Rillor

RESEARCH REPORT

How LLM forecasts compare with superforecasters and markets

What LLM forecasting benchmarks show to November 2026: top-ranked entries statistically indistinguishable from superforecasters on ForecastBench, and models mostly level with or behind professional forecasters and market prices.



12 November 2026 Forecasting and market intelligence rillor.com/insights/llm-forecasting-accuracy-benchmarks

Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

In the studies reviewed here, one frontier model's 2023 tournament forecasts did not differ significantly from a constant 50%, by July 2026 several systems were statistically indistinguishable from superforecasters on one live benchmark, and professional forecasters and market prices matched or beat the models on overall scores in most other comparisons [8, 15, 18, 22, 24, 26].

- On ForecastBench, a live benchmark run by the Forecasting Research Institute (FRI), the top-ranked entry had a Brier Index of 69.6 on 10 November 2026 and the median of its superforecasters, forecasters picked for a strong track record, 68.2, on a 0 to 100 scale where higher is better and 50 means always forecasting 50% [2, 9, 12].
- FRI reported on 16 July 2026 that several systems on ForecastBench's tournament leaderboard were "statistically indistinguishable" from superforecasters and read the results as "more consistent with superforecaster parity than with outperformance" [8]. Its superforecasters last forecast in July 2024, so later models are compared through a statistical adjustment for question difficulty [1, 8].
- Models run out of the box, without tools or a crowd forecast, ranked below the public median [3]. On ForecastBench's baseline leaderboard of 10 November 2026, the median of the general public ranked second and the top model third, and every model was significantly less accurate than the superforecasters [1, 3].
- Against ten Pro Forecasters picked by Metaculus, the bot team's head-to-head score, a per-question log-score difference where 0 means equal, went from −20.03 in the second quarter of 2025 to −1.25 in the season ending 15 April 2026, a gap that was not statistically significant ($p = 0.247$) [22].
- Against market prices, models were mostly level or behind on overall scores. On the 104 matches of the 2026 FIFA World Cup, the pre-match market scored 0.469 on a three-outcome squared-error scale where lower is better, against 0.471 for the closest of four agents [26].
- A backtest on questions that have already resolved can leak outcomes through training data or retrieval. In an August 2026 preprint, four of five flagship models failed the usual check, a score comparison before and after the training cutoff, on ForecastBench questions they could not have memorized, because models know more about times near their cutoff [30].
- Separating forecasters takes many resolved questions. A May 2026 preprint scores an agent by its Alpha, the market price's Brier score (squared error, lower is better) minus the agent's, and calculates that detecting a true Alpha of 0.02 takes about 350 resolved binary predictions at 80% power, under its stated assumptions [31].

Background

Forecasters and tournaments

ForecastBench's human groups came from one round in July 2024: 39 superforecasters, "who have a strong track record of accurate performance on a diverse set of geopolitical questions", in a 9-day experiment with at least 3 and on average 8 forecasts per question, and 500 members of the public recruited online, averaging 49 responses per question [1, 12]. Metaculus runs a different design. Its FutureEval program has a model leaderboard on a fixed prompt across most open questions, bot tournaments that developers enter, and a team of hand-picked Pro Forecasters who forecast a subset of questions each season [20]. Earlier studies compared models with the median or aggregate of a platform's whole crowd [15, 16, 17].

Scores

The Brier score of a binary forecast is the squared error of the stated probability against the outcome, 1 or 0, averaged over questions. Lower is better, and the uninformed forecast of 0.5 scores 0.25 [12]. FRI introduced the Brier Index on 4 March 2026, defined as (1 − √Brier score) × 100%, "giving a 0–100% scale where 50% = always predicting 50% and 100% = perfect foresight", and adds that it "is not a 'percentage-correct' metric" [9]. Studies of football matches score three outcomes, on a scale where a uniform forecast scores 2/3 and the worst possible forecast 2 [27]. Metaculus's head-to-head score is a log score: for a question that resolves Yes, 100 × ln(bot team probability / Pro team probability), so 0 means equal [22]. TimeSeek uses a Brier skill score with the market price at prediction time as the reference, where a negative value means less accurate than the market [25].

Live forecasts and backtests

Paleka and colleagues describe the gold standard as "running it on unresolved questions, waiting until the questions resolve, and then scoring the predictions" [29]. ForecastBench takes that route: it "is comprised solely of questions about future events that have no known answer at the time of submission", and FRI does not backfill predictions or ask models about questions that have already resolved [12, 13]. Replays on resolved questions carry the leakage risks in Finding 7 [29, 30].

Method

What was examined. Published comparisons of LLM probability forecasts with human forecasters or market prices from October 2023 to 10 November 2026, with the question set, comparison group, scale, timing and leakage control behind each figure.

Sources. ForecastBench's leaderboard page and the data files behind its tournament, baseline and preliminary boards, as updated on 10 November 2026 [1, 2, 3, 4], its about and explore pages and projection file [5, 6, 7], its changelog [10] and methodology report [11], and four FRI posts from October 2025 to July 2026 [8, 9, 13, 14]. Papers accepted at a venue, read as their arXiv or open-access versions: ForecastBench at ICLR 2025 [12], Halawi and colleagues at NeurIPS 2024 [16] and Schoenegger and colleagues in Science Advances [17]. Preprints [15, 18, 23, 24, 25, 26, 27, 28, 29, 30, 31], two of them workshop papers [25, 28]. Four Metaculus analyses posted on LessWrong and Substack [19, 20, 21, 22].

How. Each figure stays with its question set, date and scale. No figure is compared across studies or across ForecastBench snapshots. Leaderboard figures were read from the data files behind the leaderboard page. Where a source tests only one direction, as ForecastBench tests only whether superforecasters are more accurate than an entry, the report states that direction. Figures marked "derived" are arithmetic on a source's numbers. Sources name labs and models, and the report describes them by role, such as the top-ranked entry or a frontier model.

The studies

Study (date)	Questions	Comparison	Scale	Timing	Result as published
Schoenegger and Park (October 2023) [15]	23 binary, Metaculus, July to October 2023	Human forecasters, 843 participants	Brier	Live	Model 0.20, humans 0.07
Halawi and colleagues (February 2024) [16]	914 test questions from June 2023	Crowd aggregate, five platforms	Brier	Test set after model cutoffs	System 0.179, crowd 0.149
Schoenegger and colleagues (November 2024) [17]	31 binary, Metaculus, October 2023 to January 2024	925 human forecasters	Brier	Live	Ensemble of 12 models 0.20, human crowd 0.19

Study (date)	Questions	Comparison	Scale	Timing	Result as published
ForecastBench paper (February 2025) [12]	200 questions, July 2024 round	39 superforecasters, public	Brier	Live	Superforecasters 0.096, public 0.121, top model 0.122
Lu (August 2025) [18]	464 Metaculus, July to December 2024	10 experts on 157 questions	Brier	After model cutoffs, news filtered to each question's open date (main set) or collected on it (hold-out set)	Top model 0.1352, expert median 0.0225, expert mean 0.1573
Prophet Arena (December 2025) [23]	1,367 Kalshi events resolved before 11 October 2025	Market price	Brier	Live	Top model 0.184, market 0.187
Commercial forecasting system (November 2025) [24]	1,610 liquid markets, April to May 2025	Market price	Brier	Retrospective	System 0.1258, market 0.1106
TimeSeek (April 2026) [25]	150 Kalshi markets, 15,000 forecasts	Market price at each checkpoint	Brier skill score	Five checkpoints per market	All ten models below the market, pooled
FIFA World Cup 2026 (July 2026) [26]	104 matches	Pre-match market	Three-outcome Brier	Live	Market 0.469, closest agent 0.471
LLM-SoccerArena (July 2026) [27]	104 matches	Closing prices	Three-outcome Brier	Live	Top model 0.497, market 0.498
Metaculus spring 2026 (September 2026) [22]	99 questions shared with Pros	10 Pro Forecasters	Head-to-head log score	Live, Pros hidden	Bots −1.25

Study (date)	Questions	Comparison	Scale	Timing	Result as published
ForecastBench board (10 November 2026) [2]	598 for the top entry	Superforecaster median, July 2024	Brier Index	Live models, 2024 humans adjusted	Top entry 69.6, superforecasters 68.2

Findings

1. Early comparisons with human crowds (2023 and 2024)

In a Metaculus tournament from July to October 2023 with 843 participants, one frontier model's mean Brier score on 23 binary questions was 0.20 against 0.07 for the human forecasters, and it did not differ significantly from forecasting 50% on every question (p = 0.23) [15].

Halawi and colleagues built a retrieval-based forecasting system and tested it on 914 questions from June 2023 or later, a test set published after the knowledge cut-offs of its models [16]. Its mean Brier score was 0.179 against 0.149 for the crowd aggregate, with accuracy of 71.5% against 77.0%, while the top baseline model scored 0.208 and random forecasts 0.250 [16]. By platform [16]:

Platform	System	Crowd
Metaculus	0.134	0.104
Good Judgment Open	0.193	0.157
Polymarket	0.172	0.127
Manifold	0.219	0.200
INFER	0.247	0.310

On its validation set, the system "significantly outperforms the crowd when the crowd's predictions express high uncertainty", and on the test set a weighted average with four times the weight on the crowd improved the overall Brier score from 0.149 to 0.146 [16].

Schoenegger and colleagues, in Science Advances in November 2024, pooled 12 models on 31 binary questions from a Metaculus tournament run from October 2023 to January 2024 with 925 human forecasters [17]. The model crowd's mean Brier score was 0.20 and the human crowd's 0.19, with no significant difference (p = 0.850) [17]. Their equivalence bounds of 0.081 Brier points are, in the authors' words, wide, and a Metaculus review of July 2026 notes that a constant 50% forecast would also qualify within them [17, 21]. The models showed an acquiescence bias, with a mean forecast of 57.35% while 14 of 31 questions resolved Yes [17]. In a second study, seeing the human median improved model accuracy by 17 to 28%, and the authors report that "simply averaging human and machine forecasts yields more accurate results" [17].

2. ForecastBench from 2024 to 2026

FRI, a nonprofit research organization, opens a ForecastBench round every two weeks with 500 questions split evenly between market questions from prediction platforms and dataset questions generated from time series, and updates the leaderboard nightly [5]. Its ICLR 2025 paper reported the July 2024 round on a 200-question subset: superforecasters 0.096, the public 0.121 and the top model 0.122, with superforecasters significantly ahead of both ($p < 0.001$) [12]. The top models had the crowd forecast on market questions, and the top model without it scored 0.136 [12].

FRI's posts since then trace the gap:

- **8 October 2025.** The top model scored 0.101 against 0.081 for superforecasters. FRI estimated model progress at about 0.016 Brier points a year and extrapolated parity to November 2026, with a 95% interval from December 2025 to January 2028. The median public forecast had fallen from second place a year earlier to 22nd [13].
- **29 January 2026.** External submissions held second and third place, and superforecasters led by 0.017 Brier points, "about one year of LLM progress". Parity was extrapolated to November 2026 overall, June 2026 on dataset questions and August 2026 on market questions [14].
- **4 March 2026.** The board moved to the Brier Index, with superforecasters at 70.6 (Brier 0.086) and the top models at 67.9 (Brier 0.103), a gap of 2.7 points [9, 10].

The ranking method changed in October 2025 to a difficulty-adjusted Brier score, with a 50-day wait before an entry is listed [10]. FRI's methodology report explains that a plain Brier ranking "fails when forecasters predict on non-overlapping questions", and that setting a market question's difficulty from the market's own Brier score "ensures forecasters can only rank above the market by outperforming it in head-to-head comparisons" [11].

The question mix changed in 2026. The RAND Forecasting Initiative ended on 31 July 2026, Kalshi questions entered from the set due on 30 August 2026, yfinance and Manifold questions were dropped on 24 September, and new yfinance questions came on 2 October [10].

3. The July 2026 parity finding and how it was tested

FRI's post of 16 July 2026 found that "several models are statistically indistinguishable from superforecaster-level accuracy" on the tournament board [8]. The null hypothesis was that a model is as accurate as superforecasters. The one-sided bootstrap p-values were 0.41 for the top system, 0.16 and 0.15 for two submissions from one lab, and 0.14 for another lab's submission [8]. FRI wrote that the top system was the first to rank above the superforecaster median on market questions, and that systems had matched superforecasters on dataset questions in May [8]. The top system generates sub-questions and search queries, retrieves and filters recent context, and has an ensemble of models forecast before another model reviews their reasoning [8].

FRI listed four caveats. The comparison "relies on a statistical extrapolation that grows less reliable over time". Results "are inherently stochastic". Intervals overlap, so the results "are more consistent with superforecaster parity than with outperformance". And the findings do not mean the benchmark is saturated [8]. Among its next steps FRI listed, on 16 July 2026, a fresh superforecaster round, updated dataset questions and quantile questions, each for fall 2026 [8]. The leaderboard page viewed on 11 November 2026 states that the human groups were last surveyed in July 2024 [1].

The tournament board as updated on 10 November 2026 [2]:

Row on the board	Overall Brier Index
Top-ranked entry, an anonymous team, 598 questions	69.6 (95% CI 68.0 to 71.0)
Superforecaster median, rank 7, 578 questions	68.2 (66.8 to 69.8)
Public median, rank 307	63.0
Naive forecaster	57.8
Always 0.5	50.0

Six entries rank above the superforecaster median [2]. The board tests one direction, whether superforecasters are more accurate than an entry. The board labels that test "No" for 36 entries from ten teams, two of them anonymous (p of 0.10 or more), "Likely" for 40 (p from 0.01 to 0.10) and "Yes" for 429 (p below 0.01), and no column tests whether an entry beats superforecasters [2]. The superforecasters' market score of 73.9 rests on 57 questions, and two entries score above it, on 129 and 140 questions [2]. By the index formula, 69.6 corresponds to a difficulty-adjusted Brier score of about 0.092 and 68.2 to about 0.101 (derived from [2, 9]). The preliminary board, which ranks resolved dataset questions before results are final, had 17 submissions above superforecasters in July and 47 on 10 November 2026 [4, 8].

The baseline board runs models "out of the box, without extra tools, context, or scaffolding" [1]. On 10 November 2026 it put superforecasters first at 67.1, the public median second at 62.1 and the top model third at 61.1, which was not significantly above the public ($p = 0.75$) [3]. Every model was significantly below the superforecasters [3].

FRI's trend file of 10 November 2026 projects tournament parity overall in May 2027 (June 2026 to March 2028), on dataset questions in November 2026 and on market questions in May 2028 (August 2026 to December 2030) [7]. The explore page calls the line "a simple extrapolation from past data" and warns that actual progress may deviate from it [6].

4. Tools, crowd forecasts and model ensembles

The tournament and baseline boards differ in what an entry may add to a model [1]. In October 2025 FRI found that several models given market forecasts in their prompts "simply copy them" [13]. One model's forecasts correlated 0.994 with the market forecasts provided, and it submitted exact market values on 26 of 122 questions [13]. FRI concluded that "The Baseline leaderboard tells the real story on market questions" [13]. The crowd forecast FRI gives its own runs was valid 10 days before the due date [1].

Entries also check themselves against the market. In January 2026 one team compared its forecast with the market price and, on a large disagreement, let a model review and adjust, which improved on its unadjusted ensemble by nearly 0.01 Brier points [14]. Another gave a model search tools and a Python interpreter, generated eight forecasts and averaged them [14].

Ensembles gain less when members agree. A June 2026 workshop paper on 113 binary Metaculus questions found that "many frontier LLMs make highly correlated predictions" and that "the strongest ensembles combine accurate but diverse forecasters" [28]. In LLM-SoccerArena, seven models' match probabilities had a mean pairwise correlation of 0.943, and an equal-weight ensemble improved on the average member by 0.0047 Brier points [27]. In the FIFA study the four agents made the same top pick in 92% of matches [26].

Metaculus reported in June 2026 that model choice "was the largest differentiator in bot performance" and that good scaffolding is worth "around nine months of base model progress" [20]. For most of FutureEval's history a single-prompt baseline bot placed in the top 10, and in spring 2026 the top baseline bot placed 18th of 173 [22]. Recalibration moves scores too: in a Metaculus analysis summarized in July 2026, Platt scaling improved the bots' Brier score by 0.016 on binary questions [21].

5. Against expert and professional forecasters

Lu's preprint, revised in August 2025, evaluated models on 464 Metaculus questions: 334 collected from 4 July to 30 September 2024, and a 130-question hold-out set from 1 October to 1 December 2024 for which news was collected as the questions opened, to prevent leakage [18]. The top model's Brier score was 0.1352 [18]. The paper sets this beside crowd figures from two other papers, 0.149 and 0.121, and notes that "Brier scores are not directly comparable across different question sets" [18].

Ten expert forecasters hired by Metaculus took part [18]:

Expert group	Questions	Mean Brier	Median Brier
Main set	157	0.1573	0.0225
Hold-out set	41	0.1222	0.0196

The paper's finding that experts "still significantly outperform the bots" sets the experts' median of 0.0225 against the model's 0.1352 [18]. The experts' mean on their subset, 0.1573, is above that model figure, and the two cover different question counts [18].

Metaculus compares a team of ten Pro Forecasters with its top 10 bots on questions where the Pro forecasts are hidden, so bots cannot copy them [19, 22]. The bot team's head-to-head score per question, where 0 means equal [22]:

Season	Bot team score (95% CI)
Third quarter 2024	−11.3 (−21.8 to −0.7)
Fourth quarter 2024	−8.9 (−18.8 to 1)
First quarter 2025	−17.7 (−28.3 to −7.0)
Second quarter 2025	−20.03 (−28.63 to −11.41)
Spring 2026, 7 January to 15 April	−1.25 (−4.87 to 2.37)

In the second quarter of 2025 the Pros' lead had p = 0.00001, on 93 scorable questions [19]. In spring 2026, 173 bots, 111 of them run outside Metaculus, forecast 297 scored questions resolving from January to May 2026, and the Pros forecast 99 of them [22]. The Pros' edge of 1.25 points was "not statistically significant (one-sided p = 0.247)" [22]. Nine of ten individual Pros beat every individual bot, two bots slightly edged out one Pro, and 67% of the binary questions resolved No [22]. A Metaculus synthesis of 8 July 2026, written before both the July parity post and the spring results, projected single-prompt baseline bots to beat Pros around June 2027 on a naive linear extrapolation [21].

6. Against market prices

Prophet Arena, a preprint revised in December 2025, scored models on 1,367 Kalshi events covering 72,136 markets, resolved before 11 October 2025 [23]. The top model scored 0.184 (±0.006) and the market baseline 0.187 (±0.006), with the other four models in the paper's main table from 0.189 to 0.219 and all 23 models evaluated from 0.184 to 0.303 [23]. Each of those five models was better calibrated than the market, with expected calibration errors from 0.041 to 0.067 against 0.069 [23]. The market lagged several models far in advance, and near resolution markets "incorporate breaking information and news updates more rapidly than LLMs" [23].

TimeSeek, an April 2026 workshop paper, asked 10 frontier models about 150 Kalshi binary markets at five points in each market's life, with and without web search, for 15,000 forecasts [25]. Four models beat the market at the Open+1 checkpoint, early in a market's life, and all trailed it by the Close-1 checkpoint near resolution, with skill scores below −0.7 [25]. Pooled with web search, all ten skill scores were negative, from −0.068 to −0.615, which the authors present as point estimates without confidence intervals [25]. Web search raised pooled skill by 0.14 to 0.59 points, and two-model ensembles reduced error without passing the market [25].

The commercial forecasting system's technical report of November 2025 tested it retrospectively on questions that resolved between July 2024 and June 2025 [24]:

Question set	System	Reference
ForecastBench set FB-7-21, 498 questions	0.1076	Superforecasters 0.1110, public 0.1451
ForecastBench set FB-8-14, 602 questions	0.1099	Superforecasters 0.1152
ForecastBench market questions, 76	0.0753	Superforecasters 0.0740, market 0.0965
Liquid markets resolving 2 April to 23 May 2025, 1,610	0.1258	Market 0.1106

The authors call the system "statistically indistinguishable from superforecasters on ForecastBench", call ForecastBench "a relatively easy forecasting benchmark", and report that on the liquid markets it "trails the performance of the market consensus" [24]. An ensemble with the market price, with a weight of 0.33 on the system, scored 0.106 against the market's 0.111 [24]. In a live check on 64 markets that closed by 26 August 2025 it scored 0.1002 with search and 0.3609 without, and the authors could not distinguish it statistically from the market [24]. The report's description of itself as "the first work that verifiably achieves expert-level forecasting at scale" is the authors' claim [24].

On the 104 matches of the 2026 FIFA World Cup, a July 2026 preprint ran four frontier models through one search, act and reflect loop [26]. The pre-match market, with the bookmaker margin removed, scored 0.4688 on the three-outcome Brier scale, and the agents 0.4705, 0.4706, 0.4729 and 0.4828 [26]. The paper reports no significance test for that gap, and its limitations section notes wide intervals for per-agent differences in a single tournament of 104 matches [26]. The authors call the benchmark "contamination-free by construction" because every match kicked off after the models' training cutoffs [26].

LLM-SoccerArena, also from July 2026, scored seven models on the same 104 matches and 15 tournament questions [27]. Across models, Brier scores ranged from 0.506 to 0.546, and none of the 21 paired comparisons stayed significant after Holm correction [27]. For open-book forecasts made two hours before kick-off, the top model scored 0.497 against 0.498 for closing bookmaker prices with the margin removed, and web access improved scores by 0.023 [27].

7. Leakage and evaluation disputes

Paleka and colleagues, in a May 2025 preprint, wrote that some claims of models matching or beating humans call for care, and named two problems: temporal leakage, and the difficulty of extrapolating from evaluation scores to real-world forecasting [29]. Their list includes leakage through the logic of the question set, retrieval where "date metadata on documents is often inaccurate", model cutoffs that are "more of a guideline than a guarantee", and piggybacking on human forecasts, which can make claims of human-level performance circular [29]. They found logical leakage a practical concern in four benchmarks, among them Halawi and colleagues' [29]. They wrote that ForecastBench "does show fewer obvious temporal and logical leakage issues", and that Brier scores "over-emphasize performance on questions with a high base rate" [29].

Zhang and Stadie, in an August 2026 preprint, tested the usual contamination check, a comparison of scores before and after a model's training cutoff [30]. On ForecastBench questions that resolved after five flagship models' documented cutoffs, four failed the check anyway, with spurious gaps of +0.044 to +0.061 [30]. "Models legitimately know more about times near their cutoff, so recency mimics leakage", they write, and they prove that no passive backtest can separate the two from genuine skill [30]. They report that a matched clean control "identifies it globally and yields a leakage-adjusted score" [30].

The commercial system's report estimated that about 1.65% of its search results carried some form of foreknowledge, and its robustness checks stayed within 0.001 of the baseline score [24]. Metaculus wrote on 22 June 2026 that "External claims that bots already match the best humans all rest on backtesting against already-resolved questions, and have not been reproduced under live, forward-looking evaluation" [20]. On 8 July 2026 it added that parity claims had not been evaluated "with the rigor needed to rule out leakage or statistical noise", and noted that the commercial system's paper classifies its ForecastBench result as a retrospective evaluation [21]. Both posts came before FRI's July parity post, whose model forecasts are live [8, 20, 21].

8. How many questions it takes to separate skill

Foresight Arena, a preprint revised on 4 May 2026, scores an agent by its Alpha, the market's Brier score for a round minus the agent's, so "An agent that echoes market prices scores $\alpha = 0$ by construction" [31]. Under a one-sided 5% test, maximal outcome variance and a typical gap of 0.15 between forecast and market price, it calculates the resolved binary predictions needed for 80% power [31]:

True edge in Alpha	Predictions needed
0.005	5,567
0.010	1,392
0.020	348

True edge in Alpha	Predictions needed
0.030	155
0.050	56

The authors recommend at least 20 rounds, 140 or more resolved predictions, before drawing conclusions about rankings, and 50 or more rounds before claiming an edge below 0.02 [31]. These results are analytical. The numbers in the paper's Section 6 come from a calibrated Monte Carlo simulation, and the paper reports no live results [31].

The other sources show the same constraint. ForecastBench's paper put the superforecasters' 95% interval at 0.076 to 0.116 on its 200-question subset [12]. ForecastBench lists an entry only after 50 days, because rankings stabilized once the 50-day horizon had passed, and since March 2026 it requires 50 resolved market questions per round [10, 11]. Metaculus could not clearly tell its top 10 spring 2026 bots apart, with confidence intervals about 8.5 points wide on average [22]. The two football studies rest on 104 matches each, and the commercial system's live check on 64 markets [24, 26, 27].

9. Where the sources disagree

- **Parity.** FRI found statistical parity on live model forecasts compared with 2024 human forecasts through a difficulty adjustment [1, 8]. Metaculus's Pros forecast the same questions as the bots at the same time, hidden from them, and kept a lead that was not significant in spring 2026 [22].
- **Projection and board.** FRI's trend file of 10 November 2026 puts overall tournament parity in May 2027, while the board of the same day ranks six entries above the superforecaster median [2, 7].
- **The World Cup.** On the same 104 matches, one study found no agent ahead of the market and another found the top model level with closing prices [26, 27].

Implications

For a reader comparing LLM forecasting results, the studies point to these checks.

- **Ask whether each forecast existed before its outcome.** Live results and backtests answer different questions, and Metaculus wrote in June 2026 that the parity claims it disputed rested on backtests [20, 29].
- **Ask what the comparison group saw, and when.** ForecastBench's humans forecast in July 2024, while Metaculus's Pros forecast the same questions as the bots in the same season [1, 22].
- **Check whether the model saw a crowd or market forecast.** Several models given market forecasts copied them [13], and models run without tools or crowd forecasts stayed below the public median [3].
- **Compare scores only within one question set, scale and snapshot.** Binary and three-outcome Brier scores, Brier Index values and log-score differences sit on different scales, and ForecastBench's index for superforecasters differs between snapshots [2, 9, 22, 27].
- **Count resolved questions before ranking.** Under one preprint's assumptions, an edge of 0.02 over the market price needs about 350 resolved binary predictions at 80% power [31].
- **Test combinations.** Combining model forecasts with the crowd improved on the crowd in 2024, and combining a system with the market price improved on the market in one retrospective test [16, 24].

Limits of this analysis

- Most sources are preprints, and several come from teams that run a benchmark or build a forecasting system: FRI runs ForecastBench, Metaculus runs FutureEval, and a commercial lab wrote the system report [8, 20, 24]. One author of the August 2026 leakage preprint is also an author of the commercial system's report [24, 30].
- The ForecastBench figures are one snapshot of a board updated nightly, and its question sources changed during 2026 [5, 10]. Its human comparison rests on the July 2024 round [1].
- Studies differ in models, prompts, tools, questions and dates, so no figure here is averaged across them.
- Metaculus's results were read in its posts, which report its own analyses.
- Most results here are on binary questions. Metaculus's spring 2026 comparison also covered 27 numeric and 9 multiple-choice questions, and FRI notes that its 2025 analysis covered only binary questions [13, 22].

Sources

1. Leaderboards, ForecastBench. Forecasting Research Institute, viewed 11 November 2026. https://www.forecastbench.org/leaderboards/

2. Tournament leaderboard data (leaderboard_tournament_full.js), ForecastBench. Forecasting Research Institute, updated 10 November 2026, viewed 11 November 2026. https://www.forecastbench.org/assets/js/leaderboard_tournament_full.js

3. Baseline leaderboard data (leaderboard_baseline_full.js), ForecastBench. Forecasting Research Institute, updated 10 November 2026, viewed 11 November 2026. https://www.forecastbench.org/assets/js/leaderboard_baseline_full.js

4. Preliminary leaderboard data (leaderboard_preliminary_full.js), ForecastBench. Forecasting Research Institute, updated 10 November 2026, viewed 11 November 2026. https://www.forecastbench.org/assets/js/leaderboard_preliminary_full.js

5. About, ForecastBench. Forecasting Research Institute, viewed 8 November 2026. https://www.forecastbench.org/about/

6. Explore, ForecastBench. Forecasting Research Institute, viewed 8 November 2026. https://www.forecastbench.org/explore/

7. Parity date projections for the tournament leaderboard (parity_dates.tournament_leaderboard.json), ForecastBench. Forecasting Research Institute, updated 10 November 2026, viewed 11 November 2026. https://www.forecastbench.org/assets/data/parity_dates.tournament_leaderboard.json

8. AI models have likely reached parity with superforecasters on ForecastBench. Forecasting Research Institute, 16 July 2026, viewed 8 November 2026. https://forecastingresearch.substack.com/p/ai-models-have-likely-reached-parity

9. Making Forecasting Scores Easier to Interpret: Introducing the Brier Index. Simas Kučinskas, Houtan Bastani and Matt Reynolds. Forecasting Research Institute, 4 March 2026, viewed 8 November 2026. https://forecastingresearch.substack.com/p/introducing-the-brier-index

10. Changelog, ForecastBench wiki on GitHub. Forecasting Research Institute, entries from 28 February 2025 to 5 October 2026, viewed 8 November 2026. https://github.com/forecastingresearch/forecastbench/wiki/Changelog

11. ForecastBench: Updated Ranking Methodology. Simas Kučinskas, Houtan Bastani and Ezra Karger, technical report. Forecasting Research Institute, 3 October 2025, viewed 8 November 2026. https://www.forecastbench.org/assets/pdfs/forecastbench_updated_methodology.pdf

12. ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities. Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang and Philip E. Tetlock, ICLR 2025, arXiv:2409.19839, version 5, 28 February 2025, viewed 8 November 2026. https://arxiv.org/abs/2409.19839

13. How well can large language models predict the future? Houtan Bastani, Simas Kučinskas and Ezra Karger. Forecasting Research Institute, 8 October 2025, viewed 8 November 2026. https://forecastingresearch.substack.com/p/ai-llm-forecasting-model-forecastbench-benchmark

14. LLMs Are Closing the Gap on Human Superforecasters. Houtan Bastani, Simas Kučinskas and Matt Reynolds. Forecasting Research Institute, 29 January 2026, viewed 8 November 2026. https://forecastingresearch.substack.com/p/llms-are-closing-the-gap-on-human

15. Large Language Model Prediction Capabilities: Evidence from a Real-World Forecasting Tournament. Schoenegger and Park, arXiv:2310.13014, version 1, 17 October 2023, viewed 8 November 2026. https://arxiv.org/abs/2310.13014

16. Approaching Human-Level Forecasting with Language Models. Halawi, Zhang, Chen and Steinhardt, NeurIPS 2024 (Advances in Neural Information Processing Systems 37, https://proceedings.neurips.cc/paper_files/paper/2024/hash/5a5acfd0876c940d81619c1dc60e7748-Abstract-Conference.html), arXiv:2402.18563, version 1, 28 February 2024, viewed 8 November 2026. https://arxiv.org/abs/2402.18563

17. Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. Schoenegger, Tuminauskaite, Park, Bastos and Tetlock, Science Advances 10(45), eadp1528. PubMed Central, 8 November 2024, viewed 8 November 2026. https://pmc.ncbi.nlm.nih.gov/articles/PMC11800985/

18. Evaluating LLMs on Real-World Forecasting Against Expert Forecasters. Lu, arXiv:2507.04562, version 3, 4 August 2025, viewed 8 November 2026. https://arxiv.org/abs/2507.04562

19. Q2 AI Benchmark Results: Pros Maintain Clear Lead. Metaculus, LessWrong, 28 October 2025, viewed 8 November 2026. https://www.lesswrong.com/posts/Surnjh8A4WjgtQTkZ/q2-ai-benchmark-results-pros-maintain-clear-lead

20. FutureEval: Metaculus's AI Forecasting Benchmark. Metaculus, 22 June 2026, viewed 8 November 2026. https://metaculus.substack.com/p/metaculus-futureeval-ai-forecasting-benchmark

21. AI Forecasting in 2026: What 11 Analyses Say. Metaculus, LessWrong, 8 July 2026, viewed 8 November 2026. https://www.lesswrong.com/posts/a82q6yd8zKpYk56cF/ai-forecasting-in-2026-what-11-analyses-say

22. FutureEval Spring Results: Pros Beat Bots, but the Gap is Nearly Gone. Ben Wilson and John Bash for Metaculus, LessWrong, 9 September 2026, viewed 8 November 2026. https://www.lesswrong.com/posts/wZBbDqzfBjYG58CxK/futureeval-spring-results-pros-beat-bots-but-the-gap-is

23. LLM-as-a-Prophet: Understanding Predictive Intelligence with Prophet Arena. Yang and others, arXiv:2510.17638, version 2, 21 December 2025, viewed 8 November 2026. https://arxiv.org/abs/2510.17638

24. AIA Forecaster: Technical Report. Alur and others, arXiv:2511.07678, version 1, 10 November 2025, viewed 8 November 2026. https://arxiv.org/abs/2511.07678

25. TimeSeek: Temporal Reliability of Agentic Forecasters. Mostafa, Shastri and Lee, workshop paper, arXiv:2604.04220, version 1, 5 April 2026, viewed 8 November 2026. https://arxiv.org/abs/2604.04220

26. FIFA World Cup 2026 as a Contamination-Free Benchmark for LLM Forecasting Agents: Four Models, a Bookmaker, and 104 Matches. Ding, Guo and Xu, arXiv:2607.17765, version 1, 20 July 2026, viewed 8 November 2026. https://arxiv.org/abs/2607.17765

27. LLM-SoccerArena: Benchmarking LLMs on Real-World Predictions in Sports. Schröder, Schweisthal, Müller, Weinmann and Feuerriegel, arXiv:2607.24573, version 1, 27 July 2026, viewed 8 November 2026. https://arxiv.org/abs/2607.24573

28. Diversity is the Strength of the AI Crowd. Aitchison, Jeen, Shevlane and Day, ICML 2026 workshop, arXiv:2606.29661, version 1, 29 June 2026, viewed 8 November 2026. https://arxiv.org/abs/2606.29661

29. Pitfalls in Evaluating Language Model Forecasters. Paleka, Goel, Geiping and Tramèr, arXiv:2506.00723, version 1, 31 May 2025, viewed 8 November 2026. https://arxiv.org/abs/2506.00723

30. Temporal Leakage in LLM Backtesting: Measurement, Validation, and Adjusted Scores. Zhang and Stadie, arXiv:2608.02985, version 1, 4 August 2026, viewed 8 November 2026. https://arxiv.org/abs/2608.02985

31. Foresight Arena: An On-Chain Benchmark for Evaluating AI Forecasting Agents. Nechepurenko and Shuvalov, arXiv:2605.00420, version 2, 4 May 2026, viewed 8 November 2026. https://arxiv.org/abs/2605.00420

About this report

Published by Rillor on 12 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. How LLM forecasts compare with superforecasters and markets. 12 November 2026. https://rillor.com/insights/llm-forecasting-accuracy-benchmarks

Notices

Rillor is not a registered investment adviser or commodity trading advisor and does not provide investment or trading advice. Research, forecasts, data and software described on this site are for research and engineering use. Nothing here is an offer or recommendation to buy or sell any security, commodity interest or digital asset. Past or simulated results do not indicate future results.

Superintelligence (SI) outputs can be wrong. Important decisions should include human review.

Work with Rillor

Send the question you want forecast, its horizon and the data you already hold. Contact us at rillor.com/contact?type=forecasting or contact@rillor.com. More on forecasting and market intelligence: rillor.com/markets/forecasting