Proxy Scoring Enables Benchmarking LLM Forecasters Without Waiting for Outcomes

Anonymous Authors[1]

Abstract

Large language models are increasingly used as general-purpose forecasters, but benchmarking their forecasting ability remains slow and fragile. Retrospective benchmarks risk training-data contamination, while prospective benchmarks require waiting weeks or months for questions to resolve. We propose an instantaneous evaluation method based on proxy scoring rules. Rather than scoring each forecast against the eventual outcome, we score it against an extremized aggregate of the forecasts made by other models, drawing on the literature on information elicitation without verification. Empirically these proxy scores correlate strongly with resolved-outcome metrics such as the Brier score on existing LLM forecasting data and are almost as predictive as Brier scores of future performance and substantially less noisy across time. We further show the effectiveness of the method crucially depends on the aggregation method used: simple means and medians can perform poorly, while logit-mean aggregation followed by extremization yields consistently strong correlations.

1. Introduction

Forecasting has a long history in statistics and machine learning. Mathematical models have been used for decades to predict quantities such as the weather or retail sales, typically by fitting a model to historical data and evaluating it on a held-out set to assess generalization and detect overfitting.

More recently, large language models (LLMs) and agent systems built with them have broadened the scope of automated forecasting. Rather than being restricted to domains with rich historical data, LLMs are increasingly applied to arbitrary, one-off questions about the future, for example, which candidate will win an upcoming election, or whether a disease outbreak will occur in a particular region by a given date. Forecasts would then take a form like "Candidate A is 73% likely to win".

1.1. LLM Forecasting Benchmarks

More than a dozen benchmarks have been proposed for evaluating the forecasting ability of LLMs, and they broadly fall into three categories.

The first is *pastcasting*: questions are scraped or constructed about events that have already occurred, and the model's predictions are scored against the known outcomes. The difficulty is that LLMs are trained extensively on records of past events, so the answers are often already memorized and holding out test data completely would be infeasible. A common mitigation is to restrict to questions whose resolution falls after the model's training knowledge cutoff, and to provide historical web sources where additional context is needed. But this is hard to do reliably, and the resulting benchmarks can be brittle (Paleka et al.). Examples of this approach include (Halawi et al., 2024) and (Wildman et al., 2025).

The second category consists of live, prospective benchmarks: models are queried about future events before the outcomes are known, and scored once the questions resolve. This sidesteps data-contamination concerns, naturally accommodates forecasting agents that browse the live web, and mirrors how human forecasters are traditionally evaluated. The downside is that evaluation is no longer instantaneous, which is unusual for an ML benchmark and inconvenient for practitioners who want to compare models today. It also forces a focus on short-horizon questions in order to obtain results within a reasonable timeframe, even though long-horizon forecasting is often what we ultimately care about.

The third, more elusive category attempts to evaluate forecasters instantly, without waiting for outcomes. This is inherently indirect. One notable approach is that of (Fluri et al., 2024; Paleka et al., 2024), which scores forecasters by the internal consistency of their predictions: for example,

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

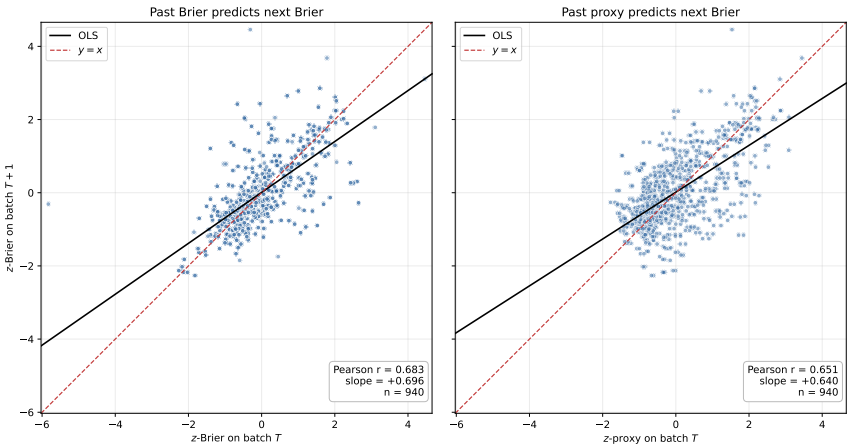

Figure 1. Comparison of how well the Brier score (left) and proxy score (right) from one batch predict the Brier score on the subsequent batch in ForecastBench. Every point represents a model tested on a particular set of questions. The fact that the regression line is shallower than the diagonal shows regression toward the mean for both measures.

$P(\text{A wins})$ and $P(\text{A does not win})$ should sum to one, and analogous identities hold for more complex logical combinations. They show that this consistency metric correlates well with real-world forecasting performance.

Our work also falls in this third category, but rather than scoring each forecaster on the internal consistency of its own predictions, we score forecasters by comparing their predictions against one another, drawing on the literature on information elicitation without verification. This cross-forecaster signal is both informative and readily available in the LLM setting, and we show empirically that it tracks ground-truth Brier rankings closely.

1.2. Proxy scoring

Before it came within reach of automated systems, this kind of very general forecasting had already been studied extensively in humans. Established methods such as the Delphi method (Dalkey & Helmer, 1963) and crowd forecasting have been used to elicit calibrated probabilities from the general public, domain experts, or "superforecasters"—people who consistently score well in forecasting tournaments (Mellers et al., 2015). In this context, there has already been empirical work on how to identify such individuals ahead of time.

Our proposed approach is based on the work of (Witkowski et al., 2017), who introduce *proxy scoring rules*: a generalization of proper scoring rules in which a forecast x is scored against a random proxy $\hat{y}$ rather than against a resolved outcome y. Their central result is that the quadratic proxy scoring rule $L(x, \hat{y}) = (x - \hat{y})^2$ is strictly proper whenever the proxy is, in expectation, equal to the true probability of the event, i.e. $\mathbb{E}[\hat{y}] = p$ where $p = \Pr(y = 1)$. As a concrete instantiation, they propose using the *extremized mean* of the forecasters' own predictions as the proxy, and show on data from a large geopolitical forecasting tournament that the resulting score ranks forecasters nearly as well as the standard quadratic score with access to true outcomes. Here extremizing means moving the estimate further from the midpoint of $1/2$ to account for information gained through aggregation and the wisdom-of-the-crowd effect (Baron et al., 2014).

It is worth noting that there is a broader literature on information elicitation without verification, which seeks to evaluate or reward forecasters for accuracy in settings without access to ground truth (Lehmann, 2026). Notable alternative methods include the Bayesian Truth Serum (Prelec, 2004) and its descendants. An attractive feature of these methods is that, under suitable assumptions, they can be shown to be incentive-compatible and to reward private information, guarantees that proxy scoring does not offer. For a benchmark, however, incentive compatibility is not essential, and these methods typically require additional elicitation questions or more elaborate mechanisms, both of which add complexity. We therefore focus on proxy scoring in this work.

2. Method

In this paper we assume a binary judgmental forecasting framework where N LLMs or agents are each individually tasked with predicting the probabilities of M heterogeneous events. We assume each participant $i = 1, \ldots, N$ outputs for each question $j = 1, \ldots, M$ a probability $x_i^{(j)} \in [0, 1]$. Each question represents an outcome in the future that may or may not occur. We can represent the resolution as $y^{(j)} = 1$ if the outcome happens, otherwise $y^{(j)} = 0$.

The traditional way of scoring the performance of a fore-

caster is with a scoring rule $L : [0,1]^2 \to \mathbb{R}$. The standard choice is the Brier score $L_{\text{Brier}}(x, y) = (x - y)^2$. A scoring rule is proper if the expected score of a forecaster is optimized by reporting their true belief, and strictly proper if this is the unique optimal strategy. The Brier score is strictly proper (Gneiting & Raftery, 2007).

When the resolution $y^{(j)}$ is not available, the proxy approach replaces it with an aggregate $\hat{y}^{(j)} = A(x_1^{(j)}, \ldots, x_N^{(j)})$ computed from the forecasters' own predictions, and scores each forecaster against $\hat{y}^{(j)}$ instead. For the Brier score, this maintains strict properness.

We test four methods of aggregating the per-question forecasts $\{x_i^{(j)}\}_{i=1}^N$ into a proxy $\hat{y}^{(j)} = A(x_1^{(j)}, \ldots, x_N^{(j)})$: the simple mean $\bar{x} = \frac{1}{N}\sum_{i=1}^N x_i$; the standard median; the extremized mean of (Witkowski et al., 2017), $\bar{x}^\alpha/(\bar{x}^\alpha + (1-\bar{x})^\alpha)$ with $\alpha = 2$; and the logit-pool extremization $\sigma\Big(d \cdot \frac{1}{N}\sum_{i=1}^N \log\frac{x_i}{1-x_i}\Big)$ where $\sigma(x) = 1/(1+e^{-x})$ from (Satopää et al., 2014) with $d = \sqrt{3}$ as suggested in (Neyman & Roughgarden, 2022), which has been shown to perform well as an aggregator in human forecasting tournaments (Sevilla, 2021).

Given an aggregator A, we define the (quadratic) proxy score of forecaster i on a batch of M questions as

$$S_{\text{proxy}}(i; A) \;=\; \frac{1}{M}\sum_{j=1}^{M}\Big(x_i^{(j)} - \hat{y}^{(j)}\Big)^2,$$

mirroring the structure of the Brier score with $\hat{y}^{(j)}$ in place of the resolution $y^{(j)}$. Lower is better. For simplicity we do not leave each forecaster out of the aggregate before calculating the score despite this breaking properness, but this effect is slight given the relatively large number of LLMs in the datasets.

3. Empirical results

3.1. Datasets

To evaluate our method, we reanalyze existing benchmark datasets in which questions have already been forecast and resolved, allowing a head-to-head comparison with live forecasting benchmarks. We primarily use ForecastBench (Karger et al., 2025), which runs regular batches in which different models forecast the same set of questions. We additionally test on the smaller dataset from (Paleka et al., 2024) (henceforth "Consistency Forecasting") to make sure the effects are not specific to one dataset. We also evaluate the same method on data from human forecasters in the multi-year Good Judgment Project (Good Judgment Project, 2016) for comparison.

For each question in each dataset, we first compute the extremized proxy. From this, we obtain a proxy score for each (question, LLM) pair, and average these within each batch of simultaneously asked questions to obtain an overall proxy score per model. Since the resolutions are also known, we additionally compute the traditional Brier score for each question and average it analogously per batch.

Brier scores are only meaningfully comparable across forecasts on the same question set. We therefore standardize: for each batch, we compute the z-score of forecasting performance for each LLM relative to the other models tested on the same batch. Because both Brier and proxy scores assign lower values to better forecasts, a lower z-score is better under either metric. Using the raw scores instead of z-scores leads to broadly similar results. Further details on the data preparation are available in Appendix A.

All data, processing scripts, and experiment code are available under the following URL will be released publicly upon acceptance: `https://anonymous.4open.science/r/ProxyScoringICMLForecasting-24C0/`

4. Experiments

4.1. Proxy–Brier correlation across batches

We first ask whether the proxy score ranks forecasters in roughly the same order as the Brier score. For every (model, batch) observation in each dataset, we z-score the Brier and proxy values within the batch and then compute the Pearson correlation across the pooled set. Table 1 reports these correlations under each of the four aggregators introduced above. The logit-pool extremization with $d = \sqrt{3}$ achieves the highest correlation on both ForecastBench ($r = 0.700$) and Consistency Forecasting ($r = 0.685$). The advantage over the simpler aggregators is small but reliable on the larger dataset and very large on the smaller one: the simple mean and median collapse to near-zero or slightly negative correlation on Consistency Forecasting, indicating that without enough extremization the aggregate is too compressed near 0.5 to discriminate forecasters at all. The extremized mean with $\alpha = 2$ sits between the two extremes, and for Brier-based ranking is decisively beaten by the logit-pool extremization.

Because logit-pool extremization performs consistently best, we use this method exclusively in the analyses that follow.

4.2. Other scoring rules

A natural concern is whether the choice of underlying scoring rule changes the picture, particularly given that both Brier and proxy scores are based on squared errors. We re-test the correlation against three further metrics: the logarithmic score $L_{\log}(x, y) = -y\ln(x) - (1-y)\ln(1-x)$ (clipping forecasts to $[\epsilon, 1-\epsilon]$ with $\epsilon = 10^{-3}$ to avoid divergences), the absolute error $L_{\text{abs}}(x, y) = |x - y|$, and the

Dataset	r(mean)	r(median)	r(extr. mean, $\alpha = 2$)	r(extremized, $d = \sqrt{3}$)
ForecastBench	$+0.555$	$+0.575$	$+0.641$	$+\mathbf{0.700}$
Consistency Forecasting	-0.086	-0.021	$+0.243$	$+\mathbf{0.685}$
GJP	$+0.756$	$+0.816$	$+0.845$	$+\mathbf{0.900}$

Table 1. Pearson correlation between each model's within-batch z-scored Brier score and its within-batch z-scored proxy score, pooled across all batches. Higher r means the proxy is a better stand-in for Brier-based ranking.

0–1 loss $L_{0\text{-}1}(x, y) = \mathbf{1}\{\mathbf{1}\{x \geq 1/2\} \neq y\}$, which simply counts how often the forecast lands on the wrong side of 0.5. Note that the latter two are not themselves proper scoring rules. The within-batch z-score Pearson correlations are reported in Table 2 in Appendix B. Logit-pool extremization remains the strongest aggregator in every row, across all datasets and all four metrics. The gap to the simpler aggregators widens sharply on the smaller Consistency Forecasting dataset, where the mean and median in fact produce negative correlations for L_{abs} and $L_{0\text{-}1}$. The logit-pool proxy still tracks both metrics, despite neither being a proper scoring rule.

4.3. Predicting next-batch Brier from past batch

Judging the quality of forecasts is always stochastic. Even subsequent batches on ForecastBench correlate at only $r = 0.683$ in within-batch z-Brier for the same model. The relevant question is therefore not whether the proxy is a perfect substitute for Brier, but whether it is competitive with Brier itself as a predictor of future performance.

Figure 1 shows this comparison on the 25 adjacent batch pairs of ForecastBench (940 model-pair observations). The left panel plots within-batch z-Brier on batch T against z-Brier on batch $T + 1$ for the same model, for models that appear in subsequent batches. The right panel plots within-batch z-proxy (extremized, $d = \sqrt{3}$) on batch T against z-Brier on batch $T + 1$. The Pearson correlations are $r = 0.683$ and $r = 0.651$ respectively.

4.4. Temporal stability of proxy vs. Brier

We restrict to ForecastBench models that appear on at least three batches (160 models) and compute, for each model, the standard deviation across batches of its within-batch z-Brier and of its within-batch z-proxy (extremized, $d = \sqrt{3}$). The mean per-model standard deviation is 0.318 for the proxy and 0.470 for Brier; the difference is highly significant (paired Wilcoxon $p = 3.5 \times 10^{-10}$). This is consistent with a prior study of proxy scores using human-generated data (Himmelstein et al., 2023).

5. Discussion and Limitations

Any attempt to assess forecast quality without access to actual outcomes is necessarily indirect, and its validity warrants scrutiny.

To illustrate, consider a hypothetical forecaster who predicts the future perfectly, always assigning probability 0 or 1 in agreement with the realized outcome. Such a forecaster would, of course, achieve a perfect Brier score of 0. Yet when we add such a forecaster to the ForecastBench dataset, they almost always rank worse than every other model under the surrogate scoring rule.

This is counterintuitive at first glance. But it follows from how the proxy is constructed. Forecast uncertainty can be decomposed into aleatoric and epistemic components, where aleatoric uncertainty reflects the inherent randomness of the world. A forecaster claiming to eliminate aleatoric uncertainty as well, by always reporting 0 or 1, will appear extremely overconfident relative to the proxy. This is arguably the right behavior: ex ante, a forecaster who reports certainty on genuinely uncertain events is indistinguishable from one who is reckless or lucky, and the proxy penalizes both. But the concern remains that an exceptionally strong forecasting model whose predictions systematically diverge from the crowd will be penalized by the proxy even when they are right. However, we did not empirically observe this in the datasets analyzed.

6. Conclusion

Proxy scoring lets practitioners rank LLM forecasters instantly, without waiting for outcomes to resolve. On ForecastBench, the proxy score from one batch predicts a model's next-batch Brier score nearly as well as the Brier score itself, and is more temporally stable across batches, suggesting it may offer a less noisy signal of underlying forecasting ability. Crucially, the choice of aggregator matters: logit-pool extremization with $d = \sqrt{3}$ consistently outperforms simpler alternatives, particularly on smaller pools of forecasters where the mean and median fail entirely.

These results have immediate practical value: new models or agent configurations can be evaluated on any set of open questions today, rather than waiting weeks or months for resolution. More broadly, our findings suggest that the literature on information elicitation without verification, which has been developed primarily for human settings, transfers well to LLMs and deserves wider attention, not only for evaluation but potentially as a source of training signal.

Impact Statement

This work aims to make evaluation of LLM forecasters faster and more reliable, including for the long-horizon questions that resolved-outcome benchmarks cannot reach in any practical timeframe. Better evaluation should in turn support more accurate probabilistic forecasts about consequential events in policy, public health, and finance, which we view as broadly beneficial.

Two caveats merit attention. First, proxy scoring rewards agreement with the aggregated crowd. This is a useful indirect signal of accuracy, but could also be misleading in individual cases particular for particularly well-performing systems that is outside the usual performance of the crowd and therefore mislead.

References

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., and Ungar, L. H. Two Reasons to Make Aggregated Probability Forecasts More Extreme. *Decision Analysis*, 11(2):133–145, 2014. ISSN 1545-8490.

Dalkey, N. and Helmer, O. An Experimental Application of the Delphi Method to the Use of Experts. *Management science*, 9(3):458–467, 1963.

Fluri, L., Paleka, D., and Tramer, F. Evaluating Superhuman Models with Consistency Checks. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pp. 194–232, Los Alamitos, CA, USA, April 2024. IEEE Computer Society. doi: 10.1109/SaTML59370.2024.00017. URL https://doi.ieeecomputersociety.org/10.1109/SaTML59370.2024.00017.

Gneiting, T. and Raftery, A. E. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Good Judgment Project. GJP Data. Technical report, Harvard Dataverse, 2016.

Halawi, D., Zhang, F., Yueh-Han, C., and Steinhardt, J. Approaching Human-Level Forecasting with Language Models. *Advances in Neural Information Processing Systems*, 37:50426–50468, 2024.

Himmelstein, M., Budescu, D. V., and Ho, E. H. The Wisdom of Many in Few: Finding Individuals Who Are as Wise as the Crowd. *Journal of Experimental Psychology: General*, 152(5):1223, 2023.

Karger, E., Bastani, H., Yueh-Han, C., Jacobs, Z., Halawi, D., Zhang, F., and Tetlock, P. E. ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://iclr.cc/virtual/2025/poster/28507.

Lehmann, N. V. Mechanisms for Belief Elicitation without Ground Truth. *Journal of Economic Surveys*, 40(1):505–527, 2026. ISSN 0950-0804.

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., and Horowitz, M. Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions. *Perspectives on Psychological Science*, 10(3):267–281, 2015.

Neyman, E. and Roughgarden, T. Are You Smarter than a Random Expert? The Robust Aggregation of Substitutable Signals. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 990–1012, 2022.

Paleka, D., Goel, S., Geiping, J., and Tramer, F. Evaluating Forecasting is More Difficult than Other LLM Evaluations. In *ICML 2025 Workshop on Assessing World Models*.

Paleka, D., Sudhir, A. P., Alvarez, A., Shen, A., and Paleka, D. Consistency Checks for Language Model Forecasters. In *Agentic Markets Workshop at ICML 2024*, 2024.

Prelec, D. A Bayesian Truth Serum for Subjective Data. *science*, 306(5695):462–466, 2004. ISSN 0036-8075.

Satopää, V. A., Baron, J., Foster, D. P., Mellers, B. A., Tetlock, P. E., and Ungar, L. H. Combining Multiple Probability Predictions Using a Simple Logit Model. *International Journal of Forecasting*, 30(2):344–356, 2014.

Sevilla, J. Principled Extremizing of Aggregated Forecasts, December 2021. URL https://forum.effectivealtruism.org/posts/biL94PKfeHmgHY6qe/principled-extremizing-of-aggregated-forecasts.

Wildman, J., Bosse, N. I., Hnyk, D., Mühlbacher, P., Hambly, F., Evans, J., Schwarz, D., and Phillips, L. Bench to the Future: A Pastcasting Benchmark for Forecasting Agents. *arXiv preprint arXiv:2506.21558*, 2025.

Witkowski, J., Atanasov, P., Ungar, L., and Krause, A. Proper Proxy Scoring Rules. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017. ISBN 2374-3468.

A. Data processing details

This appendix records the cleaning and filtering choices behind the results in Tables 1 and 2 that are not apparent from the definitions in Section 2.

The three datasets were accessed at the following URLs, current as of 2026-05-09:

- ForecastBench: https://www.forecastbench.org/datasets/
- Consistency Forecasting: https://github.com/dpaleka/consistency-forecasting/tree/main/src/data/forecasts
- Good Judgment Project: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/BPCDH5

A.1. Filtering of non-LLM forecasters

Both the ForecastBench and Consistency Forecasting datasets contain entries that are not autonomous LLM forecasters, and we exclude these from both the consensus pool and the set of forecasters being scored. Three categories are removed. The first is trivial baselines included in ForecastBench for reference, such as constant forecasters at 0, 0.5, and 1, a uniform random forecaster, and the naive and imputed forecasters. The second is human-crowd aggregates, namely the public and superforecaster median forecasts from ForecastBench. The third category is processed versions of LLMs that are themselves in the same batch. This covers any ForecastBench entry whose name contains "LLM Crowd" and the ConsistentForecaster models in Consistency Forecasting.

A.2. Cleaning of invalid forecasts

A small fraction of forecasts in each dataset are unusable, and we drop these rather than attempting to clip or repair them. Two cases arise: missing forecasts, where the model did not answer or its response could not be parsed, and forecasts outside $[0, 1]$, where a single ForecastBench entrant emits values such as 50 or values on the order of 10^6, presumably from confusing probabilities with percentages or some other scale. Drops are per-forecast, so a model with one bad forecast on one question is not removed from the analysis entirely.

The logit-pool aggregator and the log-loss metric both diverge at the endpoints of $[0, 1]$, so for these we additionally clip surviving forecasts to $[\epsilon, 1 - \epsilon]$ with $\epsilon = 10^{-3}$ before taking logs. The squared-error proxy and Brier scores use the raw probabilities.

A.3. ForecastBench

We use all batches in ForecastBench (Karger et al., 2025) for which both forecasts and resolutions were available at the time of writing, giving 26 batches between July 2024 and April 2026. We only process resolved events.

A.4. Consistency Forecasting

The Consistency Forecasting dataset (Paleka et al., 2024) has forecasts by forecaster and by split, where splits correspond either to a date range or to one of three named partitions: scraped questions, synthetic questions, and questions resolving in 2028. We use only the ground-truth resolutions provided with the dataset; the tuple-based files associated with the original consistency-checks methodology are not used here. The scraped split is not analyzed because, after removing the ConsistentForecaster ensembles (Section A.1), it contains no autonomous LLM forecasters.

B. Full per-metric correlation table

Table 2 gives the full Pearson correlations from the experiment of Section 4.2, broken out by ground-truth metric (Brier, log loss, absolute error, 0–1 loss) and proxy aggregator, for both ForecastBench and Consistency Forecasting.

Dataset	Metric	r(mean)	r(median)	r(extr. mean, $\alpha = 2$)	r(extremized, $d = \sqrt{3}$)
ForecastBench	L_{Brier}	$+0.555$	$+0.575$	$+0.641$	$\mathbf{+0.700}$
	$L_{\log}$	$+0.642$	$+0.646$	$+0.655$	$\mathbf{+0.670}$
	L_{abs}	$+0.393$	$+0.473$	$+0.608$	$\mathbf{+0.701}$
	$L_{\text{0-1}}$	$+0.334$	$+0.358$	$+0.472$	$\mathbf{+0.559}$
Consistency Forecasting	L_{Brier}	-0.086	-0.021	$+0.243$	$\mathbf{+0.685}$
	$L_{\log}$	$+0.477$	$+0.528$	$+0.698$	$\mathbf{+0.885}$
	L_{abs}	-0.493	-0.416	-0.128	$\mathbf{+0.365}$
	$L_{\text{0-1}}$	-0.228	-0.177	$+0.076$	$\mathbf{+0.541}$
GJP	L_{Brier}	$+0.756$	$+0.816$	$+0.845$	$\mathbf{+0.900}$
	$L_{\log}$	$+0.761$	$+0.772$	$+0.775$	$\mathbf{+0.803}$
	L_{abs}	$+0.570$	$+0.694$	$+0.758$	$\mathbf{+0.835}$
	$L_{\text{0-1}}$	$+0.658$	$+0.730$	$+0.766$	$\mathbf{+0.826}$

Table 2. Pearson correlation between each model's within-batch z-scored proxy score and its within-batch z-scored ground-truth loss under four different evaluation metrics, pooled across all batches. Bold marks the best aggregator in each row.

C. Leave-one-out proxy scores

Throughout the main text we include each forecaster i's own prediction in the aggregate $\hat{y}^{(j)}$ used to score that same forecaster. Conceptually, the purer alternative is the leave-one-out (LOO) variant

$$\hat{y}^{(j)}_{-i} \;=\; A\big(x^{(j)}_1, \ldots, x^{(j)}_{i-1}, x^{(j)}_{i+1}, \ldots, x^{(j)}_N\big), \qquad S^{\text{LOO}}_{\text{proxy}}(i; A) \;=\; \frac{1}{M}\sum_{j=1}^{M}\big(x^{(j)}_i - \hat{y}^{(j)}_{-i}\big)^2,$$

which removes the spurious effect of a forecaster "predicting itself" through its share of the consensus. Otherwise, the score is no longer proper.

For every (model, batch) observation in ForecastBench (1377 in total) we compute both S_{proxy} and $S^{\text{LOO}}_{\text{proxy}}$ under each of the four aggregators from Section 2. Table 3 summarizes the result.

Aggregator	mean WITH	mean LOO	mean abs. diff	mean rel. diff	max rel. diff	Spearman / Pearson
mean	0.0320	0.0333	$+0.0013$	$+4.1\%$	$+15.4\%$	$0.9996/0.9995$
median	0.0349	0.0364	$+0.0015$	$+4.9\%$	$+19.5\%$	$0.9989/0.9990$
extremized mean ($\alpha = 2$)	0.0412	0.0431	$+0.0019$	$+4.6\%$	$+17.9\%$	$0.9993/0.9991$
logit-pool ($d = \sqrt{3}$)	0.0455	0.0476	$+0.0021$	$+4.5\%$	$+17.2\%$	$0.9991/0.9988$

Table 3. Proxy scores including the focal forecaster in the aggregate (WITH) versus excluding it (LOO), pooled across (model, batch) observations in ForecastBench.

LOO is systematically larger because removing x_i from the aggregate pulls the consensus further from x_i in expectation, but the inflation is small in absolute terms (~ 0.002) and approximately uniform across forecasters. The rank correlations between the two variants are above 0.998 for every aggregator, so the relative ordering of forecasters in any single batch is unaffected, and any cross-batch statistic computed on z-scores would remain virtually unchanged.