An Adversarial Tournament Design for Efficiently Probing the Frontier of AI Forecasting

Ben Day [1] **Scott Jeen** [1] **Simion-Vlad Bogolin** [1] **Maximilian Clark** [1] **Toby Shevlane** [1]

Abstract

We show that existing forecasting tournaments are inefficient probes of the capability frontier: most questions are either trivially easy or effectively intractable, with only a small fraction in the regime where leading systems disagree. We propose an adversarial design that admits more discriminating question formats and makes question-writing itself competitive, rewarding writers in proportion to the disagreement their questions induce among the strongest forecasters. We instantiate this as a live tournament with a $25,000 prize pool and report early findings on the questions that survive adversarial selection.

1. Introduction

AI forecasting systems are improving quickly, and now perform comparably to skilled humans (Karger et al., 2025). As in any other field, progress has been made by identifying flaws with existing systems and patching them. When capabilities are weak, flaws are rampant, and identifying them is straightforward. As capabilities improve, flaws reveal themselves less readily, and more effort is required to find them.

This is particularly acute for forecasting. In most domains a wrong answer is itself evidence of a flaw, because a correct answer was available; the task is tractable by construction. Forecasting questions are not like this. A question that a system gets wrong may be hard-but-tractable, in the sense that better information or reasoning would yield a better forecast, or it may be effectively intractable, dominated by noise that no amount of skill can dissolve. The two look the same from the outside, but only the first is worth chasing. Effort spent on the second yields no improvement.

Forecasting tournaments offer a way to tell them apart. When strong systems agree on an incorrect forecast, the question was probably intractable; when strong systems *disagree*, at least one of them is leaving something on the table, and the question points at a concrete place where the frontier could move.

In this paper we argue that questions from popular forecasting tournaments now provide a weak signal on how to improve state-of-the-art systems. We demonstrate that recent questions from ForecastBench and the Metaculus AI Benchmark elicit *agreement* between strong systems, and so do not distinguish their respective strengths (Section 2). In response, we propose a new tournament that incentivises participants to write questions that elicit *disagreement* between the best systems (Section 3). We conclude with early results from a live tournament (Section 4).

2. Existing Benchmarks

2.1. ForecastBench

ForecastBench (Karger et al., 2025) asks binary questions drawn from prediction markets (`infer`, `manifold`, `metaculus`, `polymarket`) and from public time-series sources (`acled`, `dbnomics`, `fred`, `wikipedia`, `yfinance`). Every fortnight a new *question set* of $\sim$200 binary questions is released. Each participating model submits one forecast per question on the release's `forecast_due_date`; market questions resolve on a single market close, while dataset questions resolve at multiple horizons (7-day, 30-day, 90-day, ...) against the realised value of the time series. The processed forecast sets we use cover 27 `forecast_due_dates` and $\sim$410 distinct (organization, model) submitters, downloaded from the ForecastBench website.

2.2. Metaculus AI Benchmark Series

The Metaculus AI Benchmarking Tournament is a quarterly series in which participating bots forecast a mix of binary, multiple-choice, and numerical questions written and curated by Metaculus (Metaculus, 2025b). Unlike ForecastBench, the same set of bots forecasts the same set of questions, so cross-bot comparison is direct: every bot in a tournament sees every question and submits a forecast by the

[1]Mantic Technologies Ltd. Correspondence to: Ben Day <contact@mantic.com>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

Table 1. **Cross-forecaster agreement on Fall AIB 2025 and ForecastBench.** MAD and Brier are computed on the probability the panel assigns to the resolved outcome. Values are mean±std across panels.

		N	MAD	Brier
Fall AIB 2025 (by question type)				
	binary	128	0.071 ±0.070	0.078 ±0.155
	multiple choice	23	0.105 ±0.070	0.392 ±0.307
ForecastBench (by source)				
DATASET	acled	50	0.063 ±0.097	0.041 ±0.102
DATASET	fred	45	0.148 ±0.120	0.230 ±0.171
DATASET	wikipedia	50	0.041 ±0.063	0.002 ±0.004
DATASET	yfinance	49	0.034 ±0.015	0.244 ±0.017
MARKET	infer	30	0.060 ±0.052	0.007 ±0.014
MARKET	manifold	71	0.056 ±0.058	0.063 ±0.127
MARKET	metaculus	68	0.063 ±0.059	0.045 ±0.114
MARKET	polymarket	63	0.048 ±0.061	0.076 ±0.126
DATASET	all	195	0.070 ±0.093	0.126 ±0.145
MARKET	all	232	0.056 ±0.058	0.063 ±0.120

question's deadline. Multiple-choice questions present an explicit set of mutually exclusive options; numerical questions ask for a probability distribution over a bounded range. We focus on the Fall 2025 instance (Fall AIB) (Metaculus, 2025a), the most recent edition at time of writing, and report on seven of the best-performing bots.

2.3. Computing Disagreement Between Participants

For both tournaments we form panels — sets of forecasts on a common question — and compute two statistics per panel: the mean pairwise absolute deviation (MAD) on the probability the panel assigns to the resolved outcome, and the panel-mean Brier score. We report mean±std across panels in each row of Table 1.

The two tournaments differ in how panels are formed. In ForecastBench, the release cadence means that most (organization, model) submitters appear on a single due-date against a single question set, so we cannot compare across the full leaderboard directly. Instead we form panels at the level of a single question instance — the tuple `(forecast_due_date, source, question_id, resolution_date)` — and restrict to panels where at least five of the top-10 forecasters (ranked by mean Brier on resolved, non-imputed forecasts, ≥ 50 resolved each) submitted a non-imputed value. In Fall AIB the seven-bot panel is fixed by construction, and we form one panel per question.

2.4. Results

Table 1 shows that disagreement among strong forecasters is small on both tournaments. On ForecastBench, the typical pair of top-10 forecasters lies within 6 percentage points on the average question ($\text{MAD}_{\text{MARKET}} = 0.056$, $\text{MAD}_{\text{DATASET}} = 0.070$). On the cleanest sources agreement is essentially saturated — yfinance reaches MAD $= 0.034$ and Wikipedia MAD $= 0.041$ with a near-perfect panel Brier of 0.002. Prediction-market sources are uniformly tight ($\text{MAD}_{\text{MARKET}}$ between 0.048 and 0.063), so the strong overall agreement is not driven by any single source. The meaningful exceptions is the macro time-series source FRED, where both disagreement and Brier are elevated.

The same pattern shows up on Fall AIB. The seven-bot panel agrees tightly on binary questions (MAD $= 0.071$) and achieves a panel-mean Brier of 0.078, comparable to ForecastBench. Multiple-choice questions elicit somewhat more disagreement (MAD $= 0.105$), and the higher panel Brier (0.392) suggests this is partly because the questions are genuinely harder rather than because the bots have different views. Taken together with the ForecastBench numbers, the picture is that as the bots have improved they have also converged: the questions that remain are largely ones where the strongest systems already agree, and so carry little signal about how to improve any of them.

3. A New Kind of Tournament

We introduce a new kind of tournament that aims to seek out areas of disagreement among the strongest AI forecasters. This is achieved in two ways: 1) adopting more discriminating question formats (Section 3.1), and 2) scoring questions based on the disagreement they elicit between strong systems (Section 3.2).

Converting to more discriminating question formats is mostly a mechanical change to better probe differences of belief given a particular forecasting target. As AI forecasting systems improve, the preference for more discriminating formats will not change. Finding those forecasting targets about which frontier systems disagree, on the other hand, is an open-ended problem that is expected to evolve as the systems improve and the frontier advances.

3.1. More Discriminating Question Formats

There are many ways to pose forecasting questions and some of these are better able to expose differences of opinion than others. For example, we could ask 'Will the spot price of gold exceed $5,000 USD on Friday?' expecting a binary probability or we could ask 'What will the spot price of gold be on Friday?' expecting a probability distribution over some range of values, say $4,000 to $6,000 USD. The former is less able to distinguish differences of opinion than the latter because it collapses many points of view into the same forecasts, as Figure 1 illustrates.

In the Metaculus AI Forecasting Benchmarking Tournament

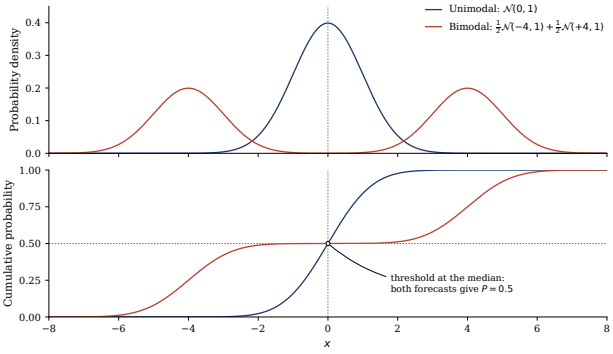

Figure 1. **Applying a threshold makes these forecasts appear to be in perfect agreement when they actually strongly disagree.** For example, the probability density functions (PDFs) of a unimodal function and bimodal function provide conflicting models of a random variable. It is possible to a select threshold such that their cumulative distribution functions (CDFs) agree.

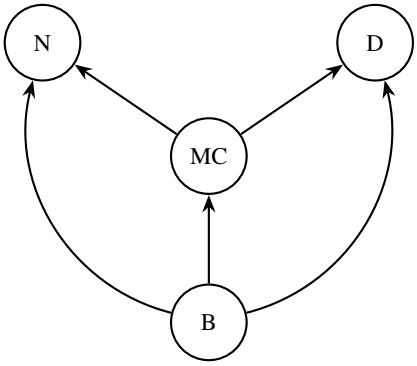

Figure 2. **Some question formats are more discriminating than others.** Where possible, it is preferable to convert from a less discriminating format to a more discriminating one because finer-grained formats expose disagreements that coarser formats collapse. Binary questions (B) may be converted to multiple choice (MC), date (D), or numerical (N) formats. Multiple choice (MC) may in turn be converted to numerical (N) or date (D) formats.

- 2025 Q2 edition, the following question was asked: *"What will Donald Trump's net approval on the issue of immigration be on June 25, 2025, according to the Silver Bulletin?"* with options *A. less than* -2.0%, *B. between* -2.0% *and* 0.0%, *inclusive*, *C. between* 0.0% *and* 2.0% and *D. greater than or equal to* 2.0%.

The resolution value was -4.2% and so the question resolved to option 1, less than -2.0%. The community had assigned 29% to that outcome, but the forecaster metaculus-o3 assigned over 68%. This seems like a strong forecast, but the system was, according to its provided explanation, failing to distinguish between overall approval and the 'approval on the issue of immigration' targeted in the question. Overall approval was nearly 8 percentage points lower than approval on the issue of immigration, and so the model was actually anticipating a far lower score but was saved by the fact that the outcome ranges binned together many different beliefs.

A less discriminating version of the approval question would be to set a single threshold e.g. Will ... be greater than 0.0% on June 25, 2025? Converting to the multiple choice format used in the tournament is an improvement over this, and we can go a step further and ask our forecasters to provide a probability distribution over the full range of possible values that approval can take, i.e. -100% to 100%. Figure 2 and Table 2 further illustrate preferable conversions with examples.

3.2. Seeking Disagreement

Forecasting tournaments require a group of willing forecasters and a process for writing questions. AI forecasting systems that are developed against a stationary question-writing process will saturate against that process as disagreements are resolved and weaknesses are patched. We propose closing the loop between the forecasters and the question-writing process by scoring questions for disagreement and rewarding question-writers based on how much disagreement they can generate.

Scoring rules. Scoring rules are proper if the forecaster maximizes the expected score for observations drawn from a distribution by issuing that distribution as their forecast. In this way, a proper scoring rule encourages the forecaster to think carefully and be honest (Gneiting & Raftery, 2007). The logarithmic score, or log score, of a prediction p given outcome ω

$$S(p, \omega) = \log p(\omega) \tag{1}$$

is a proper scoring rule.

The Metaculus Baseline Scores are a family of ergonomic scoring rules adapted for ease of human use (Metaculus, 2026). These scores are formed by taking affine transformations of the logarithmic scoring rule such that a uniform prediction over the outcomes scores 0, and so a positive score indicates beating the uniform *baseline*, and a perfect prediction scores 100. For binary questions the exact form is

$$S_B(p, \omega) = 100\,(\log_2 p(\omega) + 1) \tag{2}$$

Disagreement score. We propose measuring the disagreement of a set of forecasts $\{p_i\}$ as the variance of their log scores across forecasts, in expectation over outcomes drawn from the pooled forecast $\bar{p}(\omega) = \frac{1}{N}\sum_i p_i(\omega)$, specifically

$$D = \sum_{\omega} \bar{p}(\omega) \cdot \mathrm{Var}\left(100 \log_2 p_i(\omega)\right). \tag{3}$$

Table 2. **Question format conversion examples.**

Conversion		Example	
Start	**End**	**Original**	**Improved**
Binary	Date	Will event X happen by date D?	When will event X happen?
Binary	Multiple Choice	Will candidate C win election E?	Who will win election E?
Binary	Numerical	Will the value of series S be greater than threshold X on date D?	What value will series S take on date D?
Binary	Numerical	Will the value of series S exceed/fall below value X within period P?	What is the maximum/minimum value that series S will take within period P?
Multiple Choice	Numerical	Which bucket will politician P's approval rating fall in on date D?	What will politician P's approval rating be on date D?
Multiple Choice	Date	In which month M will event X happen?	When will event X happen?

We report $\sqrt{D}$, which lives in the same units as the baseline scores used to grade forecasters and can be read as the root-mean-square spread of log scores across forecasters at an outcome drawn from the pooled forecast. Equivalently, D is a generalized Jensen–Shannon divergence expressed in Metaculus baseline-score units, so the magnitude of disagreement is directly comparable to the magnitude of forecaster skill differences on the same questions. Further detail on applying the score to date and numerical questions is provided in the Appendix.

3.3. Further Tournament Details

Selecting the forecaster pool. Seeking disagreement between AI forecasters is useful for advancing the frontier provided the systems are of comparable strength and remain so over time. When one system is materially weaker, the questions on which they disagree are dominated by the weaker system's errors and so we should not expect to learn much about how to improve the stronger system. We therefore restrict the set of forecasters used to compute the disagreement score to a fixed panel of strong systems with established track records, reported in Table 5 in the Appendix.

Question-writing budget. In the standard forecasting tournament setup, forecasters have equal opportunity to succeed because they all answer the same questions. We extend this to question-writers by allowing them to all submit the same number of questions in total.

Immediate feedback. Because the disagreement score is a function only of the predictions and not the actual outcome, we can score as soon as the forecasts have been made and announce the score to the question-writer and the other competitors. This helps to avoid unnecessary delays in improving the question targeting.

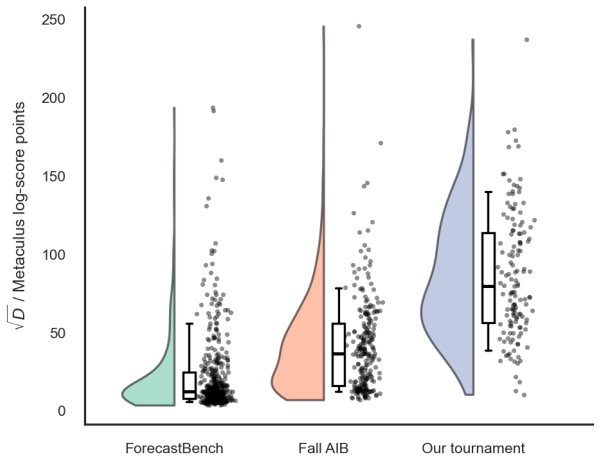

Figure 3. Per-question disagreement score on ForecastBench, Fall AIB, and our tournament. Summary statistics are provided in Table 3 in the Appendix.

4. Preliminary Results

Our preliminary results indicate that our tournament generates questions with substantially higher cross-forecaster disagreement than either AI benchmark. As shown in Figure 3, the median question on our tournament scores $\sqrt{D} \approx 80$, against $\sim$37 on Fall AIB and $\sim$12 on ForecastBench.

We consider higher disagreement to be a desirable property of the question set. A question on which the panel already agrees offers little room to improve forecasting performance: the consensus already captures most of the relevant public information, and a better forecasting method has little to contribute. A question on which the panel disagrees is the opposite case — the consensus is not yet settled, and a method that resolves the disagreement correctly turns that uncertainty into a score gain.

Impact Statement

This work aims to make AI forecasting evaluation more informative by directing community effort toward questions that actually discriminate between frontier systems, which we expect to accelerate progress in automated judgmental forecasting and, downstream, improve probabilistic decision-making in domains such as policy, public health, and economics. The same incentives could also be misused: a tournament that systematically surfaces the questions on which the best forecasters disagree highlights blind spots that adversaries could exploit, for instance to anticipate institutional responses or to time market and information operations. We mitigate this only weakly — the methodology rests on public benchmarks and standard scoring rules — and so encourage downstream users to treat the resulting forecasts as one input among several rather than as ground truth.

References

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

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*, volume 2025, pp. 93943–93980, 2025.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Metaculus. Metaculus ai benchmarks, 2025a. URL https://www.metaculus.com/futureeval/. Online; accessed 13-May-2026.

Metaculus. Ai forecasting benchmark tournament – fall aib, 2025b. URL https://www.metaculus.com/aib/2025/fall/. Online; accessed 13-May-2026.

Metaculus. Scores faq, 2026. URL https://www.metaculus.com/help/scores-faq/. Accessed: 2026-05-12.

Appendix

Disagreement score breakdown Table 3 expands Figure 3 into per-benchmark, per-question-type distributions of $\sqrt{D}$. Within every type the ranking is the same as in the chart: ForecastBench is the tightest, Fall AIB sits in the middle, and our tournament is the broadest. The gap is largest on numerical questions (median 88.1 vs. 56.1 on Fall AIB) and narrowest on multiple-choice (47.3 vs. 44.6).

Table 3. Per-benchmark, per-question-type distribution of $\sqrt{D}$ / Metaculus log-score points.

Benchmark	Question type	N	mean [95% CI]	median [p10–p90]
ForecastBench	binary	427	$22.7_{[20.2,\ 25.2]}$	$12.3_{[\ 6.1,\ 55.9]}$
Fall AIB	binary	128	$25.6_{[21.8,\ 29.4]}$	$17.2_{[10.6,\ 48.4]}$
	multiple_choice	23	$49.5_{[40.6,\ 58.4]}$	$44.6_{[26.6,\ 72.5]}$
	numerical	90	$62.8_{[56.0,\ 69.6]}$	$56.1_{[36.2,\ 93.5]}$
Our tournament	multiple_choice	12	$58.7_{[32.5,\ 84.9]}$	$47.3_{[33.4,\ \ 80.7]}$
	date	47	$83.5_{[69.5,\ 97.5]}$	$71.7_{[32.3,\ 141.2]}$
	numerical	80	$91.7_{[84.2,\ 99.2]}$	$88.1_{[53.2,\ 140.2]}$

Applying the disagreement score to continuous question formats. The sum in Equation (3) is over a discrete outcome space, which covers binary and multiple-choice questions directly. For numerical and date questions we adopt the Metaculus representation: the bounded range $[\text{range}_{\min}, \text{range}_{\max}]$ is partitioned into 200 uniformly spaced bins, with the first and last bins holding the probability mass below $\text{range}_{\min}$ and above $\text{range}_{\max}$ for open-bounded questions. A numerical or date question is therefore scored identically to a multiple-choice question over these 200 ordered outcomes, and Equation (3) applies unchanged. We note that the disagreement score does not use the ordering of the bins; it depends only on the per-bin forecasts, so a numerical question is, for scoring purposes, a fine-grained multiple-choice question.

Highest-disagreement questions Table 4 lists the five questions with the highest disagreement score recorded in the live tournament at the time of writing.

Recruiting question-writers. Question-writers applied through an open application linked from the tournament launch announcement. Applicants were screened for relevant expertise, including experience writing or forecasting questions on platforms such as Metaculus, Polymarket, Kalshi, and Manifold; designing AI benchmarks or evaluations; domain knowledge in areas such as economics, geopolitics, science, climate, energy, biosecurity, and supply chains; and backgrounds in data science, quantitative research, journalism, or intelligence analysis. We received 42 applications which were all accepted; 39 applicants signed up to participate; and 29 participants have asked at least one question at the time of writing. The $25,000 USD prize pool is distributed in proportion to the total number of points each

Table 4. The five tournament questions with the highest disagreement score.

Question	Score
How many consumer product recall announcements will the CPSC publish between 6/8/2026 and 8/12/2026?	282.94
How many Nuclear Regulatory Commission documents will the Federal Register publish between 6/8/2026 and 8/12/2026?	277.34
How many candidate lists will the Central Election Commission of BiH certify for the 2026 general election by August 5?	274.10
How many mass shootings will the Mass Shooting Tracker record in the United States between May 25 and June 8, 2026?	272.19
How many Federal Aviation Administration Airworthiness Directives will the Federal Register publish between 5/26/2026 and 8/12/2026?	272.08

participant scores across all of their questions. We allow at most 500 questions and so the average payment per question is $50, and the exchange rate from points to USD is least $1 per 10pts.

Forecaster panel Table 5 lists the bots comprising the forecaster panel over which the disagreement score is computed. All are drawn from the best-performing entrants in the Metaculus AI Benchmark series since Q4 2024.

Table 5. Bots in the forecaster panel used to compute the disagreement score. †Operated by the tournament organizers.

Bot
Mantic†
Preseen
Hayek-bot
tom_futureresearch_bot
pgodzinbot
cassi
laertes
smingers-bot
AtlasForecasting-bot
SynapseSeer
Panshul42