FORECASTBENCH: UPDATED RANKING METHODOLOGY

Simas Kucinskas
Forecasting Research Institute
simas@forecastingresearch.org

Houtan Bastani
Forecasting Research Institute
houtan@forecastingresearch.org

Ezra Karger
Forecasting Research Institute
Federal Reserve Bank of Chicago
ezra@forecastingresearch.org

ABSTRACT

This supplemental technical report documents important methodological updates to ForecastBench introduced after the publication of the original paper (Karger et al., 2025). While ForecastBench initially ranked forecasters using standard Brier scores, this approach fails when forecasters predict on non-overlapping questions—a common scenario as new models and question sets are continuously released. To solve this problem, we present the *difficulty-adjusted Brier score*, which decomposes forecasting performance into forecaster ability and question difficulty. Through extensive simulations, we find that the difficulty-adjusted Brier score produces rankings with high correlation to ground truth and outperforms alternative methods including standard Brier score, Brier Skill Score, and Peer Score. The up-to-date leaderboard employing this methodology can be accessed at www.forecastbench.org.

1 INTRODUCTION

ForecastBench is a comprehensive, dynamically-updated benchmark designed to evaluate the forecasting capabilities of large language models (LLMs) and compare them with human forecasters, including superforecasters (Karger et al., 2025). This technical report documents several important methodological updates to ForecastBench that were introduced after the publication of the original paper.

Originally, ForecastBench ranked forecasters using standard Brier scores. This approach worked well initially when all forecasters—both LLMs and human superforecasters—produced forecasts for identical question sets. However, as new language models are released, it becomes impossible to have all models predict on the same questions. For instance, OpenAI's o3-mini, released on January 31, 2025, could not have participated in the July 21, 2024 forecasting round. However, meaningful benchmarking requires comparing its performance to models like Anthropic's Claude 3.5 Sonnet that did participate in earlier rounds, as well as to human superforecasters who, due to cost constraints, only participate in select rounds.

The naive solution—calculating Brier scores only on questions each forecaster actually answered—unfortunately fails in practice. Question difficulty can vary substantially across forecasting rounds. As a result, a model might achieve an excellent Brier score simply by participating in an unusually easy round. Without accounting for these difficulty differences, the leaderboard rankings would reflect question difficulty rather than true forecasting ability.

To address this challenge, we developed and implemented the *difficulty-adjusted Brier score*. This ranking method decomposes forecasting performance into forecaster ability and question difficulty

Correspondence to forecastbench@forecastingresearch.org.

The views expressed in this paper do not necessarily reflect the views of the Federal Reserve Bank of Chicago or the Federal Reserve system.

components. By estimating and adjusting for question difficulty, we can fairly compare forecasters even when they have predicted on entirely different question sets, as long as we have intermediate models that create sufficient indirect overlap. This methodological innovation is essential for maintaining ForecastBench as a dynamic, continuously-updated benchmark.

The remainder of this report provides the technical details of this updated methodology. Section 2 presents the mathematical formulation of the difficulty-adjusted Brier score. Section 3 describes our simulation framework for validating the approach. Section 4 presents simulation results, demonstrating the superior performance of difficulty-adjusted Brier score over alternative ranking methods. Finally, Section 5 reviews practical implementation details.

This report should be read as a supplement to the original ForecastBench paper (Karger et al., 2025), documenting the key methodological changes implemented to enable continuous benchmarking in practice.

2 Difficulty-Adjusted Brier Score

2.1 Problem: Imperfect Question Overlap

The standard Brier score (Brier, 1950) for a binary question is given by

$$b_{i,j} = (f_{i,j} - o_j)^2,$$

where $f_{i,j} \in [0,1]$ is the probabilistic forecast made by forecaster i on question $j \in \mathcal{Q}$, $o_j \in \{0,1\}$ is the observed outcome, and $\mathcal{Q}$ denotes the full set of questions in the forecasting tournament. We use $\mathcal{Q}_i$ to denote the set of questions answered by forecaster i.

If all forecasters answer the same questions (i.e., $\mathcal{Q}_i = \mathcal{Q}$ for all i), we can reliably rank forecasters by their Brier score. However, when forecasters answer different questions, direct comparisons of average Brier scores are misleading. We call this scenario *imperfect question overlap*.

To see why this is true, assume that we can decompose the Brier score using a standard two-way decomposition:

$$b_{i,j} = \alpha_i + \gamma_j + \varepsilon_{i,j}. \tag{1}$$

Here, γ_j is a question fixed effect, α_i is a forecaster fixed effect, and $\varepsilon_{i,j}$ is an i.i.d. error term with mean zero and finite variance. Higher values of question fixed effects, γ_j, correspond to more difficult questions, while lower values of α_i correspond to more skilled forecasters.

Consider two forecasters, i and k. A quick calculation shows that

$$\bar{b}_i - \bar{b}_k = (\alpha_i - \alpha_k) + \underbrace{\left(\frac{1}{|\mathcal{Q}_i|} \sum_{j \in \mathcal{Q}_i} \gamma_j - \frac{1}{|\mathcal{Q}_k|} \sum_{j \in \mathcal{Q}_k} \gamma_j \right)}_{\text{question-difficulty term}} + (\bar{\varepsilon}_i - \bar{\varepsilon}_k),$$

where bars denote sample averages, e.g., $\bar{b}_i = 1/|\mathcal{Q}_i| \sum_{j \in \mathcal{Q}_i} b_{i,j}$. For a large sample of questions ($|\mathcal{Q}_i|, |\mathcal{Q}_k| \to \infty$), $(\bar{\varepsilon}_i - \bar{\varepsilon}_k) \xrightarrow{p} 0$ by the Law of Large Numbers. Hence, if all forecasters answer the same questions ($\mathcal{Q}_i = \mathcal{Q}_k$), in large samples, $\bar{b}_i > \bar{b}_k$ if and only if $\alpha_i > \alpha_k$.

However, with imperfect question overlap ($\mathcal{Q}_i \neq \mathcal{Q}_k$), the difference $\bar{b}_i - \bar{b}_k$ is confounded by the average difference in question difficulty. For example, if forecaster i is more accurate ($\alpha_i < \alpha_k$) but faces more difficult questions on average (i.e., the question-difficulty term is positive), ranking by the standard Brier score may inaccurately suggest that i is less accurate than k.

2.2 Solution: Difficulty-Adjusted Brier Score

To address the problem of imperfect question overlap, we develop a *difficulty-adjusted Brier score*.

Our key idea is to first estimate the question-difficulty fixed effects, γ_j, in the decomposition in Eq. (1). Then, we subtract the estimated question-difficulty effects from the standard Brier scores. After this adjustment, we can perform apples-to-apples comparisons between models with imperfect question overlap.

In general, there are various ways to estimate the question-difficulty fixed effects, γ_j. (We discuss our empirical choices in Section 2.3.) Suppose that we have some estimates, denoted by $\hat{\gamma}_j$. We calculate the (unscaled) difficulty-adjusted Brier score as

$$\tilde{b}_{i,j}^{\text{adj.}} = b_{i,j} - \hat{\gamma}_j.$$

For easier comparability with the standard Brier score, we rescale the difficulty-adjusted Brier scores so that always predicting 0.50 yields a difficulty-adjusted Brier score of 0.25, as it does for the standard Brier score. In particular, let $i = 0$ denote a model that always predicts 0.5, i.e., $f_{0,j} = 1/2$ for all j. Then, the average (unscaled) difficulty-adjusted Brier score for the "Always 0.5" forecaster is equal to

$$\bar{\tilde{b}}_0^{\text{adj.}} = \frac{1}{|\mathcal{Q}|} \sum_{j \in \mathcal{Q}} \tilde{b}_{0,j}^{\text{adj.}}.$$

In the final step, we rescale the scores to

$$b_{i,j}^{\text{adj.}} = \tilde{b}_{i,j}^{\text{adj.}} + \left(\frac{1}{4} - \bar{\tilde{b}}_0^{\text{adj.}} \right). \tag{2}$$

2.3 Estimating Question-Difficulty Fixed Effects

To estimate the question-difficulty fixed effects in practice, we use the following approach.

For *dataset questions* (see Karger et al., 2025, for details), we estimate the question-difficulty fixed effects using ordinary least squares (OLS).

For *market questions*, we use a weighted approach. Specifically, let $\hat{\gamma}_j^{\text{OLS}}$ denote the least-squares estimate, and let

$$b_{\text{mkt},j} = (f_{\text{mkt},j} - o_j)^2$$

denote the standard Brier score of the market forecast. Our weighted estimator is a weighted average of the two quantities:

$$\hat{\gamma}_j = w_{\text{mkt}} b_{\text{mkt},j} + (1 - w_{\text{mkt}}) \hat{\gamma}_j^{\text{OLS}},$$

where $w_{\text{mkt}} \in [0, 1]$ is the weight placed on the market.

The rationale for the weighted approach for market questions is as follows. OLS estimates from small forecaster samples are noisy, while market forecasts are generally highly accurate (Arrow et al., 2008). However, the market Brier score measures what is difficult for the market, not necessarily for the forecasters we rank. A question easy for the market but hard for most models will have its difficulty underestimated by $b_{\text{mkt},j}$ but correctly estimated by $\hat{\gamma}_j^{\text{OLS}}$. Intermediate values of w_{mkt} may optimally trade off the higher variance of OLS estimates against the potential bias of market-based estimates.

In practice, based on simulations in Section 4, we choose $w_{\text{mkt}} = 1$ for the public ForecastBench-leaderboard. This sacrifices 0.01–0.04 correlation points in some scenarios but ensures forecasters can only rank above the market by outperforming it in head-to-head comparisons, providing interpretability and robustness.

2.4 Related Literature

The proposed approach to ranking forecasters appears to be novel in the forecasting literature. Our approach is inspired by ideas in item response theory (Cai et al., 2016). In particular, Bo et al. (2017) use a model based on item response theory to estimate forecaster skill along with question features (question difficulty and discrimination). However, their model requires discretizing the underlying forecasts and employing Bayesian statistics, including setting priors. In contrast, our method can be applied to unadjusted forecasts and only requires running a single linear regression. The difficulty-adjusted Brier score is also closely linked to the Peer Score used by Metaculus, a leading online prediction platform (Metaculus, 2025). The Peer Score, in effect, uses the average score on a question to estimate question difficulty. However, since the pool of forecasters varies from question to question, such estimates are unlikely to be unbiased measures of question difficulty.

3 SIMULATION FRAMEWORK

We develop a simulation framework to evaluate different ranking methods under realistic forecasting conditions.

A central challenge in evaluating ranking methodologies is that we cannot observe the "true" ranking when forecasters answer different subsets of questions. Our simulation approach addresses this challenge by starting with a complete dataset where all forecasters answer all questions, establishing a *ground-truth ranking*, and then generating realistic incomplete samples to test how well different methods recover this truth.

3.1 DATA

For the simulation framework, we use data from the July 2024 ForecastBench round. The dataset contains forecasts for both dataset and market questions. This dataset includes 141 forecasters (comprising both human forecasters and large language models) and 473 resolved binary questions. Crucially, each forecaster provided a prediction for every question. Perfect question overlap is crucial, as it allows us to calculate a ground-truth ranking.

The dataset contains two types of questions. *Dataset questions* ($n = 422$) are created from a number of datasets (ACLED (Raleigh et al., 2023), DBnomics, FRED, Wikipedia, and Yahoo! Finance) according to a pre-specified question template. *Market questions* ($n = 51$) come from Manifold, Metaculus, Polymarket, and the Rand Forecasting Initiative.

3.2 SIMULATION PROCEDURE

Our simulation framework operates as follows.

3.2.1 STEP 1: CALCULATE GROUND TRUTH

Using the complete dataset, we calculate the average Brier score for each forecaster i as

$$\bar{b}_i = \frac{1}{|\mathcal{Q}|} \sum_{j \in \mathcal{Q}} (f_{i,j} - o_j)^2,$$

where $\mathcal{Q}$ denotes the full set of 473 resolved questions from the July 2024 ForecastBench round. We rank forecasters by their average Brier scores, $\bar{b}_i$, (lower scores are better) to establish the ground truth ranking $\mathcal{R}^*$.

3.2.2 STEP 2: GENERATE SIMULATED DATA

Next, we generate datasets with imperfect question overlap that capture the key practical challenges in the ForecastBench forecasting tournament. We employ two primary sampling mechanisms:

- *Random Sampling.* Each forecaster i (except a designated reference forecaster) answers n_i questions sampled uniformly at random with replacement from $\mathcal{Q}$. The reference forecaster answers all questions to enable calculation of the Brier Skill Score (see below). In our simulations, we use the *Naive Forecaster* as the reference model. The *Naive Forecaster* uses the market forecast for market questions, and the Prophet model (Taylor and Letham, 2018) for dataset questions.
- *Round-Based Sampling.* We simulate T tournament rounds. In each round t, a subset of questions $\mathcal{Q}_t$ and a subset of forecasters $\mathcal{F}_t$ are selected. All forecasters in $\mathcal{F}_t$ answer all questions in $\mathcal{Q}_t$, mimicking the design of ForecastBench. Similarly to *Random Sampling*, the reference model (*Naive Forecaster*) forecasts on all questions.

To model realistic tournament dynamics, we introduce several extensions to the round-based sampling setup:

- *Model Drift.* We simulate improvement in forecasting ability over time using a skill temperature parameter τ_t. When sampling forecasters for round t, we use softmax weights

$$w_i = \frac{\exp(\tau_t s_i)}{\sum_k \exp(\tau_t s_k)},$$

where $s_i = -\bar{b}_i$ represents forecaster i's true skill (estimated using the full sample of questions, $\mathcal{Q}$). Higher τ_t values bias selection toward better-performing forecasters. Model drift is important to include to capture the fact that large language models are continually improving over time.

- *Question Difficulty Variation.* We model systematic differences in question difficulty across rounds using a difficulty temperature parameter β_t. Questions are selected with softmax weights based on their empirical difficulty (average Brier score across all forecasters in the full dataset).
- *Forecaster Persistence.* We model realistic tournament participation by ensuring that a fraction $\rho \in [0, 1]$ of forecasters from round t continues to round $t + 1$, with the remainder replaced by new participants.

The simulation code is modular, allowing one to mix and match various elements of the data-generating process.

3.2.3 Step 3: Calculate Simulated Ranking

For each simulated dataset, we apply the ranking methods described below to produce rankings $\mathcal{R}_1^{(s)}, \mathcal{R}_2^{(s)}, \ldots, \mathcal{R}_M^{(s)}$ for the M different ranking methods, where s indexes different simulations.

In particular, we evaluate the following ranking approaches:

- *Standard Brier Score.* The baseline approach calculates the average Brier score using only the questions each forecaster answered:
$$\bar{b}_i^{\text{obs}} = \frac{1}{|\mathcal{Q}_i|} \sum_{j \in \mathcal{Q}_i} (f_{i,j} - o_j)^2,$$
where Q_i denotes the set of questions answered by forecaster i. Forecasters with lower Brier scores are ranked higher.
- *Brier Skill Score (BSS).* Next, we compute BSS (see, e.g. Bradley et al., 2008) relative to a reference forecaster that provides forecasts to every question. We consider two variants of BSS. First, we compute
$$\text{BSS}_{i,j}^{\text{pct}} = 1 - \frac{b_{i,j}}{b_{\text{ref},j}}$$
$$\text{BSS}_{i,j}^{\text{abs}} = b_{\text{ref},j}^{\text{obs}} - b_{i,j}^{\text{obs}}$$
at the forecaster-question level, where $b_{\text{ref},j}$ is the Brier score of the reference forecaster on question j. Then, we average across all questions to the forecaster-level to obtain $\bar{\text{BSS}}_i^{\text{pct}} = 1/|\mathcal{Q}_i| \sum_{j \in \mathcal{Q}_i} \text{BSS}_{i,j}^{\text{pct}}$ and $\bar{\text{BSS}}_i^{\text{abs}} = 1/|\mathcal{Q}_i| \sum_{j \in \mathcal{Q}_i} \text{BSS}_{i,j}^{\text{abs}}$. Forecasters with greater BSS values are ranked higher.
- *Peer Score.* Following the approach used by Metaculus (Metaculus, 2025), we calculate the peer score of forecaster i on question j as
$$\text{PS}_{i,j} = \bar{b}_j - b_{i,j},$$
where $\bar{b}_j$ is the average Brier score across all forecasters who answered question j. The forecaster's peer score, $\overline{\text{PS}}_i$, is the average of $\text{PS}_{i,j}$ across all questions they answered. Forecasters with greater peer scores are ranked higher.
- *Difficulty-Adjusted Brier Score.* As described in Section 2. Forecasters with lower difficulty-adjusted Brier scores are ranked higher.

3.2.4 Step 4: Measure Ranking Quality

Given the calculated simulated rankings, we estimate their quality using three main metrics:

- *Spearman Correlation.* We compute the Spearman's rank correlation coefficient between the true ranking $\mathcal{R}^*$ and each simulated ranking. In particular, for each simulation s, we calculate the rank correlation between $\mathcal{R}^*$ (the true ranking) and $\mathcal{R}_i^{(s)}$ (the simulated ranking for the i-th ranking method); let the resulting correlation be denoted by $r_i^{(s)}$. Then, we take the average across all simulations s.

- *Top-k Retention.* For $k \in \{20, 50\}$, we calculate the fraction of forecasters ranked in the top k positions in $\mathcal{R}^*$ who remain in the top k positions in the simulated ranking $\mathcal{R}_i^{(s)}$; denote the resulting retention rate by $\text{ret}_i^{(s)}$. Then, we take the average of $\text{ret}_i^{(s)}$ across all simulations s.
- *Median Rank Displacement.* We compute the median absolute difference between each forecaster's rank in $\mathcal{R}^*$ and their rank in the simulated ranking $\mathcal{R}_i^{(s)}$. Specifically, for each model j, denote its rank in a given ranking $\mathcal{R}$ by $r_j^{\mathcal{R}}$. Then, the difference in ranks is given by $\Delta_{j,i}^{(s)} = |r_j^{\mathcal{R}^*} - r_j^{\mathcal{R}_i^{(s)}}|$. We first take the median of $\Delta_{j,i}^{(s)}$ across all models j to get the median displacement for simulation s and method i, $\Delta_{\text{med},i}^{(s)}$, and then average $\Delta_{\text{med},i}^{(s)}$ across all simulations s.

4 Simulation Results

Table 1: Simulation Results. *Spearman*: Spearman rank correlation coefficient relative to the true ranking. *Top-20*: Top-20 retention rate relative to the true ranking. *Top-50*: Top-50 retention rate relative to the true ranking. *Med. Disp.*: Median displacement relative to the true ranking, in ranks. See Section 3 for additional details.

	Spearman	Top-20	Top-50	Med. Disp.
Random sampling				
Standard Brier	0.90	0.67	0.81	9
BSS (Pct.)	-0.05	0.14	0.32	45
BSS (Abs.)	0.88	0.65	0.79	11
Peer Score	0.94	0.76	0.87	7
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.00$)	0.94	0.76	0.87	7
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.25$)	0.94	0.76	0.86	7
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.50$)	0.93	0.74	0.85	8
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.75$)	0.92	0.72	0.84	9
Diff.-Adj. Brier ($w_{\text{mkt}} = 1.00$)	0.90	0.69	0.82	10
Round-based sampling				
Standard Brier	0.95	0.59	0.62	26
BSS (Pct.)	-0.11	0.12	0.29	44
BSS (Abs.)	0.94	0.58	0.62	26
Peer Score	0.96	0.59	0.62	26
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.00$)	0.97	0.61	0.62	26
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.25$)	0.97	0.61	0.62	26
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.50$)	0.96	0.60	0.62	26
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.75$)	0.96	0.60	0.62	26
Diff.-Adj. Brier ($w_{\text{mkt}} = 1.00$)	0.95	0.59	0.62	26
Round-based sampling with drift				
Standard Brier	0.64	0.36	0.50	36
BSS (Pct.)	-0.13	0.12	0.27	46
BSS (Abs.)	0.78	0.43	0.54	33
Peer Score	0.81	0.42	0.56	32
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.00$)	0.91	0.53	0.59	30
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.25$)	0.91	0.53	0.59	30
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.50$)	0.91	0.53	0.59	30
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.75$)	0.91	0.53	0.59	30
Diff.-Adj. Brier ($w_{\text{mkt}} = 1.00$)	0.91	0.53	0.59	30

We evaluate ranking quality across multiple simulation scenarios designed to test different aspects of the forecasting tournament environment. Tables 1 and 2 report results for seven distinct scenarios.

4.1 SCENARIO DESIGN

Our simulation scenarios vary along three key dimensions. First, we consider different sampling mechanisms: random sampling (each forecaster answers 500 randomly drawn questions) and round-based sampling (10 rounds with 500 questions per round, and persistence $\rho = 0.70$). Second, we vary the number of forecasters per round to examine sample size effects, testing configurations with 10 (small), 30 (baseline), and 50 (high) forecasters per round. Third, we introduce realistic tournament dynamics including model drift (calibrated so that the average Brier scores improve by 0.06 points from first to final round) and question difficulty variation (calibrated so that question difficulty varies by 0.09 Brier points between even and odd rounds).

Additionally, we examine scenarios in which market and forecaster difficulty measures diverge. For each question j, we calculate the *forecaster-market discrepancy* as $|b_{\text{mkt},j} - \bar{b}_j|$ where $\bar{b}_j$ is the average Brier score across all forecasters. We then construct scenarios sampling only from questions with above-median (high discrepancy) or below-median (low discrepancy) forecaster-market discrepancy values.

4.2 BASELINE SCENARIOS

In the random sampling scenario, difficulty-adjusted Brier score with $w_{\text{mkt}} = 0$ achieves a Spearman correlation of 0.94, matching the Peer Score and outperforming standard Brier (0.90) and both BSS variants (-0.05 and 0.88). The percentage-based BSS performs particularly poorly, achieving a negative correlation, while the absolute BSS is close to standard Brier score performance.

For round-based sampling without drift, all methods except percentage-based BSS perform similarly, with Spearman correlations ranging from 0.94 to 0.97. The difficulty-adjusted Brier score with $w_{\text{mkt}} = 0$ (i.e., no market adjustment, a pure two-way fixed effects estimation) achieves the highest correlation (0.97) in this scenario.

4.3 REALISTIC TOURNAMENT DYNAMICS

The most relevant scenario involves round-based sampling with model and question drift. This scenario mimics a realistic tournament environment in which model quality improves over time and question difficulty varies substantially across rounds. Here, the difficulty-adjusted Brier score substantially outperforms alternatives. Regardless of the market weight w_{mkt}, difficulty-adjusted Brier achieves a Spearman correlation of 0.91, compared to 0.81 for Peer Score, 0.78 for absolute BSS, and 0.64 for standard Brier. The top-20 retention rate is 0.53 for difficulty-adjusted Brier versus 0.42 for Peer Score and 0.36 for standard Brier. This scenario demonstrates that methods that fail to adjust for question difficulty can produce substantially biased rankings.

4.4 ROBUSTNESS ANALYSIS

As shown in Table 2, the performance of difficulty-adjusted Brier score degrades with smaller forecaster samples but remains superior to alternatives.[1] With only 10 forecasters per round, difficulty-adjusted Brier achieves a Spearman correlation of 0.93–0.94 across all market weight values, compared to 0.94 for standard Brier and 0.91 for Peer Score. With 50 forecasters per round, difficulty-adjusted Brier with $w_{\text{mkt}} = 0$ reaches 0.98, compared to 0.97 for Peer Score and 0.96 for standard Brier. Notably, top-k retention rates deteriorate substantially with small samples, dropping from 0.75–0.82 with 50 forecasters to 0.27 with 10 forecasters.

We also examine scenarios with varying forecaster-market difficulty discrepancies. In the high-discrepancy scenario, difficulty-adjusted Brier with $w_{\text{mkt}} = 0$ achieves a Spearman correlation of 0.94, decreasing to 0.90 with $w_{\text{mkt}} = 1$, as expected given the mismatch between market and forecaster difficulty. However, even in the high-discrepancy scenario, the degradation in performance with $w_{\text{mkt}} = 1$ is only 0.04 correlation points (0.94 to 0.90). In the low discrepancy scenario, all market weight specifications perform almost identically at 0.97, suggesting that OLS and market-

[1] Robustness scenarios exclude model/question drift to isolate the effect of sample size or forecaster-market discrepancy.

Table 2: Simulation Results: Additional Scenarios. *Spearman*: Spearman rank correlation coefficient relative to the true ranking. *Top-20*: Top-20 retention rate relative to the true ranking. *Top-50*: Top-50 retention rate relative to the true ranking. *Med. Disp.*: Median displacement relative to the true ranking, in ranks. See Section 3 for additional details.

	Spearman	Top-20	Top-50	Med. Disp.
Round-based sampling, low models				
Standard Brier	0.94	0.27	0.27	51
BSS (Pct.)	-0.09	0.11	0.27	52
BSS (Abs.)	0.93	0.27	0.27	51
Peer Score	0.91	0.27	0.27	51
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.00$)	0.94	0.27	0.27	51
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.25$)	0.94	0.27	0.27	51
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.50$)	0.94	0.27	0.27	51
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.75$)	0.94	0.27	0.27	51
Diff.-Adj. Brier ($w_{\text{mkt}} = 1.00$)	0.93	0.27	0.27	51
Round-based sampling, high models				
Standard Brier	0.96	0.74	0.81	12
BSS (Pct.)	-0.13	0.12	0.29	46
BSS (Abs.)	0.96	0.73	0.80	13
Peer Score	0.97	0.76	0.82	12
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.00$)	0.98	0.78	0.82	12
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.25$)	0.98	0.77	0.82	12
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.50$)	0.97	0.77	0.82	12
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.75$)	0.97	0.76	0.81	12
Diff.-Adj. Brier ($w_{\text{mkt}} = 1.00$)	0.96	0.75	0.81	12
Round-based sampling, high discrepancy				
Standard Brier	0.93	0.57	0.63	26
BSS (Pct.)	-0.04	0.19	0.33	41
BSS (Abs.)	0.90	0.56	0.62	26
Peer Score	0.93	0.58	0.63	26
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.00$)	0.94	0.58	0.63	25
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.25$)	0.94	0.58	0.63	25
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.50$)	0.93	0.58	0.63	26
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.75$)	0.92	0.57	0.63	26
Diff.-Adj. Brier ($w_{\text{mkt}} = 1.00$)	0.90	0.56	0.62	26
Round-based sampling, low discrepancy				
Standard Brier	0.93	0.58	0.61	28
BSS (Pct.)	0.65	0.40	0.51	33
BSS (Abs.)	0.96	0.61	0.62	27
Peer Score	0.95	0.60	0.62	27
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.00$)	0.97	0.62	0.62	27
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.25$)	0.97	0.62	0.63	27
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.50$)	0.97	0.62	0.63	27
Diff.-Adj. Brier ($w_{\text{mkt}} = 0.75$)	0.97	0.62	0.63	27
Diff.-Adj. Brier ($w_{\text{mkt}} = 1.00$)	0.97	0.62	0.63	27

based difficulty estimates converge when they measure similar constructs. Peer Score achieves 0.93 and 0.95 in the high and low discrepancy scenarios, respectively.

5 Practical Implementation Details

5.1 Excluding Stale Models from Difficulty Estimation

The difficulty-adjusted Brier score decomposes Brier scores into forecaster ability and question difficulty. For dataset questions, we estimate question difficulty using a two-way fixed effects model, as described in Section 2. A key assumption underlying this approach is that forecaster quality remains stable over time.[2]

However, LLM forecasting performance is expected to degrade as questions extend beyond the model's training cutoff date. This temporal decay biases difficulty estimates. A question long past a model's training cutoff may appear difficult not because it is inherently challenging, but because the model's knowledge has become stale. Such contamination would bias our difficulty estimates.

To address this issue, we only include models within 1 year of their training cutoff date when estimating question fixed effects via OLS. After this threshold, we continue to evaluate models in the benchmark but exclude them from the difficulty estimation procedure.

5.2 New Model Inclusion Timing: 50-Day Waiting Period

A key operational challenge for ForecastBench is determining when to add newly released models to the public leaderboard. Adding models too early risks displaying unreliable performance metrics based on a small number of resolved questions. Conversely, waiting too long reduces the benchmark's timeliness and relevance.

To address this challenge, we conducted a *stability analysis* to identify the optimal waiting period before including new models in the leaderboard. Our analysis examines how model rankings stabilize over time as more questions resolve. Based on this analysis, we use a 50-day waiting period for adding new models to the leaderboard.

5.2.1 Methodology

To establish ground-truth rankings, we first restrict the sample to forecasters that had been active for at least 180 days. This choice is made to ensure a reliable ground-truth ranking. For each forecaster, we compute the difficulty-adjusted Brier score using questions resolved within various time windows (e.g., 0–30 days, 0–50 days) from the forecaster's first prediction, re-estimating the $\hat{\gamma}'_j s$ within each window.

We then compare these early rankings to the final rankings based on all resolved questions. To evaluate how well the early rankings match the final rankings, we use the same metrics as in the simulation study (see Section 3). The only addition is the *score correlation*: Pearson's correlation between difficulty-adjusted Brier scores computed from the early versus final datasets.

5.2.2 Results

Figure 1 shows the stability results for dataset questions (Baseline leaderboard).[3] Since dataset questions resolve at fixed forecast horizons (in particular, at 7, 30, 90, and 180 days out), the graphs exhibit discontinuities at these corresponding dates. Rankings stabilize rapidly for dataset questions. Within 7 days, score and rank correlations exceed 0.90, while the top-25% retention rate is higher than 80%. Median rank displacement drops from 3 positions initially to 2 positions at 30 days and 1 position by day 90. This rapid stabilization occurs because dataset questions resolve at fixed horizons, with 250 dataset questions resolving at each forecasting horizon.[4]

[2] This assumption is implicit in Eq. (1), which does not include a time index for the forecaster fixed effect α_i.

[3] The Baseline leaderboard excludes models with market values in their prompts and any future external submissions; the Tournament leaderboard includes all models. See the benchmark website for more details.

[4] See Karger et al. (2025) for the sampling details. We have 250 questions per horizon because we filter out so-called combination questions. Note that the correlations and top-25% retention rate do not reach exactly 1.0 at day 180 because some forecasters have more than 180 days of resolved questions. The ground-truth ranking incorporates all available data (e.g., up to day 360 for some forecasters), while the day-180 ranking uses only questions resolved within the first 180 days. For the same reason, median rank displacement does not converge to exactly zero.

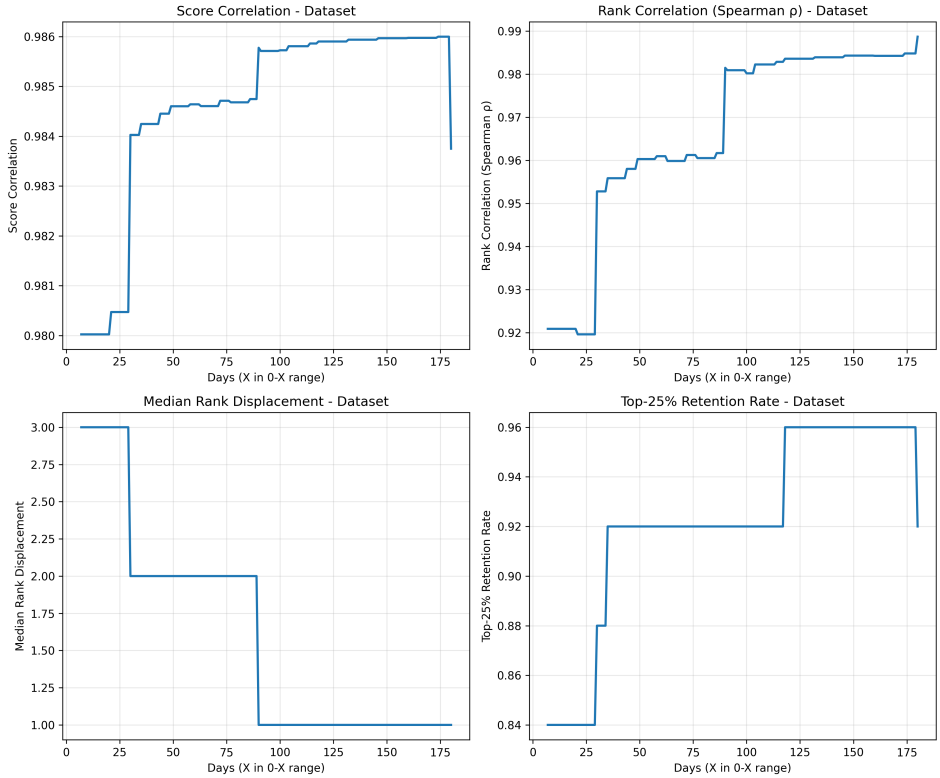

Figure 1: Stability results: Dataset questions, Baseline leaderboard.

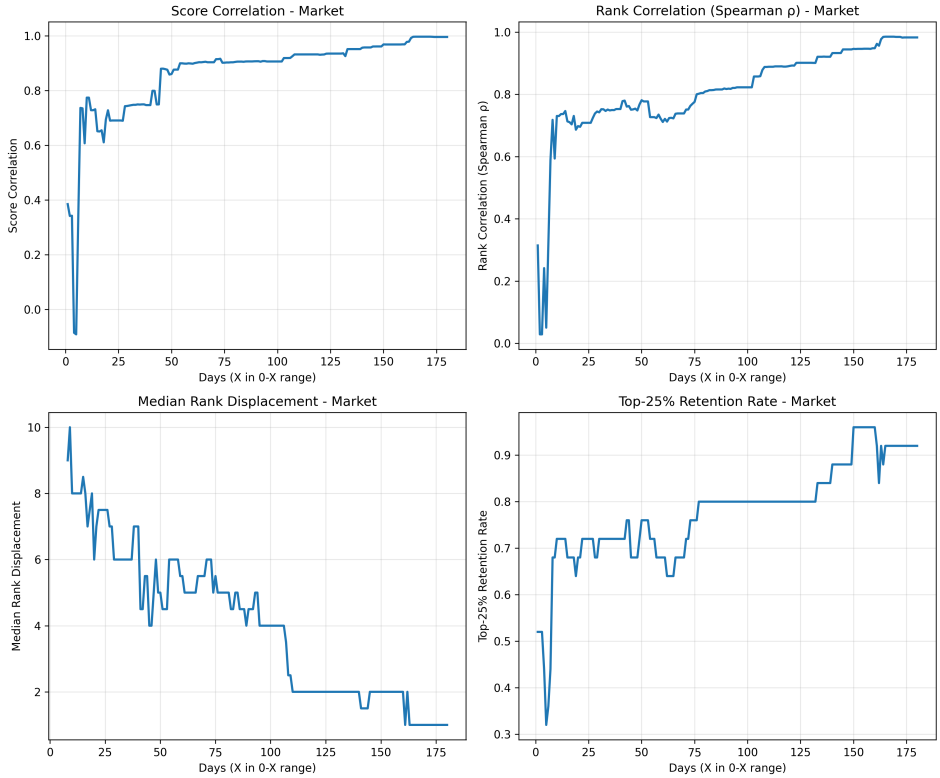

Figure 2: Stability results: Market questions, Baseline leaderboard.

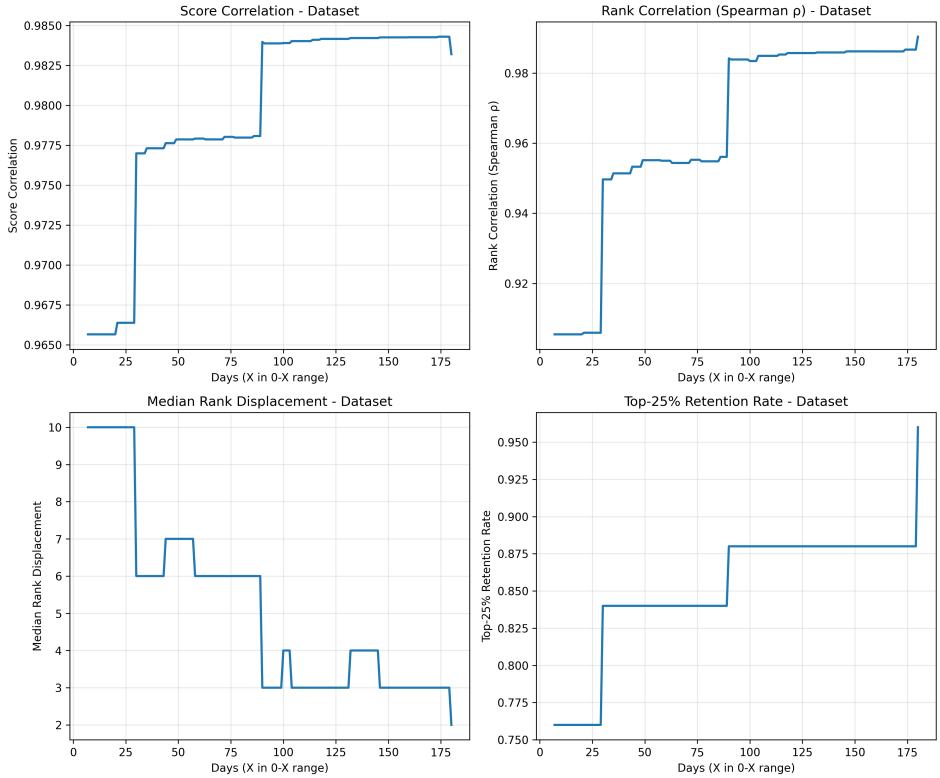

Figure 3: Stability results: Dataset questions, Tournament leaderboard.

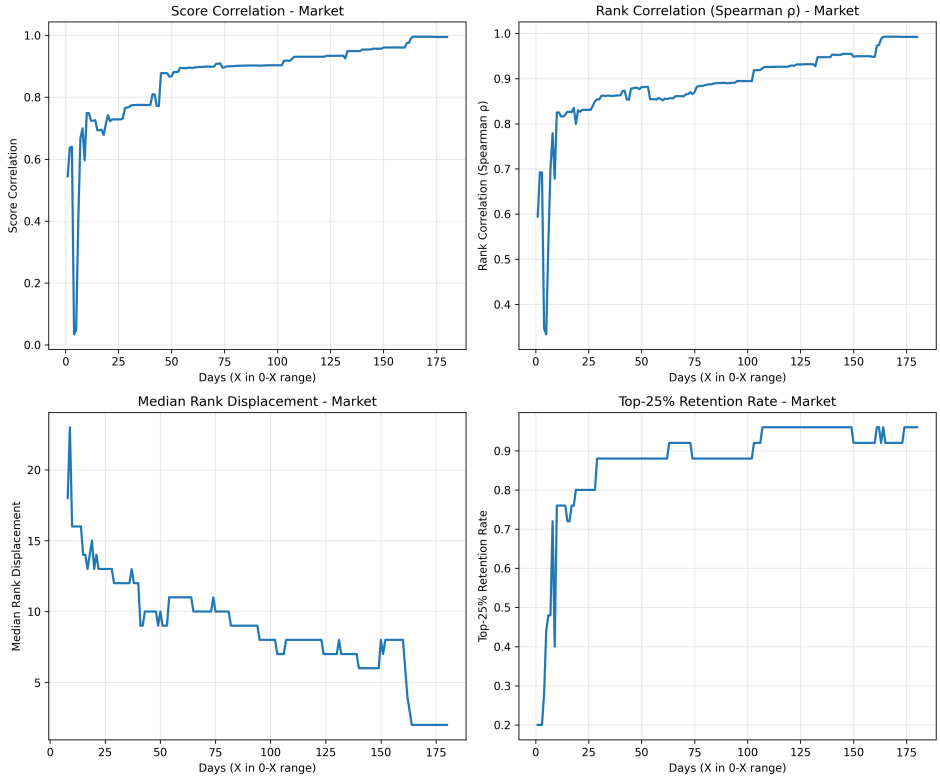

Figure 4: Stability results: Market questions, Tournament leaderboard.

Figure 2 shows the stability results for market questions, which present a markedly different pattern. Market questions exhibit substantially lower initial stability due to their smaller sample size and variable resolution times. Score and rank correlations increase rapidly until approximately 50 days and then plateau at 0.8–0.9. The top-25% retention rate follows a similar trajectory, reaching around 70% after 50 days. Figures 3 and 4 show that the results are very similar for the Tournament leaderboard.[5]

Based on this analysis, we implemented a 50-day waiting period before adding new models to the ForecastBench leaderboard. This threshold balances two competing objectives: ensuring reliable performance metrics while maintaining benchmark timeliness. By day 50, market question rankings achieve sufficient stability (correlations of 0.8–0.9), while dataset question rankings have already stabilized (correlations exceeding 0.95).

5.3 Other Changes

For a list of other changes, please consult the changelog at www.github.com/forecastingresearch/forecastbench/wiki/Changelog.

6 Limitations

Estimating question-difficulty fixed effects relies on assumptions that may not always hold in practice. In particular, the two-way fixed effects decomposition in Eq.(1) assumes forecaster ability is constant across domains—for example, a forecaster's relative performance on geopolitical questions should be identical to their relative performance on financial questions. In reality, forecasters may have heterogeneous skill profiles, excelling in certain domains while underperforming in others. This factor structure is absorbed into the error term in the current model, potentially biasing difficulty estimates. Future extensions could include forecaster-by-domain fixed effects or hierarchical models that cluster questions by topic.

Our choice to use market-based difficulty adjustment ($w_{\text{mkt}} = 1$) prioritizes interpretability and robustness. Simulation results show this specification can underperform pure OLS adjustment ($w_{\text{mkt}} = 0$) when forecaster and market difficulty measures diverge substantially. The market-based approach ensures forecasters can only rank above the market by directly outperforming it, providing a clear benchmark. However, this comes at the cost of potential bias when markets and forecasters find different questions difficult.

We apply a uniform 50-day waiting period before adding new models to the leaderboard, despite dataset questions stabilizing much faster. This choice simplifies communication but introduces a delay. To address this, we display preliminary dataset question rankings for models between 7–50 days old on the benchmark website.

[5]The only major difference is that the median rank displacement metric increases. This change is mechanical: The Tournament leaderboard has roughly twice as many models as the Baseline leaderboard, and hence for the same relative stability, absolute stability numbers become worse.

REFERENCES

Kenneth J. Arrow, Robert Forsythe, Michael Gorham, Robert Hahn, Robin Hanson, John O. Ledyard, Saul Levmore, Robert Litan, Paul Milgrom, Forrest D. Nelson, et al. The promise of prediction markets, 2008.

Yuanchao Emily Bo, David V. Budescu, Charles Lewis, Philip E. Tetlock, and Barbara Mellers. An IRT forecasting model: Linking proper scoring rules to item response theory. *Judgment and Decision Making*, 12(2):90–103, 2017.

A. Allen Bradley, Stuart S. Schwartz, and Tempei Hashino. Sampling uncertainty and confidence intervals for the brier score and brier skill score. *Weather and Forecasting*, 23(5):992–1006, 2008.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Li Cai, Kilchan Choi, Mark Hansen, and Lauren Harrell. Item response theory. *Annual Review of Statistics and Its Application*, 3(1):297–321, 2016.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. ForecastBench: A dynamic benchmark of ai forecasting capabilities. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://iclr.cc/virtual/2025/poster/28507.

Metaculus. Scores FAQ, 2025. URL https://www.metaculus.com/help/scores-faq/. Accessed: 2025.

Clionadh Raleigh, Roudabeh Kishi, and Andrew Linke. Political instability patterns are obscured by conflict dataset scope conditions, sources, and coding choices. *Humanities and Social Sciences Communications*, 10:74, 2023.

Sean J. Taylor and Benjamin Letham. Forecasting at scale. *The American Statistician*, 72(1):37–45, 2018. doi: 10.1080/00031305.2017.1380080.