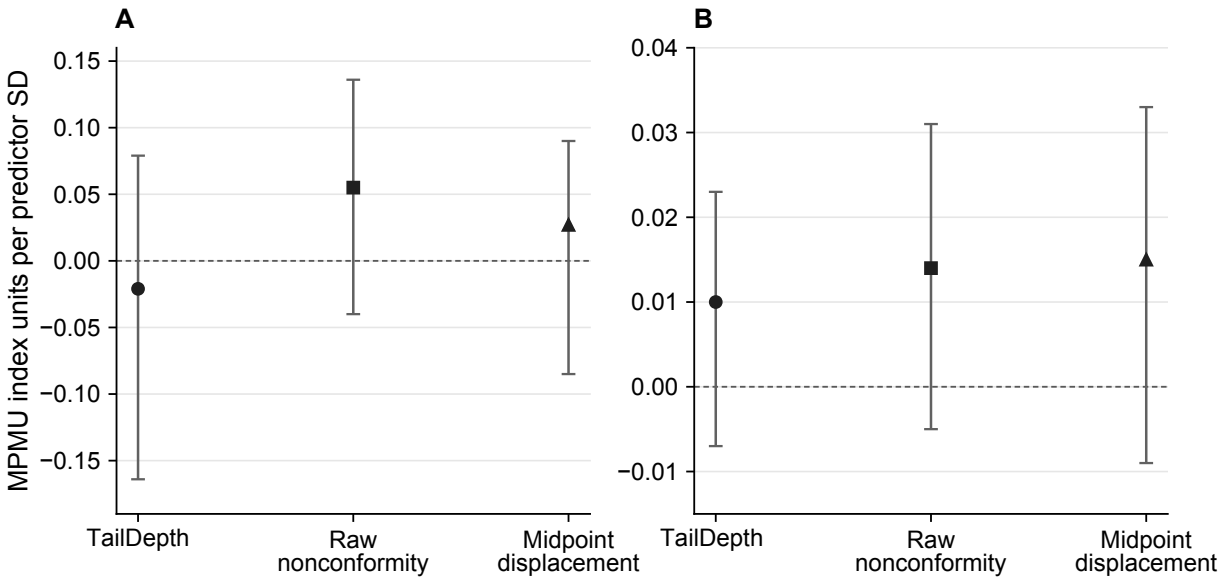


Figure IA.1: Options-Based Monetary-Policy Uncertainty around FOMC Announcements

Notes: Panel A reports coefficients for the pre-survey 12-month MPMU level; Panel B reports coefficients for the subsequent FOMC-related change. Points correspond to TailDepth, raw nonconformity, and midpoint displacement. Vertical bars are event-bootstrap 95% confidence intervals; every displayed interval includes zero.

D5. TailDepth and Forecast-Center Sensitivity

Table IA.26 reports the center comparison; Figures IA.2, IA.3, and IA.4 summarize the same-sample coefficients, finite-event influence, and center sensitivity.

Table IA.26: TailDepth as a same-sample calibration sensitivity

Specification	N	Coefficient (bp/SD)	95% bootstrap CI	Boot. p	Path p
Panel A. Same-sample calibration comparison					
Midpoint displacement	22	2.05^{*}	[-0.24, 3.11]	0.076	< 0.001
TailDepth	22	1.66^{**}	[0.27, 2.83]	0.016	0.007
TailDepth + midpoint	22	0.84	[-0.25, 1.91]	0.104	0.402
Panel B. Core recursively rebuilt center sets					
Participant only	22	1.01	[-0.54, 2.20]	0.299	0.072
Participant + Dealer	22	1.06	[-0.54, 2.26]	0.248	0.045
Participant + machine consensus	22	1.76^{***}	[0.41, 2.96]	0.008	0.004
Dealer + machine consensus	22	1.67^{**}	[0.33, 2.78]	0.012	0.003
Participant + Dealer + machine consensus	22	1.66^{**}	[0.27, 2.83]	0.015	0.006

Notes: All rows use the same 22 events. Panel A shows that TailDepth is positive alone but does not add reliably to midpoint displacement. Panel B rebuilds every recursive source scale and calibration distribution. The Participant-plus-Dealer interval crosses zero; the three-center result is therefore center-dependent and is not interpreted as a stable latent forecast set. *, **, and *** denote statistical significance at the 10, 5, and 1 percent levels, respectively; p-values are based on the two-sided event bootstrap.

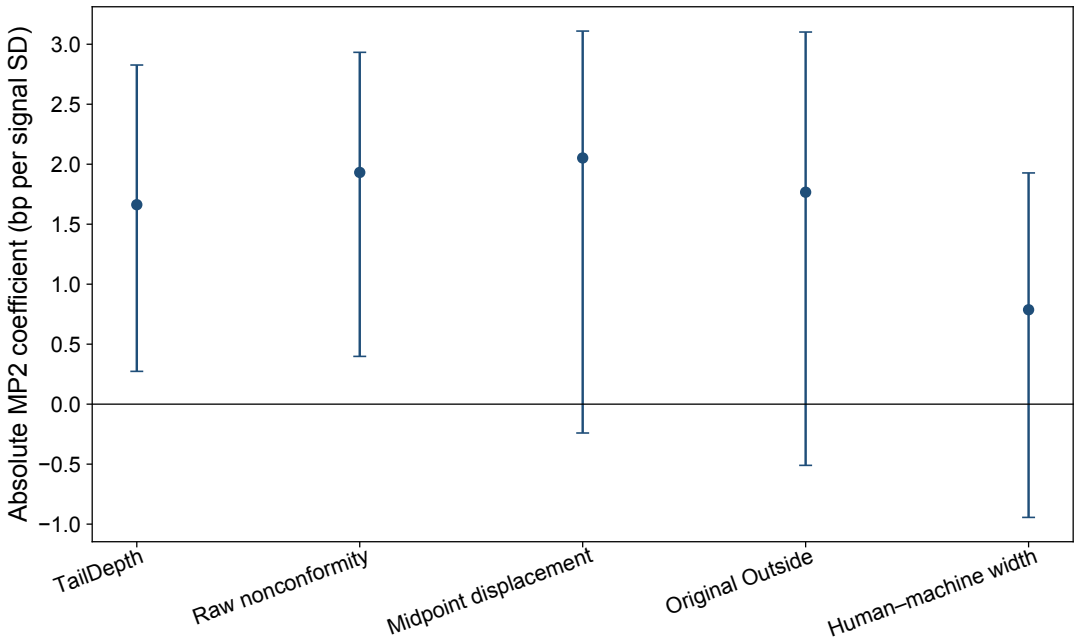


Figure IA.2: Same-Sample Calibration and Market-Geometry Coefficients

Notes: Points are coefficients from regressions of absolute MP2 on TailDepth, raw nonconformity, midpoint displacement, original outside status, and human–machine width in the common event sample. Vertical bars are event-bootstrap 95% confidence intervals.

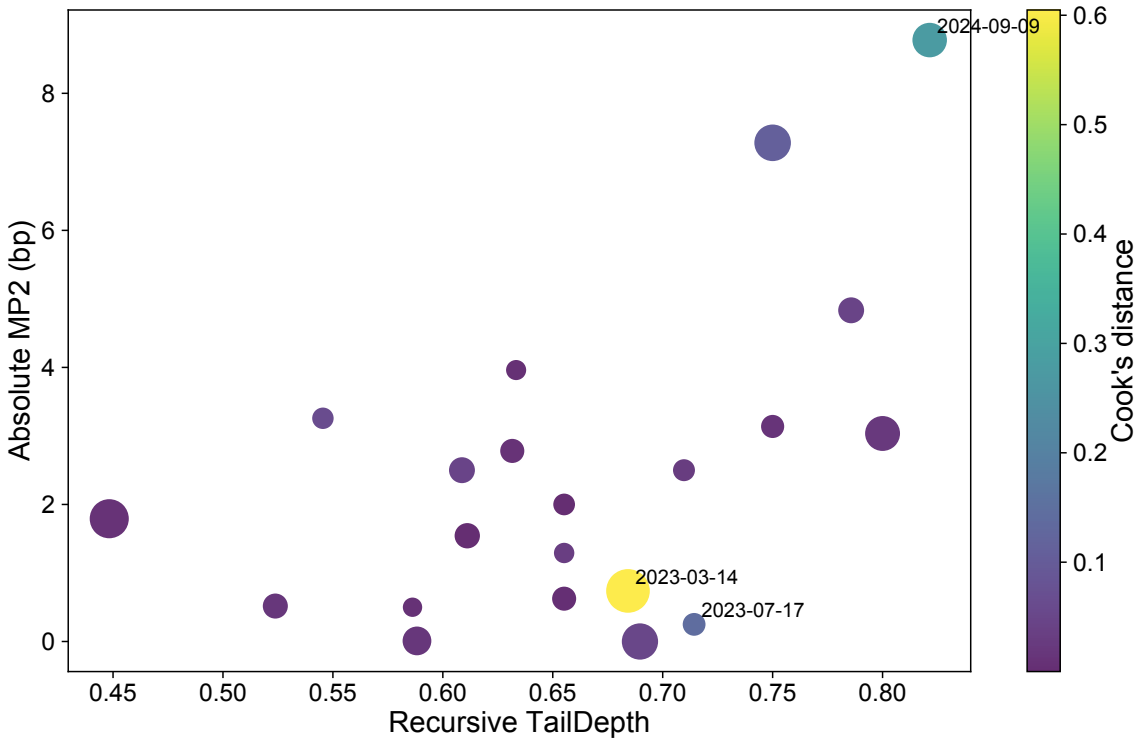


Figure IA.3: TailDepth and Finite-Event Influence

Notes: The scatterplot relates recursive TailDepth to absolute MP2 across 22 events. Point size and color represent Cook's distance. The September 2024 observation has the highest TailDepth and absolute MP2, while several labeled 2023 observations have greater influence.

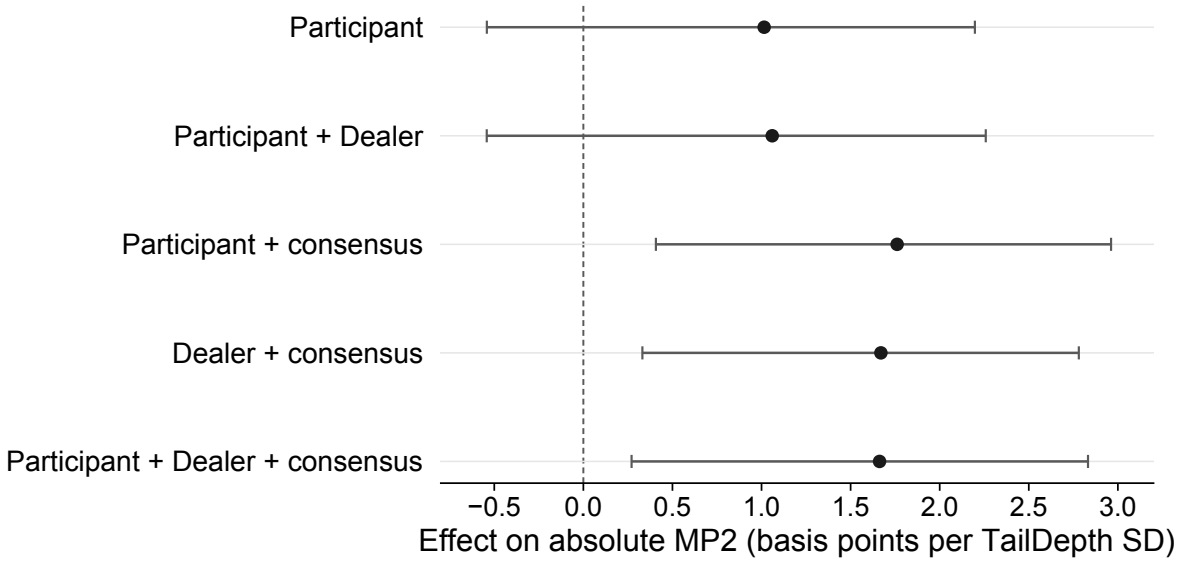


Figure IA.4: TailDepth Estimates across Forecast-Center Definitions

Notes: Dots report coefficients from regressions of absolute MP2 on TailDepth in the common 22-event sample; horizontal lines report event-bootstrap 95% confidence intervals. Each row reconstructs TailDepth using the forecast centers named at left. Consensus denotes the four-model machine consensus.

D6. Full Forecast-Center Ablation Grid

The complete 21-specification grid is retained as sensitivity evidence. No model or center is selected using MP2; five economically interpretable core sets are highlighted within the figure. Figure IA.5 reports the complete grid.

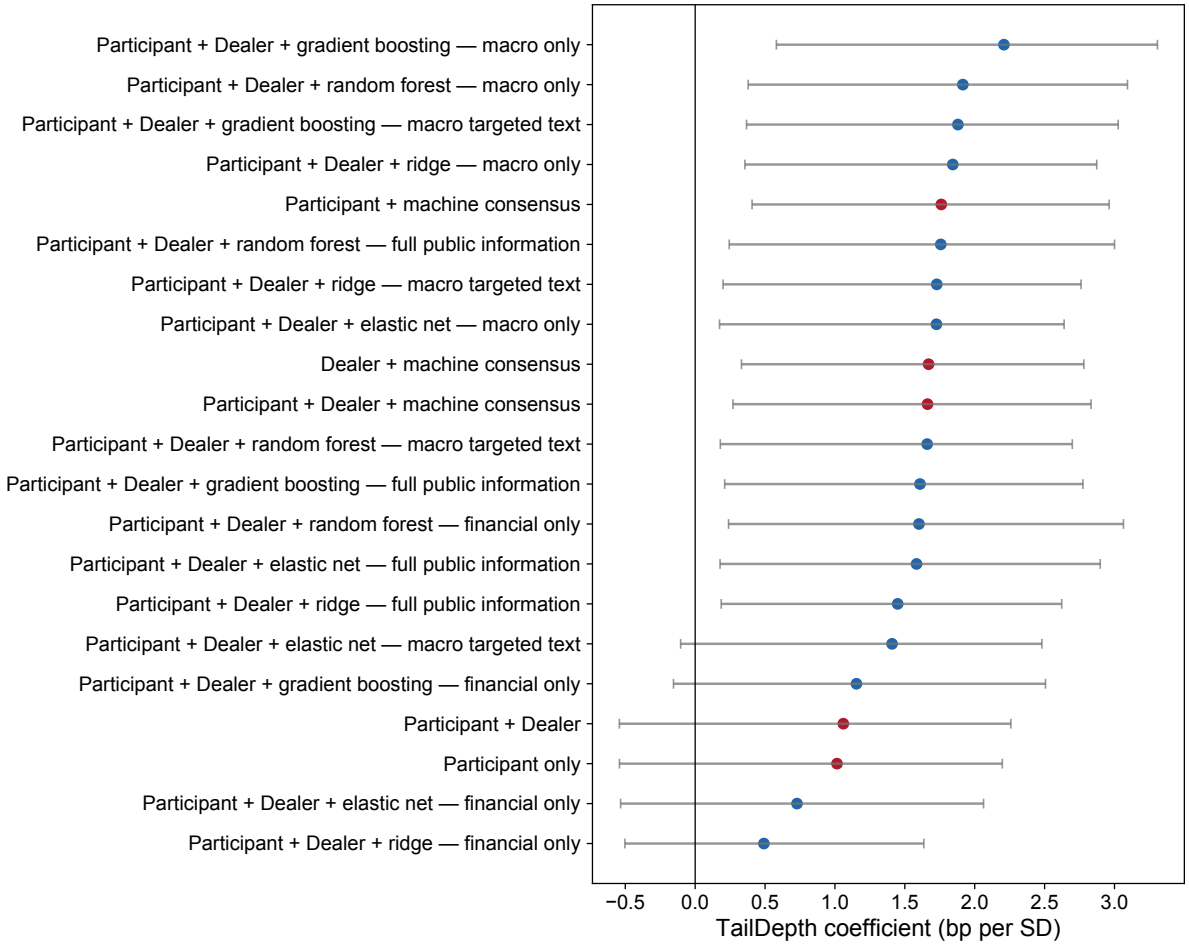


Figure IA.5: Full recursive forecast-center ablation grid.

Notes: Each row rebuilds source scales, realized-score calibration, and TailDepth on the same 22 events. The full grid is reported for completeness and is not used to select a preferred center set.