PREDICT THE FUTURE: A REVIEW OF LANGUAGE MODEL EVENT FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

A language model can appear to predict the future for very different reasons: it may recall an outcome already encoded in its training data, retrieve evidence available at forecast time, reason about an unresolved event, or echo a human or market estimate. Because these capabilities are tested across historical, live, structured, and open-ended benchmarks with different information boundaries and scoring rules, a reported gain can be difficult to interpret or compare. This review asks what reported improvements in real-world event forecasting actually establish. It organizes the field around probabilistic forecasts, structured event predictions, and open-ended future generation, distinguished by their temporal and outcome-resolution requirements. We compare retrieval, reasoning, calibration, training, aggregation, and agentic updating as distinct interventions, focusing on the controls needed to attribute gains to each. A source-traceable evidence map links reported results to information boundaries, forecast horizons, outcome-resolution rules, metrics, and aggregation populations. Its central distinction is between the information a system receives and what it can infer from that information; probability quality, scenario coverage, and decision utility are separate evaluation targets. We close with research priorities for temporally valid comparisons, calibrated uncertainty, reliable forecast updating, and decision-aware evaluation.

1 INTRODUCTION

Suppose a model predicts that a company will release a product by June. Whether this counts as a forecast depends on what "release" means, which source resolves the outcome, and what information was available when the model answered. If it answers after launch, memory or retrieval may reconstruct the outcome, but that success does not show anticipation. A fluent explanation can hide the difference.

For language models, this distinction is difficult because one prediction may draw on fresh reporting, quantitative indicators, and information retained in model parameters. A high score does not identify which source supported the prediction: a system may reason from permitted evidence, retrieve useful information, or reproduce a probability supplied by a human or market. Historical benchmarks may expose models to outcomes that were unresolved when questions were written, while retrieved pages can include revisions published after a nominal cutoff (Jin et al., 2021; Paleka et al., 2026). Prospective evaluation reduces some temporal risks, but a system may still receive market probabilities or human forecasts, changing what its score measures (Karger et al., 2025). Evaluation therefore needs to trace the information path from question construction, through tool use, to forecast submission.

Research spans historical question collections as well as prospective forecasts committed before outcomes resolve, but these designs support different temporal claims. ForecastQA and Autocast evaluate predictions against resolved historical events, while ForecastBench, FutureX, and Prophet Arena collect prospective or live forecasts (Jin et al., 2021; Zou et al., 2022; Karger et al., 2025; Zeng et al., 2025; Yang et al., 2025). MIRAI evaluates structured event relations, while OpenForecaster studies open-ended event generation (Ye et al., 2024; Chandak et al., 2025). Because target forms,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

information boundaries, resolution procedures, and comparison populations differ, an accuracy value or leaderboard position cannot be compared across settings without qualification.

Several reviews and position papers already synthesize important parts of this field. Su et al. cover LLM forecasting and anomaly detection; Lee et al. discuss large-scale training for event forecasting and its broader impacts; Xu et al. provide the closest broad synthesis, spanning numerical and event targets, methods, training, evaluation, and applications. Paleka et al. focus on temporal leakage and on limits to extrapolating benchmark performance to real-world forecasting (Su et al., 2024; Lee et al., 2025; Xu et al., 2026; Paleka et al., 2026). Xu et al. already review forecasting evaluation, so we do not frame evaluation itself as an unaddressed gap. Our narrower unit of synthesis is the reported comparison: we trace its target, information boundary, resolution, metric, and evaluated population to the source, then ask what that comparison supports. Table 1 makes the overlap explicit.

Rather than rank forecasters across incompatible settings, we record the evaluation contract behind each selected result: its target, information available at issuance, resolution rule, metric, and evaluated population. A 53-record evidence ledger traces these fields to source locations and separates reported values from our deterministic recomputations. We use this map to audit within-benchmark metric trade-offs and the sensitivity of leaderboard order to aggregation. The contribution is a bounded, inspectable account of what these comparisons establish, not a new forecasting method or an exhaustive census.

Within-study audits show why these records matter: ForecastQA accuracy improves while class-summed Brier loss worsens, and SCATTER Pass@16 coverage rises even as mean Pass@1 falls, separating repeated-sample coverage from one-response performance (Jin et al., 2021; Chang et al., 2026). From the pinned ForecastBench archive, we reconstruct scores for 130 configurations and find that dataset weights of 0.25 and 0.75 reverse 9.1% and 6.9% of 8,354 strict pairwise orderings, respectively, relative to equal weighting (ForecastBench team, 2025). This deterministic sensitivity analysis uses published, rounded components; it does not rerun the forecasting systems. Taken together, these audits show why a single leaderboard can obscure both the capability being measured and the conditions that produced the score. Figure 1 connects task contracts, forecasting mechanisms, and result evidence; its dashed controls are proposed comparisons, not experiments completed in this review.

Section 2 describes the review scope and selection procedure; Section 3 defines event-forecasting task contracts; Sections 4 and 5 organize mechanisms and benchmarks; Section 6 analyzes what reported results establish; Section 7 presents the research agenda, and Section 8 discusses limitations and concludes.

2 Scope and review procedure

We include systems that forecast events whose outcomes remain unresolved at issuance, whether outputs describe closed outcomes, structured relations, open-ended future events, or time-to-event distributions. Numerical extrapolation is included only when it clarifies an event-forecasting boundary or component. Event extraction, temporal ordering and ordinary retrieval remain adjacent tasks: performance on them alone does not establish forecasting ability.

We built a targeted corpus from benchmark papers and related surveys, then traced citations and searched arXiv, ICLR proceedings, OpenReview and ACL Anthology. Queries combined event forecasting, language models, probability prediction and live evaluation. We retained work specifying an event target or evaluating a forecasting mechanism, while recording adjacent numerical-forecasting studies separately. Conference, workshop and preprint status remain distinct. Appendix A records the September 2026 search follow-up and inclusion decisions.

For quantitative claims, we checked the original table and retained its version, target, model setting, metric and population. Source-reported values are not presented as replications, and unreported capabilities or protocol fields remain unknown. Because exhaustive database screening and independent duplicate extraction are incomplete, this review does not claim PRISMA-complete coverage. That limitation affects scope, while each included claim remains traceable.

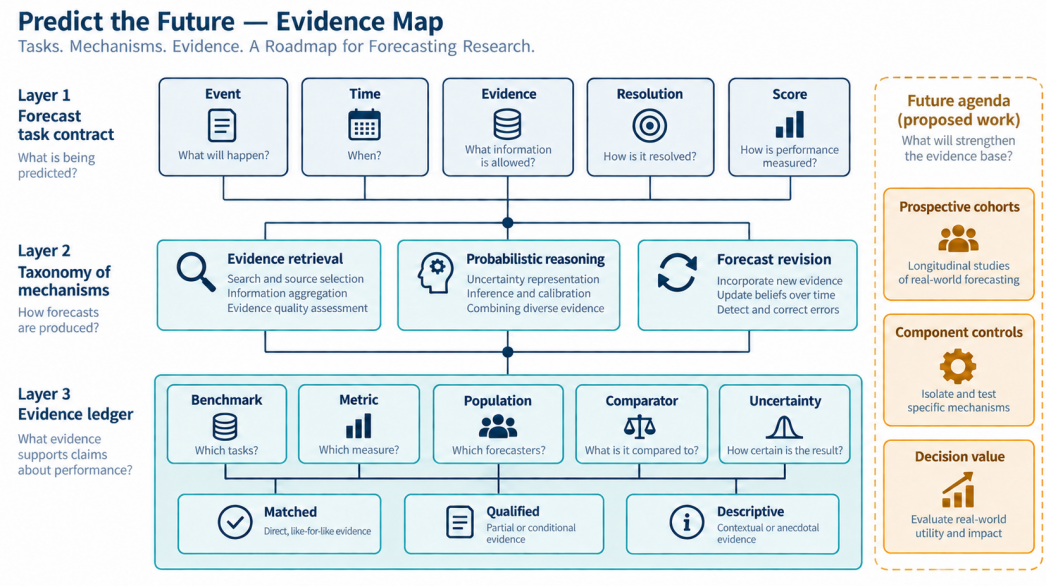


Figure 1: Evidence map for interpreting event-forecasting results. The task contract specifies the event, horizon, permitted evidence, resolution, and score; the mechanism taxonomy separates retrieval, probabilistic reasoning, and revision; and the evidence ledger records benchmark, metric, population, comparator, and uncertainty. The dashed panel marks proposed research, not completed experiments.

Table 1: Relationship to existing syntheses. The final column identifies overlap and the bounded role of this review; it does not imply that prior work omitted the listed analyses.

Work	Scope	Relation
Zhao	Event prediction across methods and applications	Historical foundations and target breadth
Benzin & Rinderle-Ma	Event prediction from a systems perspective	Pipeline and integration context
Deng et al.	Human-event modeling, graphs to LMs	KDD tutorial; structured-event lineage
Su et al. (2024)	LLM forecasting and anomaly detection	Broad survey; adjacent to event-focused forecasting
Lee et al. (2025)	Event-forecasting training, data and societal impacts	Position-paper perspective; complements result-level audits
Xu et al.	Numerical and event forecasting agents, methods, training, evaluation and applications	Substantial overlap; focus here is a bounded result-level ledger and aggregation-sensitivity audit
Paleka et al.	Temporal leakage and extrapolation limits	Core evaluation analysis
This review	Selected event results traced to evaluation contracts	Claim provenance, metric conflicts and aggregation sensitivity

3 What is an event forecast?

An evaluation instance can be written as

$$\mathcal{C} = (q, \Omega, t, \tau, \mathcal{I}_t, \rho, S, \mathcal{U}), \tag{1}$$

where q is the question, Ω the outcome space, t the forecast time, τ the resolution horizon, $\mathcal{I}_t$ the permitted information, ρ the resolution rule, S the scoring rule and $\mathcal{U}$ the population and aggregation unit. These terms define an evaluation contract, not a new forecasting model: two results with the same model and metric may still measure different tasks when one component changes.

A binary question asks for $p = P(Y = 1 \mid q, \mathcal{I}_t)$, with Y resolved by ρ. Categorical forecasts require a normalized distribution over mutually exclusive, collectively exhaustive alternatives; a residual category can cover outcomes outside the named options. Exact resolution rules matter. "Announced," "available to selected users" and "publicly released" can assign different outcomes to the same product story. A model should not be rewarded or penalized using a definition introduced after its prediction.

Beyond closed-outcome probabilities, a relational task can ask which action or relation will occur between named entities during a future interval. MIRAI asks which CAMEO-coded relation will connect two countries at a future timestamp, using a cleaned GDELT event database and multiple forecast horizons (Ye et al., 2024). Because the labels depend on that event coding and relation hierarchy, relation F1 measures agreement with recorded labels, not a calibrated probability for one binary event. News coverage and extraction omissions also affect what becomes a label.

Temporal knowledge graph (TKG) forecasting is a related but distinct structured subfamily. A TKG represents timestamped facts as subject–relation–object–time tuples, and forecasting may rank candidate entities or facts for a future timestamp. Recent systems such as RECIPE-TKG and critic-guided rule induction evaluate this task with ranking metrics such as Hits@k (Akgül et al., 2026; Ning et al., 2026). These scores answer whether a known future fact is highly ranked among candidates; they are not calibrated event probabilities or relation-extraction F1. We therefore include TKG forecasting in the task boundary while keeping its ranking results in a separate evidence stratum.

Removing a predefined ontology leads to open-ended forecasting: the model must describe possible future events without a closed answer set (Wang et al., 2025; Yu et al., 2025). Evaluation must decide whether two descriptions refer to the same event, whether details are sufficiently specific and whether unsupported additions count as errors. A confident but vague answer may match many possible futures.

Time-to-event forecasts have a separate target: when a specified event will occur. If observation ends before the event, its time is right-censored rather than a negative outcome; excluding or imputing such cases can forfeit calibration guarantees (Bostrom et al., 2023). A benchmark that includes this target should report the observation window, censoring status, and scoring treatment for unresolved cases.

A historical prompt can place an event in the future even when the evaluated model already knows its outcome. A backtest is scored after resolution and therefore evaluates a retrospective cohort; separately, the model or its retrieval system may receive post-forecast information, creating temporal leakage (Paleka et al., 2026). Conversely, an outcome may already be determined in the world but remain unknown to the system. Prospective commitment addresses outcome availability at evaluation, but it does not by itself control question selection, external probability inputs or later changes to resolution rules.

4 How forecasting systems use language models

Methods differ in where they act along the information path: elicitation, evidence acquisition, inference, aggregation, or updating. Training can change several stages at once, while tools may alter both available evidence and subsequent actions. A system-level score difference therefore cannot identify a reasoning gain by itself; attribution requires controls matched to the components that changed.

At the input stage, a question-only forecaster tests how the model responds to the prompt and its retained knowledge; a date instruction cannot erase memorized outcomes. Retrieval adds external evidence, but displayed publication dates do not establish that page contents were fixed at that time (Paleka et al., 2026). Log retrieved versions and record the forecast timestamp after tool use so that the prediction's evidence boundary can be audited.

Given acquired evidence, decomposition makes prerequisites, base rates, and alternative scenarios explicit. Correlated reports or rationales can repeat the same signal, so aggregation should not assume that additional samples supply independent evidence (Jin et al., 2026). Model-only ensembles should also be separated from combinations with human or market forecasts: the latter may be useful but do not establish independent human-level performance (Paleka et al., 2026).

PRISM illustrates an inference pipeline that models repeated LLM extractions as noisy measurements of latent factors. A Bayesian model combines them into probabilistic forecasts and propagates extraction uncertainty; its factor-contrast summaries are associations, not causal explanations (Jin et al., 2026). Training changes a model's mapping from evidence to forecasts, whereas retrieval and tools change the evidence available at forecast time (Xu et al., 2026).

Sequential systems produce several forecasts for one event, so evaluation should retain the trajectory and cluster uncertainty at the event level. Scoring only the terminal update hides earlier errors; treating dependent updates as independent observations overstates the effective sample size. Equation 1 specifies whether events, forecasts or event–horizon pairs receive equal weight, and abstention requires coverage alongside loss. Benchmarks that automate question generation or resolution should report those stages separately from forecasting, lest event-generation or judging performance be conflated with forecast quality (Bosse et al., 2026; Wang et al., 2025).

4.1 What changes when a forecasting system improves?

A pipeline diagram alone does not identify the source of a gain. Retrieval changes available evidence; aggregation changes how estimates are combined; training changes the mapping from evidence to predictions. Halawi et al. combine retrieval with performance-selected fine-tuning, while OpenForecaster also changes question construction and the confidence objective (Halawi et al., 2024; Chandak et al., 2025). FutureBench already proposes separating framework, tool and model comparisons by holding the other components fixed (Bianchi et al., 2025). Matched inputs and output budgets remain necessary when interpreting a particular reported gain. Appendix B.1 provides the detailed study-level comparisons.

Three contrasts illustrate the distinction. PRISM's calibration and statistical baselines show why direct LLM probabilities are an insufficient reference for attributing a full-system gain (Jin et al., 2026). Competence pooling improves on some external forecasts, but the study finds no significant difference from its calibrated logistic comparator (Tiwari et al., 2026). MIRAI's matched query subset shows different interface contrasts across backbones; its oracle Max@K uses realized labels and must not be treated as deployable aggregation (Ye et al., 2024). These are different empirical questions, even when all systems contain retrieval, reasoning and aggregation.

The prediction object and evaluated cohort also matter. SCATTER increases hypothesis-set coverage but lowers one-round success in one comparison (Chang et al., 2026); greater set coverage is not probability quality. Specialist roles do not ensure complementary evidence: one sports study finds substantial output overlap (Varghese et al., 2026). FutureX reports model-specific missing predictions; because strict alignment would substantially reduce its sample, it allows slight cohort differences and estimates added score variability by simulation (Zeng et al., 2025). This trade-off does not make the cohorts interchangeable.

These examples motivate four audit questions: does a component change evidence, the mapping to forecasts, the scored object, or the evaluated population? Controls are shared evidence, a common inference or calibration reference, a fixed prediction object, and common cohort and aggregation rules. Appendices report negative and qualified findings with populations, metrics, and uncertainty. These controls guide future matched tests; they do not identify causes of published gains.

5 Benchmark landscape

Benchmarks vary in when forecasts are committed, what information systems may use, and how targets and scores are defined. Table 2 maps representative tasks along these dimensions; the discussion distinguishes repeatability, temporal validity, and cross-benchmark comparability rather than ranking unlike evaluations.

Historical datasets support repeatable experiments, but the forecast-time information boundary needs separate scrutiny. ForecastQA pairs event questions with temporal text, while Autocast includes heterogeneous targets and crowd baselines (Jin et al., 2021; Zou et al., 2022). For both, ask what the model and retrieval system could know about the outcome at the nominal forecast time. This question narrows the claim a retrospective result supports, but does not by itself invalidate the comparison.

Table 2: Representative event-forecasting benchmarks classified by target, temporal protocol, and evaluation distinction.

Benchmark or task family	Target and output	Evaluation distinction
ForecastQA (Jin et al., 2021)	Historical question answers and probabilities	Time-filtered context; model outcome knowledge remains relevant
Autocast (Zou et al., 2022)	Binary, multiple-choice and numerical questions	Heterogeneous scores and historical crowd references
MIRAI (Ye et al., 2024)	Future international relation labels	Structured event ontology, tool access and relation F1
Temporal KG forecasting (Akgül et al., 2026; Ning et al., 2026)	Future timestamped facts or candidate entities	Ranking (e.g., Hits@k); separate from probability and relation-F1 scores
ForecastBench (Karger et al., 2025)	Binary future-event probabilities	Outcome and interim crowd-reference scores; multiple horizons
Forecast-Dojo (Ye et al., 2026a)	Resolved market events replayed at dated-news checkpoints	Shared longitudinal steps; immediate outcome feedback for evaluation and training
FutureBench (Bianchi et al., 2025)	News- and market-derived questions	Project report; tool-assisted agent comparisons
FutureX (Zeng et al., 2025)	Diverse live event questions	Task-specific scores, tier weighting and unmatched coverage
Prophet Arena (Yang et al., 2025)	Prediction-market events	Probabilities, calibration and market-relative decisions
OpenForecast (Wang et al., 2025)	Historical events: generation, arguments and closed-choice tasks	Retrieval-augmented evaluation of generated events
ForestCast (Yu et al., 2025)	Next event in an open-ended news development chain	Occurrence judged using retrieved news; separate LLM-rated forecast quality
OpenForesight (Chandak et al., 2025)	Open-ended forecasting questions	Offline news corpus; training data for OpenForecaster
ForecastBench-Sim (Lee et al., 2026b)	Binary and continuous forecasts in Freeciv	Controlled world reports and interventions; real-world transfer unestablished

Prospective benchmarks commit forecasts before resolution. ForecastBench compares models with humans; FutureX collects live questions; Prophet Arena links forecasts to markets (Karger et al., 2025; Zeng et al., 2025; Yang et al., 2025). Forecast-Dojo replays resolved events at dated-news checkpoints; FutureBench is a public project report (Ye et al., 2026a; Bianchi et al., 2025). In the 2026 AI World Cup export, GPT-5.5 Thinking ranks first on tournament composite points, whereas Claude Sonnet 4.6 has the highest group-stage outcome accuracy (46 of 72 matches). These are distinct summaries of ten submitted configurations in one tournament: the composite also includes source-reported group-standing and knockout points, so this contrast is descriptive rather than a general ranking (Shianifar & Faiud, 2026). Event source, horizon, evidence, selection and scoring differ, so live status alone does not make results comparable.

The adjacent boundary includes event-conditioned numerical forecasts and controlled simulations. ForecastBench-Sim uses fixed state reports in Freeciv, allowing interventions and rapid resolution; transfer to real events remains unestablished (Lee et al., 2026b). LEAF includes event probabilities, trend and time-series forecasts, supported by time-restricted retrieval and dual-agent validation (Tan et al., 2026b). TemporalBench spans structure interpretation, context-free, contextual and event-conditioned time-series prediction (Weng et al., 2026). These settings clarify the boundary: numerical-series accuracy and event-probability quality require different metrics.

OpenForecast adds three open-ended tasks: argument-level prediction, event prediction on a specified date, and long-term event evolution (Wang et al., 2025). For these tasks, its LLM-based Retrieval-Augmented Evaluation (LRAE) splits forecasts into atomic predictions, retrieves web evidence, and semantically matches predictions to events. This can capture developments missing from reference timelines, but the retrieved pages belong to the evaluator and are not necessarily

forecast-time inputs. Reported scores therefore depend on the judge, retrieved evidence, and matching procedure as well as on the forecasts themselves.

ForestCast organizes news into story trees and predicts the next event along each branch (Yu et al., 2025). NewsForest contains 12,406 chains across six months, a span limited over stated leakage concerns; experiments use a 30-day post-forecast window to determine whether predictions occur. Table 7 reports Qwen2.5-7B Pass@1 of 16.22%, 18.27% with DPO-irrelevant, and 17.34% with DPO-fake (Pass@5: 56.22%, 62.22%, and 59.56%). These occurrence-based Pass@k and LLM-rated quality measures are protocol-specific, not proper probability scores, and remain separate from Brier and log-loss results.

Recent benchmarks add distinct sources of evidence. Curating the Future's free-form confidence score is not directly comparable to binary Brier loss; PROPHET reports a Brier score scaled by 100 on January 2025 Polymarket questions paired with retrieved news (Chandak et al., 2026; Tao et al., 2026). FutureSearch automates question generation and later resolution, making both stages part of benchmark construction (Bosse et al., 2026). Foresight Arena's v2 preprint reports a deterministic, seed-controlled simulation calibrated to prior Brier ranges and explicitly defers results from its deployed benchmark (Nechepurenko & Shuvalov, 2026). These evidence types belong in separate strata, not one leaderboard.

6 What the reported performance measures

6.1 Scoring rules define the claim

For a binary outcome y, the usual Brier loss is

$$B_{\text{binary}} = (p - y)^2. \tag{2}$$

Summing both class errors doubles Equation 2:

$$B_{\text{class}} = (p - y)^2 + ((1 - p) - (1 - y))^2 = 2B_{\text{binary}}. \tag{3}$$

The identity in Equation 3 permits a scale conversion only when both expressions refer to the same binary predictions and population. It does not justify rescaling an aggregate that mixes binary and multiclass questions, or comparing different questions. ForecastQA sums class errors over two- and four-choice questions, whereas OpenForecaster adapts Brier loss to confidence in a free-form answer and explicitly distinguishes this reward from the classical binary score (Jin et al., 2021; Chandak et al., 2025). Placing their values in one column called "Brier" would hide differences in range, direction and meaning.

Calibration alone does not ensure informative predictions: repeatedly reporting the base rate can be calibrated yet uninformative (Ovadia et al., 2019; Gneiting & Raftery, 2007). Expected calibration error (ECE) is not a proper scoring rule, and its estimate depends on binning choices; it is therefore best read as a diagnostic alongside proper scores (Ovadia et al., 2019).

Pointwise scores do not capture every property of a forecast. Consistency checks ask whether predictions on logically related questions fit together; Paleka et al. report correlation with later Brier scores, which makes these checks a possible early signal rather than a substitute for resolved-outcome evaluation (Paleka et al., 2025). Economic return answers a different question because it depends on actions, payoffs and costs; it should be reported under an explicit policy and separately from probability scores.

6.2 Historical and prospective evaluations answer different questions

On ForecastQA's test set, BERT$_{\text{LARGE}}$++ reaches 60.1% accuracy; human majority-vote accuracy is 71.2% with the top-ten retrieved articles and 79.4% with timestamp-constrained Google Search (Jin et al., 2021). These human scores come from seven annotators answering 150 sampled questions. Annotators could skip questions when they could not set aside future knowledge, and did so for 28.3% of questions on average. Because the human comparison uses a sampled cohort, skips some questions, and varies information access, it does not establish a fully paired model–human gap over the test set.

Autocast reports 65.4% true/false and 35.8% multiple-choice accuracy for FiD Static 2.8B, compared with 92.4% and 81.0% for the crowd (Zou et al., 2022). The crowd score is the final aggregate before question close, after forecasts can update over time; it is therefore an end-of-window reference rather than a matched information snapshot (Zou et al., 2022).

On the full test set, Halawi et al. report Brier losses of 0.179 for their system and 0.149 for the crowd (Halawi et al., 2024). A system–crowd blend with four times the weight on the crowd scores 0.146; the authors select this weight on validation data, so the result measures a hybrid forecast rather than the system alone.

The paper separately evaluates a selective cohort using rules motivated by validation results: the crowd probability lies between 0.3 and 0.7, the forecast is from one of the first three retrieval dates, and at least five relevant articles are available. In this subset, system and crowd Brier losses are 0.240 and 0.247, while their unweighted blend scores 0.237. The subset retains 22% of forecasts and 43% of questions, so the system's small advantage applies to a selected population and does not establish that it beats the crowd on the full test set (Halawi et al., 2024).

The report evaluates 177 resolved Metaculus questions, averaging five forecasts per question. It reports system and crowd accuracy of 87.7% and 87.0%, respectively; Brier scores are 0.0999 before and 0.0934 after Platt scaling, versus 0.0952 for the crowd (Phan et al., 2024a). The CAIS post reports 87.7% "± 1.4" without defining the spread (Phan et al., 2024b); the technical report does not state whether Platt scaling uses held-out data (Phan et al., 2024a). These retrospective results do not establish prospective performance.

ForecastBench's ICLR 2025 Table 2 compares the superforecaster median, public median, and LLMs on 498 evaluation units from the human question set. It reports Brier scores of 0.096, 0.121, and 0.122 for the superforecasters, public, and Claude 3.5 Sonnet with scratchpad and freeze values, respectively (Karger et al., 2025). Its 422 dataset outcomes and 76 market questions contribute equally by source group, so market questions (15.3% of units) receive half the aggregate. Twenty-seven market questions remain unresolved and are scored against community forecasts rather than outcomes. For market questions, a freeze value is the crowd forecast recorded when the question set is created; the model therefore receives a human-market signal in this condition. The 0.122 result is specific to this mixed cohort and protocol, not a question-only estimate of the model's forecasting ability.

ForecastBench's human references use different elicitation protocols: the public median comes from a one-hour survey, while the superforecaster median comes from a nine-day exercise with a later group-discussion stage (Karger et al., 2025). Because participant cohort, time allocation, and collaboration vary together, the score gap cannot isolate expertise. Appendix C records the cohort sizes, question assignments, subgroup results, and source versions; the later ranking update is kept distinct from the original ICLR comparison (Karger et al., 2025; ForecastBench team, 2026).

6.3 Result-level audits expose comparison limits

Aggregation can change model order even within a single benchmark. Reconstructing the published rounded freeze-value rows in ForecastBench Table 2, Claude 3.5 Sonnet with scratchpad has lower Brier loss than GPT-4 Turbo zero-shot under equal dataset–market weighting; the ordering reverses when the dataset receives less than one third of the total weight (Karger et al., 2025). This crossing is an arithmetic comparison of two reported rows, not a significance test.

The same issue appears across the pinned human-cohort leaderboard. Its 130 configurations form 8,354 pairs that are not tied under equal weighting. Moving the dataset weight to $w = 0.25$ or 0.75 reverses 758 (9.1%) or 578 (6.9%) pairwise orderings and creates 20 or 19 new ties; at $w = 0$ or $w = 1$, 1,582 (18.9%) or 1,037 (12.4%) orderings reverse (ForecastBench team, 2025). Under a nearest-0.001 rounding assumption, a conservative bound certifies 613 of the 758 reversals at $w = 0.25$ and 452 of the 578 at $w = 0.75$. The pairs share configurations, so these are deterministic score-level descriptions rather than independent tests. The rounding bounds and their assumptions are detailed in Appendix D.1; no weighting is designated as correct.

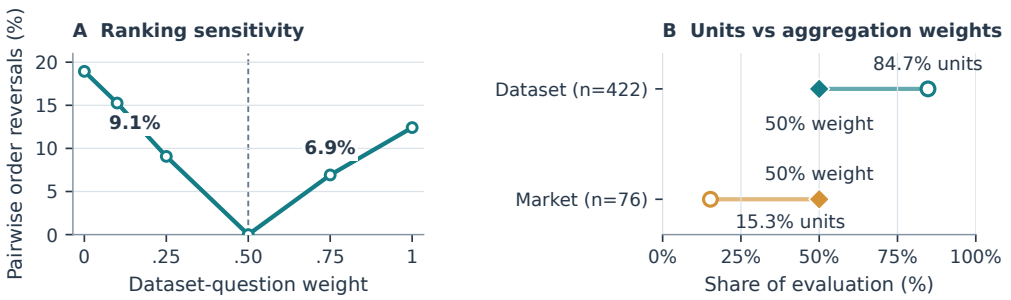


Figure 2: ForecastBench rankings vary with aggregation. Left: reversals among 8,354 untied pairs relative to equal weighting. Right: dataset outcomes are 422/498 units (84.7%); market units are 76/498 (15.3%, 49 resolved and 27 unresolved). Each group receives 50% weight; scores use published rounded components, without rerunning systems.

These within-study contrasts show why gains need matched controls. ForecastQA accuracy rises while class-summed Brier worsens; SCATTER's Pass@16 rises while Pass@1 falls, although the latter evaluates a ten-hypothesis set (Jin et al., 2021; Chang et al., 2026). PRISM's calibration and system-level Brier gaps do not isolate a causal effect (Jin et al., 2026). Appendix G lists the numerical contrasts; Appendices B.1 and F document study-specific systems and evaluation details.

Evidence quality matters: preregistered prompt studies found no reliable primary benefit after multiple-comparison correction, while BLF reports residual leakage (Schoenegger et al., 2025; Murphy, 2026). Other studies trade endpoints: LEAF's augmentation worsens Brier, OpenForecaster's consistency gains coincide with more negation errors, and OpenForecast's closed-choice gains accompany lower horizon F1 (Tan et al., 2026b; Chandak et al., 2025; Wang et al., 2025). ForecastDojo finds research tools lower Brier for all 12 models, but all trail an information-unmatched historical market reference in both Brier and accuracy; the market may draw on evidence outside the agents' archive (Ye et al., 2026a).

The appropriate control follows the claim. Reasoning claims require fixed evidence, calibration claims require a separate calibration split, and system-level claims need a common target cohort, commitment schedule and scoring rule. We classify evidence as matched, qualified or descriptive and preserve unknown protocol fields; Appendix H gives the full comparison rules.

7 Open problems and future research

Prospective cohorts should freeze the event set, evidence cutoffs, resolution rules, and scoring before forecasts are issued. They should retain missing, failed, and unresolved cases, then test frozen systems on later cohorts (Karger et al., 2025; Paleka et al., 2026; Yang et al., 2025). This design tests temporal transfer separately from downstream policy value.

Matched-cohort experiments should vary retrieval, inference, calibration, and update frequency while holding the event set and evidence budget fixed; crowd and market inputs should be separate conditions (Halawi et al., 2024; Jin et al., 2026; Chang et al., 2026). For open-ended outputs, preregister event matching, validate judges, and report coverage and unsupported details alongside probability quality (Wang et al., 2025; Yu et al., 2025). Appendix L gives falsifiable designs.

8 Limitations and conclusion

Forecasting scores are interpretable only under a defined event cohort, information boundary, resolution rule, metric, and aggregation. Our targeted audits show how metric conflicts, limited coverage, and weighting choices bound the supported comparisons; without exhaustive coverage or system reruns, this review offers a source-traced account rather than a universal ranking. Future tests should separate forecast skill from evidence access and decision value (Jin et al., 2021; Chang et al., 2026; ForecastBench team, 2025).

AI-USE DISCLOSURE

AI tools assisted search, extraction, analysis, code, figures and drafting. The numerical literature results are source-reported; the weighting curves are a deterministic secondary analysis, and the coverage example is a synthetic sensitivity construction. No newly resolved real-world forecast is claimed. Human authors must verify the completed paper and its disclosure before submission.

REPRODUCIBILITY STATEMENT

Appendix A records the targeted search procedure, and Appendix I defines the result-level evidence fields and provenance labels. The supplementary archive contains the 53-row evidence ledger, tabular outputs, and the script, source URLs, and hashes for the AI World Cup group-stage recomputation. The 43 literature records are source-reported; the ten recomputed records cover released group-stage predictions only. The original forecasting systems were not rerun, and source papers are not redistributed.

REFERENCES

Ömer Faruk Akgül, Feiyu Zhu, Yuxin Yang, Rajgopal Kannan, and Viktor Prasanna. RECIPE-TKG: From sparse history to structured reasoning for LLM-based temporal knowledge graph completion. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1943–1965, Rabat, Morocco, March 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.86. URL https://aclanthology.org/2026.eacl-long.86/.

Janik-Vasily Benzin and Stefanie Rinderle-Ma. A Survey on Event Prediction Methods from a Systems Perspective: Bringing Together Disparate Research Areas. ACM Computing Surveys, 2025. URL https://arxiv.org/abs/2302.04018.

Federico Bianchi et al. Back to the Future: Evaluating AI Agents on Predicting Future Events. Project article, July 17, 2025. URL https://huggingface.co/blog/futurebench.

Nikos I. Bosse, Peter Mühlbacher, Jack Wildman, Lawrence Phillips, and Dan Schwarz. Automating Forecasting Question Generation and Resolution for AI Evaluation. arXiv:2601.22444, 2026. URL https://arxiv.org/abs/2601.22444.

Henrik Bostrom, Henrik Linusson, and Anders Vesterberg. Mondrian Predictive Systems for Censored Data. In *Proceedings of the Twelfth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 204 of *Proceedings of Machine Learning Research*, pp. 399–412. PMLR, 2023. URL https://proceedings.mlr.press/v204/bostrom23a.html.

Jingyi Chai, Bingyang Zheng, Xiangrui Liu, Hao Lu, Zihang Zhou, Tianchen Wang, Kemeng Zhang, and Siheng Chen. Agentic Time Machine as an Infrastructure for Future-Event Forecasting. arXiv:2606.21013v1; preprint, 2026. URL https://arxiv.org/abs/2606.21013.

Nikhil Chandak, Shashwat Goel, Ameya Prabhu, Moritz Hardt, and Jonas Geiping. Scaling Open-Ended Reasoning to Predict the Future. arXiv:2512.25070, 2025. URL https://arxiv.org/abs/2512.25070.

Nikhil Chandak et al. Curating the Future: A Scalable Recipe for Training Open-Ended Forecasters. ICML 2026 poster; OpenReview forum SiMYGtHfxT, 2026. URL https://openreview.net/forum?id=SiMYGtHfxT.

He Chang, Zhulin Tao, Lifang Yang, Xianglin Huang, and Yunshan Ma. Scattered hypothesis generation for open-ended event forecasting. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 17288–17304, 2026. doi: 10.18653/v1/2026.findings-acl.855. URL https://aclanthology.org/2026.findings-acl.855/.

Pu Cheng, Juncheng Liu, and Yunshen Long. Polybench: Benchmarking llm forecasting and trading capabilities on live prediction market data. arXiv:2604.14199v1, 2026. URL https://arxiv.org/abs/2604.14199v1.

Yizhou Chi, Eric Chamoun, Zifeng Ding, and Andreas Vlachos. Worldreasoner: Evaluating whether language model agents forecast events with valid reasoning. arXiv:2606.11816v1, 2026. URL https://arxiv.org/abs/2606.11816.

Songgaojun Deng, Maarten de Rijke, and Yue Ning. Advances in Human Event Modeling: From Graph Neural Networks to Language Models. KDD tutorial; proceedings article, pp. 6459–6469, 2024. URL https://doi.org/10.1145/3637528.3671466.

ForecastBench team. Archived Human-Cohort Leaderboard. Repository snapshot 601f6d9; retrieved September 16, 2026, 2025. URL https://github.com/forecastingresearch/forecastbench-datasets/tree/601f6d9/leaderboards/csv.

ForecastBench team. ForecastBench: Updated Ranking Methodology. Technical report, 2026. URL https://www.forecastbench.org/assets/pdfs/forecastbench_updated_methodology.pdf.

Tilmann Gneiting and Adrian E. Raftery. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Shashwat Goel, Nikhil Chandak, Arvindh Arun, Ameya Prabhu, Steffen Staab, Moritz Hardt, Maksym Andriushchenko, and Jonas Geiping. Futuresim: Replaying world events to evaluate adaptive agents. arXiv:2605.15188v1, 2026. URL https://arxiv.org/abs/2605.15188.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching Human-Level Forecasting with Language Models. arXiv:2402.18563, 2024. URL https://arxiv.org/abs/2402.18563.

Chengyuan Jin, Daojian Zeng, Kang Liu, Jun Zhao, and Yubo Chen. PRISM: Calibrated Bayesian Fusion and Auditable Attribution for Reliable LLM Event Prediction from Text. In *Proceedings of UAI*, volume 337, pp. 2595–2617. PMLR, 2026. URL https://proceedings.mlr.press/v337/jin26a.html.

Woojeong Jin, Rahul Khanna, Suji Kim, Dong-Ho Lee, Fred Morstatter, Aram Galstyan, and Xiang Ren. ForecastQA: A Question Answering Challenge for Event Forecasting with Temporal Text Data. Archived paper, 2021. URL https://arxiv.org/abs/2005.00792.

Ezra Karger, Houtan Bastani, Yueh-Han Chen, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities. ICLR; proceedings version, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ea74e45a229dac70b5b63b28d8934db6-Paper-Conference.pdf.

Dennis Lee, Hamza Mostafa, and Om Shastri. TimeSeek: Temporal Reliability of Agentic Forecasters. ICLR Workshop on Advances in Financial AI; arXiv:2604.04220v1, 2026a. URL https://arxiv.org/abs/2604.04220.

Jaeho Lee, Nick Merrill, and Ezra Karger. ForecastBench-Sim: A Simulated-World Forecasting Benchmark. arXiv:2606.18686, 2026b. URL https://arxiv.org/abs/2606.18686.

Sang-Woo Lee, Sohee Yang, Donghyun Kwak, and Noah Y. Siegel. Advancing Event Forecasting through Massive Training of Large Language Models: Challenges, Solutions, and Broader Impacts. arXiv position paper arXiv:2507.19477, 2025. URL https://arxiv.org/abs/2507.19477.

Haoxuan Li, He Chang, Yunshan Ma, Yi Bin, Yang Yang, See-Kiong Ng, and Tat-Seng Chua. Thinktank-me: A multi-expert framework for middle east event forecasting. In *Proceedings of the ACM Web Conference 2026*, pp. 8541–8544. ACM, 2026. doi: 10.1145/3774904.3792905. URL https://doi.org/10.1145/3774904.3792905.

Kevin Murphy. Agentic Forecasting using Sequential Bayesian Updating of Linguistic Beliefs. arXiv:2604.18576v4, 2026. URL https://arxiv.org/abs/2604.18576.

Maksym Nechepurenko and Pavel Shuvalov. Foresight Arena: An On-Chain Benchmark for Evaluating AI Forecasting Agents. arXiv:2605.00420v2; version-qualified record, 2026. URL https://arxiv.org/abs/2605.00420.

Yingsong Ning, Fu Zhang, Jingwei Cheng, Jiashun Peng, and Xiaoke Wang. Critic rule induction: Improving temporal knowledge graph forecasting with generator-critic language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 29436–29448, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1471. URL https://aclanthology.org/2026.findings-acl.1471/.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift. In *Advances in Neural Information Processing Systems*, volume 32, pp. 13969–13980, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

Daniel Paleka, Abhimanyu Pallavi Sudhir, Alejandro Alvarez, Vineeth Bhat, Adam Shen, Evan Wang, and Florian Tramèr. Consistency checks for language model forecasters. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=r5IXBlTCGc.

Daniel Paleka, Shashwat Goel, Jonas Geiping, and Florian Tramer. Pitfalls in Evaluating Language Model Forecasters. ICLR, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/b9d69c72e4587ddf3fa6352890aefc4f-Paper-Conference.pdf.

Long Phan, Adam Khoja, Mantas Mazeika, and Dan Hendrycks. LLMs Are Superhuman Forecasters. Technical report, 2024a. URL https://drive.google.com/file/d/1Tc_xY1NM-US4mZ4OpzxrpTudyo1W4KsE/view.

Long Phan, Andrew Zeng, Mantas Mazeika, Adam Khoja, and Dan Hendrycks. Superhuman Automated Forecasting. Center for AI Safety blog, September 8, 2024, 2024b. URL https://safe.ai/blog/forecasting.

Philipp Schoenegger, Cameron R. Jones, Philip E. Tetlock, and Barbara Mellers. Prompt Engineering Large Language Models' Forecasting Capabilities. arXiv:2506.01578, 2025. URL https://arxiv.org/abs/2506.01578.

Jonaid Shianifar and Iias Faiud. AI World Cup 2026: Benchmarking Large Language Models for End-to-End Football Tournament Prediction. arXiv:2608.03416v1, 2026. URL https://arxiv.org/abs/2608.03416v1.

Jing Su, Chufeng Jiang, Xin Jin, Yuxin Qiao, Tingsong Xiao, Hongda Ma, Rong Wei, Zhi Jing, Jiajun Xu, and Junhong Lin. Large Language Models for Forecasting and Anomaly Detection: A Systematic Literature Review. arXiv preprint arXiv:2402.10350, 2024. URL https://arxiv.org/abs/2402.10350.

Mingtian Tan, Mike A. Merrill, Zack Gottesman, Tim Althoff, David Evans, and Thomas Hartvigsen. Inferring events from time series using language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3469–3490, 2026a. doi: 10.18653/v1/2026.acl-long.157. URL https://aclanthology.org/2026.acl-long.157/.

Mingtian Tan, Mihir Parmar, Palash Goyal, Chun-Liang Li, Nanyun Peng, Thomas Hartvigsen, Jinsung Yoon, and Tomas Pfister. LEAF: A Living Benchmark for Event-Augmented Forecasting. arXiv:2605.16358v1; HTML full text, 2026b. URL https://arxiv.org/abs/2605.16358.

Zhengwei Tao et al. PROPHET: An Inferable Future Forecasting Benchmark with Causal Intervened Likelihood Estimation. arXiv:2504.01509v2, 2026. URL https://arxiv.org/abs/2504.01509.

Aditi Tiwari, Aashrith Bandaru, and Heng Ji. Competence-gated pooling of language models and priors for event forecasting. arXiv:2609.12101v1, 2026. URL https://arxiv.org/abs/2609.12101.

Julian Varghese, Lucas Bickmann, and Sarah Sandmann. Information specialization and constrained synthesis in multi-agent llm forecasting: A prospective live-study of the 2026 fifa world cup, 2026. URL https://arxiv.org/abs/2609.12495.

VargheseLab. Fifa2026-pred: Prospective forecasting ledger and evaluation code, 2026. URL https://github.com/VargheseLab/Fifa2026-Pred/tree/ecc62a5c18909408e0c04aecf39d7180b6fcd5c2. Branch update-files-branch, commit ecc62a5c18909408e0c04aecf39d7180b6fcd5c2.

Zhen Wang, Xi Zhou, Yating Yang, Bo Ma, Lei Wang, Rui Dong, and Azmat Anwar. Openforecast: A large-scale open-ended event forecasting dataset. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 5273–5294, 2025. URL https://aclanthology.org/2025.coling-main.353/.

Zhenran Wang, Zhonghan Bian, Jinsong Li, and Zhangyang Qi. Societybench: Forecasting counterfactual social-world evolution. arXiv:2608.04009v2, 2026. URL https://arxiv.org/abs/2608.04009.

Muyan Weng, Defu Cao, Wei Yang, Yashaswi Sharma, and Yan Liu. TemporalBench: A Benchmark for Evaluating LLM-Based Agents on Contextual and Event-Informed Time Series Tasks. arXiv:2602.13272; adjacent task, 2026. URL https://arxiv.org/abs/2602.13272.

Xiaogang Xu, Jiaqi Tang, Jianmin Chen, et al. LLM-based Agents for Forecasting and Prediction: Methods, Training, Evaluation, and Applications. arXiv preprint arXiv:2608.23058v1, posted 24 August 2026, 2026. URL https://arxiv.org/abs/2608.23058.

Fan Yang and Rina Foygel Barber. Contraction and uniform convergence of isotonic regression. arXiv:1706.01852v3, 2018. URL https://arxiv.org/abs/1706.01852v3.

Qingchuan Yang, Simon Mahns, Sida Li, Anri Gu, Jibang Wu, and Haifeng Xu. LLM-as-a-Prophet: Understanding Predictive Intelligence with Prophet Arena. Numerical source: arXiv v2, December 2025; later ICLR 2026 paper, 2025. URL https://arxiv.org/abs/2510.17638.

Chenchen Ye, Ziniu Hu, Yihe Deng, Zijie Huang, Mingyu Derek Ma, Yanqiao Zhu, and Wei Wang. MIRAI: Evaluating LLM Agents for Event Forecasting. arXiv:2407.01231, 2024. URL https://arxiv.org/abs/2407.01231.

Liqin Ye et al. Forecast-Dojo: Replayable Environments for Benchmarking and Training LLM Forecasting Agents. arXiv:2609.28876v1; preprint, 2026a. URL https://arxiv.org/abs/2609.28876.

Xiao Ye, Jacob Dineen, Evan Zhu, Shijie Lu, Kevin Song, and Ben Zhou. Hindcast: Replaying Prediction Markets to Evaluate LLM Forecasters. arXiv:2607.14051v1; preprint, 2026b. URL https://arxiv.org/abs/2607.14051.

Zi Yu, Shaoxiang Wang, Guozheng Li, Yu Zhang, and Chi Harold Liu. ForestCast: Open-Ended Event Forecasting with Semantic News Forest. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 12667–12681, 2025. doi: 10.18653/v1/2025.findings-emnlp.678. URL https://aclanthology.org/2025.findings-emnlp.678/.

Zhiyuan Zeng, Jiashuo Liu, et al. FutureX: An Advanced Live Benchmark for LLM Agents in Future Prediction. Technical report; arXiv:2508.11987v3, 2025. URL https://arxiv.org/abs/2508.11987.

Jaden Zhang, Gardenia Liu, Oliver Johansson, Hileamlak Yitayew, Kamryn Ohly, and Grace Li. Prediction arena: Benchmarking ai models on real-world prediction markets. arXiv:2604.07355v1, 2026. URL https://arxiv.org/abs/2604.07355.

Meiru Zhang, Auss Abbood, Zaiqiao Meng, and Nigel Collier. The power of simplicity in llm-based event forecasting. In *Proceedings of the 1st Workshop for Research on Agent Language Models*, pp. 454–470, 2025. doi: 10.18653/v1/2025.realm-1.32. URL https://aclanthology.org/2025.realm-1.32/.

Liang Zhao. Event Prediction in the Big Data Era: A Systematic Survey. *ACM Computing Surveys*, 54(5):1–37, 2021. doi: 10.1145/3450287. URL https://doi.org/10.1145/3450287.

Andy Zou, Tristan Xiao, Ryan Jia, et al. Forecasting Future World Events with Neural Networks. arXiv:2206.15474, 2022. URL https://arxiv.org/abs/2206.15474.

A Search and verification history

GAMETime (Tan et al., 2026a) asks models to recover missing event descriptions from numerical time series. Its Sections 3.1–3.2 supply the numerical values at the missing-event timestamps; the prompt can also include surrounding event descriptions. This differs from committing a forecast before those observations arrive. We retain it as adjacent event understanding, without merging its accuracy into prospective forecasting comparisons. The relevant boundary is the relation between the prediction time and the supplied evidence, not whether the output describes an event or the dataset can be refreshed.

The current corpus was assembled from core benchmark papers, related surveys, their references and a targeted search of arXiv, ICLR proceedings, OpenReview and ACL Anthology. The follow-up searches on 16–17 September 2026 used combinations of "LLM event forecasting survey," "forecasting benchmark," "probability forecasting" and "live evaluation." The earlier query snapshot and a dated follow-up inclusion ledger are retained. The follow-up adds the UAI 2026 PRISM paper and full-text LEAF evidence; TemporalBench is logged as an adjacent numerical-forecasting benchmark. A further primary-source search added SCATTER (ACL Findings 2026) and a REALM 2025 workshop comparison of retrieval and ReAct. The 2026 AI World Cup preprint is retained as a narrow prospective tournament case because its path-weighted total and group-match accuracy produce different leaders; its single run per submitted system and composite points keep it separate from proper-probability comparisons. GAMETime remains an adjacent event-understanding task because the supplied series includes observations at the timestamps of the events to be inferred. The 19 September follow-up adds competence-gated pooling and checks the public FutureX scoring implementation against the retained paper version. ForecastQA result conventions and human sampling are checked against the original table and methods. Subsequent follow-up examines the COLING 2025 OpenForecast evaluation protocol and the PolyBench preprint, separating reference-answer verification from forecaster retrieval and execution-based returns from probability scores. LEAF's archived prompt and the public repository state are checked separately.

The 26 September follow-up verified Xu et al.'s closest agent-focused survey as arXiv v1 (posted 24 August 2026): it covers forecasting-agent methods, training, evaluation and applications across numerical and event targets. We retain this overlap explicitly and distinguish the present review by its result-level claim audits under evaluation contracts, rather than an architecture taxonomy. The same follow-up adds Forecast-Dojo, a 24 September arXiv preprint on replayable longitudinal forecasting; it is included as a benchmark contribution, not a review. The ledger remains a bounded screening record, not the total number of retrieved studies. Preprints are included with their status distinguished from conference publications. A workshop paper or project article is not silently counted as a main-conference result.

Xu et al. (2026) already define forecasts relative to information admissible before the decision time, distinguish feature from parametric leakage, and avoid cross-paper leaderboards when protocol reporting is incomplete. Paleka et al. (2026) separate trust in retrospective evaluation from extrapolation to real-world use, documenting bias in questions generated from post-outcome news, unreliable date-filtered retrieval, possible access to human or market forecasts, and contest incentives that favor correlated risk.

Benzin & Rinderle-Ma (2025) treat uncertainty quantification, probabilistic quality and data quality as requirements for event-prediction systems. In their review of 260 papers, fewer than 5% evaluate predicted probabilities with an appropriate scoring rule (Benzin & Rinderle-Ma, 2025). We take

these requirements as prior art; the remaining question is how specific source protocols shape the interpretation of reported comparisons.

These cases show why a reported score needs its protocol attached before it can support a comparison. ForecastQA reports human accuracy on 150 randomly sampled test questions and permits annotators to skip questions when they cannot set aside knowledge of the outcome; this does not establish a paired human–model gap over the full test set (Jin et al., 2021). Halawi et al.'s matched base-GPT-4 ablation is more informative about retrieval (Brier loss 0.186 with retrieval and 0.206 without) than the full-system score of 0.179, which also uses a different model mixture and fine-tuning; its losses are averaged over retrieval dates within each of 914 questions (Halawi et al., 2024). ForecastBench, in turn, defines its headline score as the equal-weighted mean of dataset and market Brier scores, so its ranking answers a particular aggregation question (Karger et al., 2025). Our evidence map records these source conditions alongside the results and labels deterministic reweighting of the archived, rounded components as a secondary analysis (ForecastBench team, 2025). It therefore makes explicit which comparisons are supported, restricted or unresolved, leaving extraction accuracy, coverage and reproducibility as the tests of its value.

B Detailed mechanism comparisons

B.1 From a pipeline description to a mechanism comparison

Autocast makes the distinction between static retrieval and forecast updating explicit (Zou et al., 2022). FiD Static retrieves ten articles across the available period; FiD Temporal processes one article per day and uses intermediate crowd forecasts as auxiliary training targets. Its Table 2 averages combined scores across different model sizes: the temporal family averages 37.8 versus 37.2 for the static family, while the largest static model scores 40.6 versus 40.1 for the largest temporal model. These aggregates do not establish a uniform benefit from temporal supervision. The comparison also changes architecture and parameter counts: the static family uses T5, while the temporal family uses GPT-2 with a T5 encoder. It therefore does not isolate the effect of temporal updating. The crowd baseline is the final aggregate before closing, rather than a forecast at a common early horizon. The combined score also mixes binary accuracy, multiple-choice accuracy and numerical error; it is not a probability loss.

An architecture label such as retrieval-augmented forecasting can hide several interventions. Halawi et al.'s system generates queries, retrieves and filters historical news, and summarizes the selected articles; its training pipeline keeps candidate forecasts that outperform the crowd, while its final predictor combines six base- and fine-tuned-model forecasts with a validation-selected trimmed mean (Halawi et al., 2024). The end-to-end system therefore changes the evidence supplied to the forecaster, the learned mapping from evidence to probabilities, and the aggregation of multiple outputs. Its reported gain cannot identify which change produced the improvement. A retrieval comparison must hold the forecaster fixed, a reasoning comparison must provide the same articles, and an aggregation comparison must use the same questions and component forecasts. Without these controls, the result supports a system-level comparison rather than attribution to a particular module.

OpenForecaster addresses one source of temporal leakage by using monthly CommonCrawl News snapshots for both question generation and retrieval, rather than a live search index whose documents or rankings may change (Chandak et al., 2025). Its pipeline generates questions from news, filters and rewrites them to remove answer cues, trains on events through the Qwen3 release in April 2025, and evaluates on later resolutions from May to August. This makes the document and model time boundaries inspectable, but leaves separate evaluation choices: an LLM judges semantic equivalence between free-form answers and outcomes, while reinforcement learning combines answer correctness with a confidence-sensitive reward adapted from the Brier score (Chandak et al., 2025). The resulting gain therefore depends both on the dated evidence available to the forecaster and on how generated questions, answer matches and confidence are scored. We treat the snapshot corpus as a concrete temporal control, not as a substitute for auditing those labels and reward definitions.

Taken together, these studies are not successive versions of one forecasting agent. Autocast studies predictions that update as daily evidence arrives; Halawi et al. combine retrieved news with

performance-selected reasoning and an ensemble; OpenForecaster changes the training objective while using a dated news corpus (Zou et al., 2022; Halawi et al., 2024; Chandak et al., 2025). A score from any one system describes that configured package under its own test protocol, not a cross-paper ranking of which intervention is strongest. The useful synthesis is to trace which change is actually compared within each study and which other design choices move with it before interpreting the reported gain.

B.2 Calibration and incremental predictive value

PRISM's Table 1 separates calibration from the larger system comparison. Isotonic calibration lowers direct Llama-3-70B's Brier loss from 0.094 to 0.070 and ECE from 0.176 to 0.072, while AUROC remains 0.742; the two-stage Dawid–Skene/Bayesian baseline scores 0.067 Brier, and full PRISM scores 0.062 with ECE 0.058 (Jin et al., 2026). The paper marks the full-system comparison with direct LLM forecasting as significant under its paired country-level procedure, but these rows do not identify the causal contribution of every component. PRISM separately reports 90.8% mean empirical coverage at a 90% conformal target, with six-month rolling coverage from 86.9% to 93.5% and mean set size 1.38. The authors present this as an empirical diagnostic under temporal dependence, not a finite-sample distribution-free guarantee. Their factor contrasts are controlled predictive associations; an intervention-style interpretation requires additional identification assumptions (Jin et al., 2026).

Tiwari et al. ask whether a language model adds value to a forecast already available (Tiwari et al., 2026). Their competence gate lowers Brier loss from 0.0771 to 0.0732 across 2,357 resolved questions, but the main external baseline is source-dependent: it retains market or crowd probabilities where available and uses training-fold outcome rates for structured questions. On 1,294 official ForecastBench market questions, the gate does not significantly improve over the market forecast ($p = 0.42$), nor does it significantly outperform calibrated domain-aware logistic stacking ($p = 0.53$). These tests do not establish equivalence; moreover, the paired bootstrap resamples questions and does not explicitly model dependence among related events. The paper's historical results therefore support selective complementarity in the studied cohorts, not a universal routing gain. The authors acknowledge residual contamination and identify prospective evaluation as stronger evidence.

B.3 Tool interfaces and selection of generated forecasts

The clearest interface comparison in MIRAI is its balanced 100-query subset (Ye et al., 2024). GPT-3.5-turbo scores 27.9 versus 14.4 second-level relation F1 with Single Function and Code Block actions, respectively, while GPT-4o scores 32.6 versus 32.8. This descriptive reversal does not establish a model-by-interface interaction. On the full test set, GPT-3.5-turbo scores 29.6 versus 12.6, but those figures cannot be directly compared with GPT-4o's subset results because the query populations differ. The paper also reports more execution errors with Code Block across models and fewer errors for GPT-4o, making code generation a plausible mechanism; the experiments do not isolate error reduction as the cause of the interface contrast.

For Mistral-7B, MIRAI contrasts self-consistency, which retains entries repeated in more than two samples, with Max@K, which selects each query's highest-F1 output after comparing it with the realized label (Ye et al., 2024). Max@10 reaches 33.9 F1, but this is a per-query oracle score among sampled outputs, not the performance of a deployable inference-time selector. It therefore measures how often sampling produces a strong candidate, not whether a system can identify that candidate before resolution.

A 2025 REALM workshop study tests a fixed RAG pipeline against iterative ReAct on MIRAI (Zhang et al., 2025). With GPT-4o-mini and the same retrieved information, RAG reaches comparable or higher F1 at about one tenth of the token cost. Across backbones, structured relation distributions and event-code histories generally help, whereas article titles alone often hurt; larger models can benefit when semantic context is combined with structured inputs. The result is a conditional cost–F1 trade-off for this benchmark, not evidence that iterative agents are generally unnecessary.

B.4 HYPOTHESIS COVERAGE AND COMPLEMENTARY INFORMATION

SCATTER trains a forecaster to generate sets of plausible hypotheses, rewarding semantic validity together with within-set and across-set diversity (Chang et al., 2026). On OpenForecast, Qwen2.5-3B raises GPT-5o-mini-judged Pass@16 from 25.66 with standard GRPO to 30.26, while mean Pass@1 falls from 11.84 to 8.55. Pass@K counts problems for which at least one hypothesis is judged correct across K sampled responses; each response contains multiple hypotheses, and the reported Pass@1 mean averages single-response scores over 16 repeated rounds. The result therefore indicates a trade-off between success within one sampled set and coverage accumulated through repeated sampling, not a change in single-answer accuracy or calibrated probabilities. SCATTER also reports embedding-based SoftPass, SoftRecall and valid-hypothesis ratio, whose values depend on similarity thresholds, embedding models and filtering rules; cross-embedding and threshold-sensitivity checks probe, but do not remove, that dependence. Comparisons of hypothesis-generation methods should consequently match the number of hypotheses and sampling rounds, and keep set coverage distinct from probability quality.

Varghese et al. prospectively test four role-specialized agents, all using one model, on the final 56 FIFA World Cup matches (Varghese et al., 2026). The news specialist has the highest mean probability-weighted Top-3 utility (1.066, versus 1.033 for the market), while both record 23 exact Top-3 hits; the market leads on the Top-1 measures, and the synthesis agent does not exceed the strongest specialist on any summary metric. These are descriptive comparisons: the authors report no inferential tests for the prespecified tournament cohort. Their utility weights probabilities over three scorelines using UEFA points, so it is not a proper loss over the full score distribution.

The study also finds substantial overlap: the two specialists share at least two of their three scorelines in 50 matches, and the synthesis agent retains at least two scorelines from their combined set in all 56 matches, including all three in 39. This single architecture therefore shows that assigning distinct information roles need not yield distinct predictions, while leaving shared model priors, evidence and synthesis instructions entangled.

A read-only inspection of the cited versioned ledger found a concrete record-level contamination: the row labelled `Columbia-Portugal` (the Colombia–Portugal fixture) contains the Panama–England forecast in its Agent B field, which the same row's meta-agent text says it discarded as irrelevant (VargheseLab, 2026). This does not by itself invalidate the paper's reported aggregates, but it means the ledger still requires systematic row-level auditing before those aggregates can be independently reproduced. The paper also enforces its exclusion of betting information through prompts rather than technical controls. We treat this September 2026 arXiv preprint as bounded evidence about one prospective system, not as evidence that multi-agent forecasting generally fails.

Table 3: Mechanisms, retained evidence and the next useful comparison. Evidence describes the cited studies; the final column gives our recommended controls. The recommended controls identify attribution limits, rather than asserting an absence of prior work.

Intervention	Evidence retained here	Comparison needed for attribution
Retrieval (Halawi et al., 2024)	Same base model has lower Brier loss with retrieval	Preserve issue time and dated corpus; estimate paired uncertainty
Decomposition (Chai et al., 2026)	Planner–solver system exceeds self-consistency; components change jointly	Isolate planning with matched evidence and output budgets
Aggregation (Halawi et al., 2024)	Combined forecasts also contain crowd information	Same questions; distinguish model-only from crowd-assisted forecasts
Forecast training (Chang et al., 2026)	Coverage gains coexist with lower one-round set success	Same inference budget; held-out events and reward-independent scores
Calibration (Jin et al., 2026)	Calibrated direct forecasts narrow the descriptive system gap	Separate calibration data; report discrimination too
Repeated updating (Zou et al., 2022)	Temporal family mean and largest-model contrasts differ	Matched event–horizon cohorts; separate temporal supervision from updating

B.5 SPECIALIZATION, CONSISTENCY AND TASK TRANSFER

ThinkTank-ME frames forecasting as predicting a missing object in a timestamped event relation and fine-tunes country-specific Llama-3.1-8B experts on POLECAT-FOR-ME; its authors set the test partition with the model's December 2023 knowledge cutoff in mind (Li et al., 2026). Table 1 reports 24.6%/25.0% micro/macro accuracy for the trained-router Elite Ensemble with Weighted Best-of-N, compared with 22.7%/22.6% for all-data training and 18.6%/19.4% for Expert Routing alone. The figures support the combined specialization-and-aggregation system on this entity-prediction task, but do not isolate the contribution of expert diversity, top-k selection or generation-probability-weighted aggregation. The table reports accuracy point estimates, not calibrated event probabilities, and does not give an inferential test for these contrasts.

Published in WWW '26, ThinkTank-ME adds evidence on country-conditioned specialization for structured geopolitical relations, not on general gains from adding agents (Li et al., 2026).

Paleka et al. (2025) ask whether logically related forecasts agree before their outcomes resolve. In their Table 1, the cross-forecaster correlation between Brier loss and arbitrage violations is 0.92 for conditional-probability checks on scraped questions, but -0.36 for negation checks on NewsAPI questions. These are not matched estimates: both the check and the question corpus change. The analysis also excludes forecasters with Brier loss above 0.25, including a basic Llama-3.1-8B setup, so the association is conditional on a selected set of forecasters rather than evidence that consistency predicts accuracy across the full model population. The authors themselves note that model capability may drive both quantities.

Their intervention results make a separate point. Recursive ArbitrageForecaster configurations improve consistency on the checks they target, but the study reports no reliable gains in resolved-outcome forecasting or in untargeted checks. The intervention covers only a small subset of checks and recursion settings, in part because a full configuration would require 49 model calls per question at each depth. The evidence therefore supports a narrow mechanism claim: targeted consistency can be improved without demonstrating a general improvement in prediction. Correlation and intervention require separate evaluation, with the forecaster population, targeted checks and call budget stated explicitly.

OpenForecast's full test-set results show a sharp task-dependent trade-off after multi-task fine-tuning Llama3-8b (Wang et al., 2025). Table 2 reports MCNC accuracy rising from 55.7% to 96.1%, while short- and long-term forecasting F1 each fall to 22.9%, from 56.8% and 49.5%, respectively; argument-question accuracy changes more modestly, from 63.0% to 61.5%. The list-forecasting scores use LRAE's modified metric, which credits additional predictions judged true from retrieved evidence or reasoning, rather than probability scores. The contrast supports a limited conclusion: gains on constrained answer choices need not transfer to open-ended event coverage. The authors discuss label incompleteness and task imbalance as possible explanations, but this comparison does not isolate either mechanism or show that fine-tuning generally harms forecasting.

The paper's separate human-validation study samples 100 questions per open-ended task, collects forecasts from nine models, and filters cases where the two annotator groups' F1 scores differ by more than 0.1 before adjudication. This protocol is used to compare LRAE with human judgments, not to define the full test-set Table 2 contrast. Keeping the two populations distinct matters: the filtered sample supports a human-alignment check for the evaluator, but it does not establish that the reported fine-tuning trade-off has the same size on every test question.

C HUMAN REFERENCE PROTOCOLS AND SOURCE SNAPSHOTS

C.1 FORECASTBENCH: INFORMATION AND AGGREGATION BOTH MATTER

The ICLR 2025 version of ForecastBench provides a particularly instructive example (Karger et al., 2025). Its Table 2 reports 0.096 for the superforecaster median, 0.121 for the public median and 0.122 for Claude 3.5 Sonnet with scratchpad and freeze-value inputs. The table concerns the human question set and 498 evaluated units across specified horizons, rather than the full LLM-only cohort. Its aggregate equally weights the dataset-question mean and market-question mean. Unresolved market questions are compared to a community aggregate, whereas resolved ones use the realized outcome.

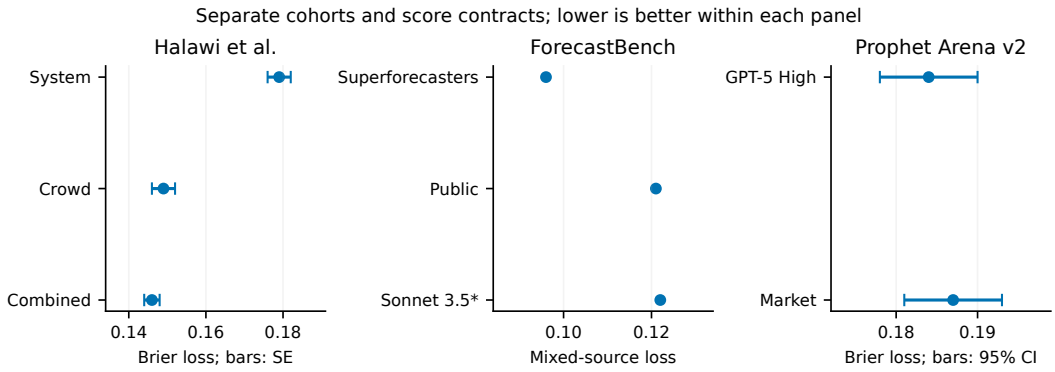


Figure 3: Selected source-reported results remain in separate experimental settings. Halawi error bars are reported standard errors; Prophet Arena bars are reported 95% interval half-widths. ForecastBench uses its published mixed-source aggregate and no interval bars here; the marked Sonnet 3.5 configuration uses scratchpad and freeze-value inputs. Prophet uses GPT-5 with High reasoning. The panels are not a pooled leaderboard and their axes do not define a common evaluation population.

In the ICLR 2025 protocol, 500 public participants recruited through Prolific and Facebook each answered 20 randomly assigned questions in a one-hour survey. The separate superforecaster cohort comprised 39 participants who forecasted 20 questions over nine days; halfway through, they entered a group stage where they could see one another's forecasts and rationales and discuss the questions. These protocols differ in cohort, time, and collaboration, so their score gap does not isolate expertise. The subgroup tables also show that the aggregate human advantage is not uniform: the retained Sonnet configuration has lower reported loss on environment/energy (0.126 versus 0.141; 81 evaluated units) and at the 90-day horizon (0.151 versus 0.172; 110 units). These subgroup comparisons are descriptive, without a subgroup significance claim (Karger et al., 2025).

The best model settings in this table can receive market crowd probabilities through freeze values. Their reported performance therefore combines inference with a strong supplied signal. This is useful end-to-end evidence, but it should not be restated as independent prediction without that signal. The paper also states that its human surveys did not directly elicit combination-question forecasts: the human reference in that comparison assumes independence between component forecasts. It is therefore a constructed reference, not directly measured human performance on the combined questions. The later combination-question table and updated leaderboard methodology introduce other constructions (ForecastBench team, 2026); their numbers should not replace the original table without a version and cohort explanation.

C.2 Halawi: isolate retrieval and preserve the question-level unit

Retrieval ablations and ensemble results answer different questions. The former compare access to evidence; the latter describe how several estimates become one forecast. Their interpretation requires keeping the intervention separate from the unit over which forecast errors are averaged.

Halawi et al.'s Table 5 reports Brier loss 0.186 for base GPT-4 with retrieval and 0.206 without retrieval, each with reported standard error 0.002 (Halawi et al., 2024). This is a more informative retrieval comparison than the full-system score of 0.179, which also changes the reasoning-model mixture and fine-tuning. Marginal standard errors do not determine the uncertainty of the paired difference. The retrieval pipeline combines direct queries with queries generated from subquestions, but the inspected ablation does not isolate this decomposition step. Its benefit should not be inferred from the full retrieval contrast.

Halawi et al.'s system combines three GPT-4-1106-Preview forecasts with three forecasts from a fine-tuned GPT-4-0613 using a trimmed mean (Halawi et al., 2024). These are components of a prediction, not independent test questions. The full test split contains 914 binary questions. Section 3.2 first averages losses across retained retrieval dates within each question, then averages across

questions. Pooling all question–date forecasts instead would weight questions by their number of retained dates and can change the reported quantity. The system-only ensemble must also remain distinct from the separately reported combination with crowd forecasts. Recording these levels prevents the number of model calls, issued predictions and evaluated questions from being treated as interchangeable sample sizes.

D TWO-CONFIGURATION WEIGHTING DERIVATION

Both configurations receive freeze values in ForecastBench Table 2. Using its rounded components, let w be the weight on dataset questions. Sonnet 3.5 with scratchpad has component scores 0.138 and 0.107; GPT-4 Turbo with zero-shot prompting has 0.162 and 0.095. Their reconstructed aggregates are

$$S_{\text{Sonnet}}(w) = 0.107 + 0.031w, \qquad S_{\text{Turbo}}(w) = 0.095 + 0.067w. \tag{4}$$

Equation 4 crosses near $w = 1/3$. At the published equal weighting, the Sonnet configuration has the lower score; a sufficiently market-heavy weighting reverses that ordering. Figure 4 is our secondary analysis of published rounded summaries. It is not a new evaluation of raw forecasts, a claim that the official weighting is wrong, or evidence of statistical significance between the systems. It demonstrates why the aggregation population belongs in the result contract.

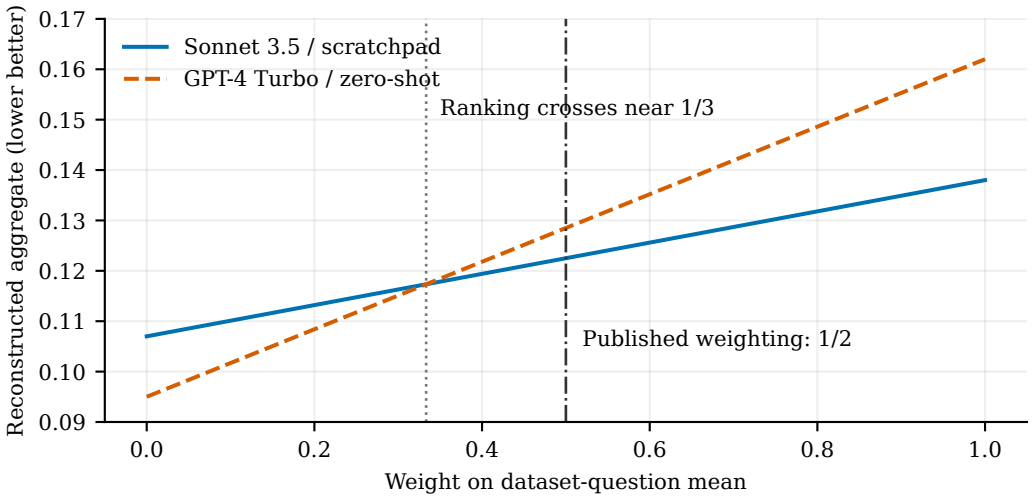


Figure 4: Changing the dataset–market weight changes the ordering of two reported systems. Curves reconstruct aggregates from rounded ForecastBench Table 2 components; the vertical lines mark the approximate crossing and the official equal weighting. No underlying model predictions are rerun or changed.

D.1 SENSITIVITY TO ROUNDED COMPONENT SCORES

The archived components have three decimal places. Conditional on rounding to the nearest 0.001, each component differs from its unrounded value by at most 0.0005. A convex weighted average has the same bound, and a difference between two configurations has error at most 0.001. Therefore, an observed reversal is guaranteed to persist under this rounding model if both the equal-weight difference and the changed-weight difference have opposite signs and magnitudes strictly above 0.001.

Of the 758 observed reversals at dataset weight 0.25, 613 satisfy this sufficient condition; at weight 0.75, 452 of 578 do. The corresponding counts at weights 0 and 1 are 1,444 of 1,582 and 927 of 1,037. Thus rounding can leave individual near-ties unresolved without explaining all observed sensitivity. Uncertified pairs are not necessarily rounding artifacts: the separate bounds are conservative and do not optimize the shared component errors across weights. The original denominator remains

8,354 non-tied pairs under the reconstructed baseline. These are deterministic bounds under the stated rounding assumption, not confidence intervals, sampling-uncertainty estimates or guarantees under a different score-export convention.

An independent implementation represents the published components as integer thousandths and computes ordering signs without floating-point tolerances. It reproduces all reported reversal counts and sufficient-bound counts. This checks the arithmetic of the archived-score analysis; it neither reconstructs the original forecasts nor establishes statistical uncertainty over future events.

At dataset weights 0.25 and 0.75, respectively, 129 and 125 of the 130 archived configurations participate in at least one strict reversal relative to equal weighting. Thus the pair counts are not driven by a single exceptional configuration. This is a descriptive participation count: multiple rows can share the same underlying model with different prompts or information settings, and one reversal suffices for inclusion. It does not imply that most model families differ significantly or that every configuration's rank changes materially.

E Coverage-conditioned scores: a synthetic sensitivity example

FutureX permits model-specific question coverage and examines random prediction deletion (Zeng et al., 2025). Random deletion can quantify sampling variability under its assumptions, but selective nonresponse changes the population being scored. Let $S \in [0, 1]$ be a higher-is-better score defined on the intended population, and let A indicate whether a prediction is observed. For coverage $q = \Pr(A = 1) > 0$, the elementary identity

$$\mathbb{E}[S \mid A = 1] - \mathbb{E}[S] = \frac{\mathrm{Cov}(S, A)}{q} \tag{5}$$

separates a population shift from sampling error. Increasing the number of events does not force the covariance to vanish. This is a sensitivity argument, not a claim that the missingness process in any cited benchmark has been identified. When an unanswered question has no well-defined counterfactual forecast, the potential full-population score itself requires an explicit intervention or abstention rule.

An exact synthetic example illustrates the distinction. Consider equally frequent easy and hard strata, on which system A has expected scores 0.9 and 0.3. It answers every easy question and one quarter of hard questions, independently of score within each stratum. Its intended-population mean is 0.60, its coverage is 0.625, and its answered-question mean is 0.78. A fully answering system B with score 0.65 is better on the intended population but appears worse when compared with A's answered-question mean. These numbers are constructed assumptions, verified with rational arithmetic; they are not extracted results, newly resolved forecasts, or estimated effects for FutureX. The example establishes possibility rather than prevalence.

If only coverage q and the answered-question mean μ_{obs} are known, bounded scores imply the identification interval $[q\mu_{\mathrm{obs}},\ q\mu_{\mathrm{obs}} + 1 - q]$ for a well-defined full-population score. For A this is $[0.4875, 0.8625]$, which does not determine its ordering relative to B. The interval is a worst-case bound, not a confidence interval. Reporting coverage, task-conditioned nonresponse and a common-cohort comparison therefore complements the conditional score. The common cohort still selects events that both systems answer; it does not recover population performance. An abstention-aware utility can instead define the cost of missing predictions prospectively, provided that choice is distinguished from the original forecasting score.

F Detailed benchmark result contracts

F.1 Calibration, aggregation and online feedback

The PRISM comparison in Section B.1 concerns next-month state-based violence in country-month units constructed from UCDP and GDELT (Jin et al., 2026). Its reported variations use country-level bootstrap resampling. They do not constitute a ranking against general forecasting agents. The reproducibility appendix describes train/calibration/test separation but does not give exact calendar boundaries in the inspected passage, leaving temporal replication incomplete in our ledger.

The source distinguishes the aggregation used for main-table performance from the contrast used for paired inference. Its metric appendix pools country-month instances for the main tables, whereas its uncertainty protocol first computes within-country method differences and then averages those differences over resampled countries (Jin et al., 2026). With unequal country support, the latter need not equal the difference between pooled scores. The table's marginal bootstrap deviations also cannot supply a paired interval for every baseline contrast. We therefore retain the descriptive calibration comparison without inferring its statistical significance from those deviations. This distinction concerns the quantity being estimated; it does not imply that either aggregation is inherently incorrect.

Temporal evaluation also requires specifying when feedback becomes available. PRISM's algorithm permits rolling conformal prediction to incorporate earlier test labels once their outcomes are observed, while model and calibration choices follow the stated split protocol (Jin et al., 2026). This is a prescribed online update rule, not by itself evidence of selecting a method after inspecting test performance. A forecasting comparison should therefore record both the parameter-selection boundary and the schedule for incorporating newly resolved events. Treating every use of an earlier outcome as leakage would exclude legitimate deployment behavior; leaving that feedback schedule unspecified would make adaptive and frozen systems difficult to compare.

F.2 Forecast horizon and outcome-conditioned construction

LEAF's Table 8 reports event Brier scores at six time-to-resolution cutoffs (Tan et al., 2026b). For Gemini 3.1 Pro, the event-augmented score changes from 0.457 at 100% remaining time to 0.042 at 10%; corresponding no-event values are 0.602 and 0.622. These are horizon-dependent, source-reported measurements. The domain description specifies 30 curated Polymarket tasks, with a February 2025–February 2026 collection span in Table 2; this span does not establish a training/test split. Table 8 presents point estimates without intervals. Its six cutoffs agree with Section 4.1, whereas the Table 2 caption describes ten. Exact cross-event weights, missing-response handling and per-cell event identities remain unverified, so the tabulated averages should not be treated as a fully reproduced evaluation.

Section 4.1 defines a squared probability error and describes eliciting the probability of the realized outcome. Figure 11 clarifies that the target outcome is selected as the eventual Polymarket winner, while the user template asks for its probability at the cutoff without explicitly stating which outcome won. Outcome-conditioned target selection is different from committing to a proposition before resolution and evaluating both positive and negative realizations. We retain the source score label without pooling these values with fixed-proposition benchmarks; this distinction does not establish that the model was told the winner or that all LEAF tasks leak outcomes. The linked repository at inspected commit `08fc347` contains only a README announcing forthcoming code and data, so the executed scoring implementation cannot be checked from that artifact.

Improvement is not universal: Claude 4.6 Sonnet at 80% remaining time changes from 0.460 without events to 0.476 with them. LEAF's separate sentiment-consistency diagnostic constructs context using the known future trajectory (Section 3.2). That diagnostic tests context sensitivity under outcome-conditioned construction; it must not be represented as a deployable forecast condition or used to label the entire benchmark as leaking.

F.3 ForecastQA: accuracy and probability quality can disagree

ForecastQA provides an early within-benchmark example of metric-dependent conclusions (Jin et al., 2021). In Table 3, closed-book BERT-Large achieves 57.6% accuracy and 0.543 Brier loss, whereas the integrated retrieval model reaches 60.1% accuracy but 0.589 Brier loss. The integrated configuration improves answer accuracy while worsening the reported probability loss; this configuration-level contrast does not isolate a causal effect of retrieval. The paper sums squared errors over two or four answer classes, with range $[0, 2]$; these values cannot be pooled directly with single-probability binary Brier losses. Its human comparison uses 150 sampled questions, asks participants to avoid remembered future information, and reports an average 28.3% skip rate. The reported 71.2% accuracy uses the top-ten retrieved articles, whereas 79.4% permits time-restricted Google Search; both aggregate answers by majority vote. These are retrospective, filtered protocols.

They are not a prospective same-cohort human–model gap. This example motivates retaining both the scoring convention and the human sampling procedure in a result-level comparison.

ForecastQA Table 2 lists 1,092 test questions: 584 binary and 508 four-choice questions (Jin et al., 2021). The human study starts from a sample of 150 questions and permits skipping. Its average annotator skip rate does not determine the denominator of the final majority-vote score. Consequently, neither subtracting the two reported accuracies nor normalizing their numerical ranges creates a paired human–model comparison. The source populations and treatment of unanswered questions must be aligned first.

F.4 AUTOCAST: QUESTION INSTANCES AND INDEPENDENT EVENTS

Autocast's Appendix B.1 distinguishes 6,707 original questions from the expanded dataset after adding negated true/false questions (Zou et al., 2022). Its baseline experiments use 4,411 resolved training examples and 1,292 resolved test examples; Table 5 divides the test set into 775 true/false, 176 multiple-choice, and 341 numerical questions. These are question instances after balancing, not necessarily independent underlying events. Consequently, a follow-up uncertainty analysis should retain links between original and negated questions rather than assume every row is independent. The Table 2 *Average* aggregates across model sizes, not random seeds; it supplies no estimate of run-to-run uncertainty. Auxiliary training on unresolved questions is a separate possible setting, which the reported baselines do not use.

F.5 FUTUREX: COMPOSITE SCORES AND UNMATCHED COVERAGE

FutureX combines four task-specific scores with tier weights 0.1, 0.2, 0.3, and 0.4 (Zeng et al., 2025). Its retained July 20–August 3, 2025 snapshot ranks Grok-4 above Gemini-2.5-flash Deep Research and GPT-o4-mini Think&Search. This is a weighted composite, not a common accuracy or proper probability loss. The experiment roster is broader than the overall leaderboard: the paper excludes AgentOrchestra from the aggregate because it was tested only on the easier tiers. Its absence cannot be interpreted as a low overall score. The source figure has no exact numeric labels, so we retain its reported ordering without inventing precise bar values. Section 3.4.3 specifies exact-match, set-F1, rank-overlap, and volatility-normalized numerical scores. The earlier multi-choice description differs from the later F1 formula. The public evaluator at commit `2626f7a` uses F1 for multi-label ground truth and exact match for singleton ground truth. For example, predicting only A when the truth is {A,B} receives $2/3$, rather than half credit. This August 2026 implementation clarifies the current rule but does not establish which code produced the 2025 leaderboard.[1]

Missing predictions also affect comparison. The paper permits slightly different question sets across models and studies random prediction removal by simulation. That simulation does not establish that actual API failures or refusals are unrelated to event difficulty. The human study uses 40 experts and 300 questions, but explicitly notes that humans and models did not receive identical question sets. Its human comparison is therefore contextual evidence, not a matched performance gap (Zeng et al., 2025).

Increasing the number of scored events can reduce sampling variability without removing selective coverage. FutureX's random-removal experiment addresses variability under the simulated deletion process; it does not bound bias when a model systematically omits difficult questions. A useful extension would report both each model's answered cohort and the common answered cohort, together with missingness by task type and forecast horizon. Neither conditional score alone represents performance on every intended question. This distinction also matters when comparing models with different refusal or tool-failure behavior.

F.6 AGENTS, MARKETS AND CONSISTENCY

MIRAI provides both a full-cohort experiment and a common-subset model comparison (Ye et al., 2024). Table 1 uses 705 queries, whereas Table 2 compares backbones on the same 100-query subset. For second-level relation F1, switching from single-function to code-block actions changes GPT-3.5 from 27.9 to 14.4 and GPT-4o from 32.6 to 32.8. Thus the common-subset comparison

[1] FutureX public evaluation code, commit 2626f7a (August 24, 2026).

avoids the cohort mismatch incurred by combining Tables 1 and 2. These descriptive contrasts do not establish a statistically reliable model-by-interface interaction; relation F1 is also distinct from probability loss. Section B.1 retains both experiments separately.

MIRAI's test construction also defines the scope of this comparison (Ye et al., 2024). Targets are directed country-pair relation sets from November 2023. Test events must have at least 100 daily mentions and five downloaded news articles; the resulting 705 queries contain multiple possible relations per query. The 100-query subset further balances dates, countries across continents, and first-level relation types. These filters select well-reported events, so the reported performance does not establish coverage of low-reporting or rare events. A useful extension would stratify a temporally controlled evaluation by reporting intensity while auditing label completeness: reducing the reporting threshold alone could confuse weak forecasting with missing ground truth. Restricting retrieved information to the forecast cutoff is necessary, but does not independently rule out pretraining exposure.

The released MIRAI evaluator clarifies the aggregation unit. At commit `badda88`, it computes relation precision, recall and F1 within each query, averages those values equally across queries, and then averages run-level metrics. Its nearby micro-averaging comment conflicts with the called function. A separate output selects a prediction for each query using its gold-label F1; that oracle selection must not be substituted for ordinary run averages. We inspected these paths and exercised the metric functions on a constructed fixture, without reproducing the historical table artifacts. This distinction illustrates why a survey must record the executed aggregation and output branch, rather than infer them from a metric name or code comment. The same implementation retains an unparseable answer as an empty relation prediction, which receives zero query F1 when gold relations are present. Missing files and malformed file-level JSON instead raise errors before that fallback. Our controlled fixtures verify this distinction; they do not estimate a model failure rate. Such cases differ from evaluators that exclude unsuccessful predictions and report a conditional score on the surviving cohort.

A follow-up on MIRAI reuses observations gathered during ReAct to compare one-step generation with and without the preceding thoughts (Zhang et al., 2025). On its balanced 100-query subset, Table 2 reports second-level F1 of 30.79 without thoughts and 27.33 with them for GPT-4o-mini at a one-day horizon. For Llama-3-70B, the corresponding values are 32.73 and 32.51 at one day, but 30.49 and 31.62 at seven days. These contrasts support model- and horizon-dependent effects, without reported uncertainty establishing significance. They control the retrieved observations more closely than the paper's separate RAG–ReAct comparison, which also restricts retrieval scope; thought inclusion still changes context content and length. We retain table values because the prose reverses the one-day and seven-day GPT-4o-mini observation-only comparison. Neither result establishes that reasoning generally helps or harms forecasting.

The archived v2 Prophet Arena Table 2 reports GPT-5 with high reasoning effort at Brier loss 0.184, compared with 0.187 for its market reference, each with a 0.006 95% interval half-width (Yang et al., 2025). The analysis covers 1,367 events resolved before 11 October 2025 and excludes forecasts within three hours of resolution. Marginal intervals and a 0.003 difference do not establish a paired Brier-loss advantage. Table 6 identifies this row as high reasoning effort; its unsuffixed and minimal-reasoning GPT-5 variants score 0.187 and 0.188 respectively. The abbreviated main-table label should not erase this configuration distinction.

Probability loss and economic performance also diverge within this source. Prophet Arena reports an average return ratio of 0.943 for its high-reasoning GPT-5 configuration, compared with 0.899 for the market baseline; their reported bootstrap 95% interval half-widths are 0.042 and 0.043. Both point estimates are below the break-even ratio of one. A lower Brier loss therefore does not establish profitable execution under the source's betting policy. The main metric definition averages policy-specific payoffs with a unit budget per market; this is not an annual portfolio return or a paired significance test of the return difference (Yang et al., 2025).

OpenForecaster's consistency evaluation compares Qwen3-8B with its RL-trained counterpart on long-term forecasting question tuples (Appendix B.4, Table 2). The aggregate arbitrage violation changes from 0.031 to 0.026, while negation violation changes from 0.043 to 0.063 (Chandak et al., 2025). Thus an average improvement can coexist with a component regression. Neither value is an

ordinary Brier loss. Keeping unfavorable components in the evidence map prevents a synthesis from converting a conditional result into a general success claim.

F.7 OpenForecast: improvements need not transfer across answer formats

OpenForecast Table 2 reports that Llama3-8b fine-tuning increases next-event multiple-choice accuracy from 55.7% to 96.1%, while short-term event-generation modified F1 falls from 56.8% to 22.9% and long-term generation F1 from 49.5% to 22.9% (Wang et al., 2025). These are different endpoints, not three interchangeable accuracy measurements. Section 6.1 uses sampled human evaluation for open-ended tasks and automatic answer matching for closed-ended tasks. The contrast illustrates that success at choosing a supplied continuation need not transfer to generating a supported event set. It does not isolate a single cause of the training tradeoff or establish statistical significance from the table alone.

The scoring modification is substantive. Appendix B.3 adds missing true predictions, denoted MTP, when external evidence or reasoning supports an event absent from the reference labels. Precision becomes $(TP+MTP)/(TP+MTP+FP)$ and recall becomes $(TP+MTP)/(TP+MTP+FN)$. These scores depend on the verification procedure as well as the fixed labels. Table 2 uses sampled human evaluation for open-ended tasks; the retrieval-augmented automatic evaluator is studied separately. Evidence retrieved to judge a completed forecast is not necessarily information supplied to the forecaster. The two access boundaries should be recorded separately (Wang et al., 2025).

The open-ended cohort also requires a separate denominator. Section 6.1 samples 100 questions per task, but Appendix B.2 removes samples when the two annotation teams' absolute F1 difference exceeds 0.1 before adjudication. The inspected passages do not establish the final retained count or Table 2's exact averaging weights. This is a different evaluation population from the automatically scored closed-ended tasks. The temporal split places training data before June 30, 2023, validation in July–August 2023, and testing from September 1, 2023 through March 31, 2024. A chronological split alone does not verify the absence of relevant information in every evaluated model's pretraining. A reproducible comparison should preserve the retained question identifiers, annotation exclusions and aggregation rule alongside the model forecasts (Wang et al., 2025).

F.8 PolyBench: selection and execution define the return denominator

PolyBench distinguishes 38,666 collected markets from 5,097 prediction-ready markets and 36,165 reported model predictions (Cheng et al., 2026). These counts are not interchangeable sample sizes. Moreover, seven predictions per prediction-ready market would total 35,679, not 36,165; the inspected counts do not establish a common evaluation cohort across all seven models. Evaluation excludes skipped predictions and discards decisions with stated confidence below 0.6. For executed buys, the confidence-scaled budget is filled against the recorded ask book, and CWR divides summed settlement profit by actual invested capital. Comparisons therefore depend on participation and execution, not just whether the predicted outcome occurs. The displayed profit formula does not separately deduct fees, so we do not relabel CWR as an all-cost net return.

The paper elects not to use Brier scores, citing the lack of unbiased continuous probability references. A Brier score on resolved binary events requires realized outcomes rather than an oracle probability reference. Proper-score evaluation and execution simulation could therefore be complementary on a specified resolved cohort; neither substitutes for the other. Its reported Sharpe definition uses the distribution of individual trade returns, which should also be distinguished from a calendar-time portfolio-return series. We retain these definitions without pooling them with other benchmarks' probability or portfolio metrics.

PolyBench Table 5 reports accuracy of 75.0% for Gemini-3-Flash and 62.1% for MiMo-V2-Flash, whereas their CWR values are 6.2% and 17.6%, respectively, at a base lot size of \$10 (Cheng et al., 2026). This descriptive ordering reversal does not establish better performance under matched event and execution conditions: participation, prices and invested-capital weights enter the return comparison. The same table reports unweighted Non-CWR values of 4.1% and 11.1%. Higher confidence-weighted returns alone do not identify probability calibration; a profitable allocation can depend on prices and liquidity even when reported probabilities are miscalibrated. A discriminating

follow-up would hold the attempted cohort and execution policy fixed, report a proper probability score on that cohort, and separately vary the allocation rule. These reported returns are not relabelled as fee-adjusted performance.

F.9 SCATTER: DIFFICULTY RELATIVE TO A REFERENCE MODEL

SCATTER defines hard subsets using questions that GPT-4o-mini fails across all sampled rounds (Chang et al., 2026). The retained hard sets contain 335 of 456 OpenForecast test questions and 61 of 90 OpenEP questions. These are adapted populations: the former uses short-term forecasting questions, while the latter re-extracts answer spans and excludes time and location queries. A gain on these subsets measures transfer to a reference model's failure set, rather than difficulty independent of model identity. The reference's zero success rate is imposed by selection and cannot independently demonstrate a competing model's unconditional advantage. A complementary comparison would retain the full cohort and report overlap between models' failure sets.

SCATTER's aggregation preserves the question as the scoring unit: Appendix A.2 averages sample-level scores equally across test questions (Chang et al., 2026). Section 5.1 reports Pass@1 means and standard deviations over 16 sampling rounds. For this main-results setup, each response contains ten hypotheses; Pass@1 therefore measures success of one generated set, not one answer, and its reported variability is not variation across independently trained models. Pass@K uses an LLM correctness judge, whereas soft matching uses embeddings. The sample, generation and verification levels must remain separate: additional hypotheses expand a question's candidate set without increasing the number of test questions, and a semantic match does not establish a calibrated event probability.

Our inspection of SCATTER's released evaluation driver also identifies a reproducibility boundary distinct from model accuracy. A question that exhausts the outer evaluation retries is saved with an empty result but omitted from the fresh-run aggregation list. When the driver resumes from that saved file, the empty result enters the aggregation list and the question is skipped. A controlled probe of the original driver, replacing model evaluation with a deterministic success-or-failure fixture, verifies that the aggregation denominator changes across these execution paths. This probe uses no model calls and does not establish that the published results encountered such failures. The practical requirement is to preserve intended, completed and scored cohorts, and to apply the same failure policy when a run is resumed. The inspected code revision and probe are retained with our evidence artifacts.

F.10 OPENFORECASTER: CONFIDENCE REWARD AND PROBABILITY LOSS

OpenForecaster defines a higher-is-better free-form answer score $S' = 2qz - q^2$, where q is confidence and z indicates whether the generated answer matches the reference. Its training reward may additionally include correctness, giving $R = z + S'$ (Chandak et al., 2025). Algebraically, $S' = z - (q - z)^2$, so an aggregate reward difference combines a change in answer accuracy with a change in squared confidence error when systems generate different answers. It is therefore insufficient to negate this reward and compare it directly with binary event Brier loss. The correctness variable is tied to the generated answer and semantic matching procedure, rather than a single fixed event proposition shared by every system. For a fixed answer with correctness probability p, expected S' is maximized at $q = p$; this truthful-confidence property does not make its cross-system aggregate identical to an ordinary fixed-target Brier comparison. Both accuracy and confidence error should be retained when interpreting this training objective.

The reward ablation and final evaluation also use different cohorts. Design choices use a 207-question validation set drawn from July 2025 Guardian articles, whereas the final free-form test retains 302 questions from an initial set of 1,000 across five other news sources (Chandak et al., 2025). The latter undergoes search-assisted answerability checks, earliest-resolution checks and manual filtering. Training ends in April 2025, but the reported May–August test-construction range overlaps the validation month. Distinct source populations and delayed inspection of test results should therefore be distinguished from a test period wholly after validation. These facts do not establish leakage; they specify which temporal and population transfer claims the experiment can support.

The retained test population is also conditioned on a search-enabled model being able to answer the retrospectively constructed questions. This filter may improve label quality while changing which events remain eligible. A prospective follow-up could commit questions before resolution and report performance separately by evidence availability, retaining unresolved or unanswerable cases in coverage accounting. Such a comparison would test whether gains transfer beyond the selected, retrospectively answerable cohort; the existing filtered results alone do not settle that question.

G WITHIN-STUDY CONTRAST TABLE

Table 4: Within-study contrasts and their synthesis status. Status attaches to the contrast in each row, not to the paper as a whole. Arrows follow the reported intervention; values retain their original units and are not a cross-benchmark ranking.

Study / level	Retained contrast	Supported reading and boundary
ForecastQA (Jin et al., 2021)		
Qualified	BERT-large closed-book → integrated: accuracy 57.6% → 60.1%; class-summed Brier .543 → .589	Same benchmark cohort and metrics, but the integrated configuration does not isolate retrieval; supports a metric trade-off, not retrieval attribution.
PRISM: calibration (Jin et al., 2026)		
Matched reference	Direct → isotonic calibration: Brier .094 → .070	Same country-month task and scoring; this is the post-hoc calibration reference. The source does not mark this contrast significant.
PRISM: full system (Jin et al., 2026)		
Qualified bundle	Direct → full PRISM: Brier .094 → .062	The full system also changes measurement and inference. The source marks only this Direct-LLM contrast significant (paired country bootstrap, $p < .05$); it does not isolate calibration's causal share.
SCATTER (Chang et al., 2026)		
Matched within study	Qwen2.5-3B, OpenForecast: GRPO → SCATTER, Pass@16 25.66 → 30.26; Pass@1 11.84 → 8.55	Same benchmark protocol; repeated-sample coverage and one-response set success are separate endpoints. Preserve the sampling budget and semantic judge.
Competence pooling (Tiwari et al., 2026)		
Matched comparator	Gate vs external baseline Brier .0771 → .0732; logistic stacker .0740	Five-fold question evaluation; the gate does not significantly outperform the logistic comparator ($p = .53$).
FutureX (Zeng et al., 2025)		
Descriptive	Tier-weighted ordinal ranking	A heterogeneous composite supports a ranking under its weights, not a common proper-loss comparison; evaluator identity remains unresolved.

Table 4 collects comparisons whose interpretation changes when their evaluation contract is retained. The binary scale conversion in Equation 3 does not align mixed two-/four-class aggregates, different events or different information sets. Likewise, Pass@16 is a sampling-budget endpoint, not a probability score. A useful synthesis compares matched contrasts; changing units alone cannot repair a different prediction task.

In PRISM, the direct-to-calibrated Brier difference is 0.024, while the direct-to-full-system difference is 0.032 on the same country-month conflict task. The calibrated condition is the explicit post-hoc reference; the full-system comparison also changes measurement and statistical inference. Table 1 marks the Direct-LLM versus full PRISM comparison significant under paired country-level bootstrap resampling ($p < 0.05$), but does not mark the calibration-only contrast. Neither aggregate gap identifies calibration's causal share. The table's marginal bootstrap deviations are not paired in-

tervals and should not be relabelled as standard errors or used as independent error bars for a paired difference (Jin et al., 2026).

H CLAIM-ALIGNED COMPARISON CONTROLS

A comparison should specify what is allowed to differ. For an end-to-end system claim, retrieval and aggregation may differ, but systems still need a common question cohort, commitment schedule and scoring rule. For a claim about reasoning, the available evidence must additionally be held fixed; otherwise better retrieval is an alternative explanation. For a claim about calibration, the same forecast targets and a calibration split separate from evaluation are needed. The appropriate control therefore follows from the claim, rather than from a universal requirement that every component be identical.

We use three levels of synthesis. *Matched contrasts* retain the same targets, scoring and evaluation population while identifying the changed system component. *Qualified contrasts* retain an explicit mismatch, such as different human time budgets or a system comparison that changes several components together, and support only the narrower observation under those conditions. *Descriptive juxtaposition* places results from different tasks side by side without interpreting their numerical difference as a performance gain. These labels attach to individual reported contrasts, not to papers as a whole: one study can contain a matched calibration reference alongside a bundled system comparison. Unknown protocol fields remain unknown: neither equality nor incompatibility can be inferred from missing metadata. Score conversions, including Equation 3, change units but do not promote a qualified or descriptive comparison into a matched one.

Difficulty-conditioned comparisons require an additional check: how was the difficult subset selected? A cohort defined by one reference model's failures can show complementary capabilities, but its reference-model score is constrained by that selection. Full-cohort performance and cross-model failure overlap are needed before interpreting such a subset as a general difficulty scale (see the SCATTER example in the appendix).

I EXTRACTION RULES AND ILLUSTRATIVE RECORDS

Each numerical record distinguishes reported, derived and reproduced evidence. Reported values come from original sources, derived values apply an explicitly stated transformation, and reproduced values require a retained implementation and inputs. The present survey has reported and derived records; it does not label the original systems as reproduced.

A consolidated artifact preserves 43 source-extraction records from seventeen archived source files, including records with multiple contrasts. A versioned extension adds ten system-level AI World Cup records independently recomputed from a frozen public prediction, fixture and leaderboard export. This is an artifact inventory, not the number of papers screened or an estimate of field coverage. The recomputation recovers group-stage outcome counts, exact-score counts and additive match points for all ten submissions; the leaderboard's group-standing and knockout components remain source-reported. No model was rerun. All source hashes were checked, but that check does not establish complete protocol metadata. Missing structured fields remain explicit, and the original records are retained.

For auditability, the normalized companion ledger records each row's local source hash, version or identifier, locator, target and output description, temporal protocol, population, metric, aggregation rule, uncertainty statement and comparison boundary. Fields absent from the archived material are encoded as "unknown" rather than inferred. The original 43 rows remain source-reported; the ten added rows are marked as recomputed from released forecasts and fixtures, explicitly without reproducing any original forecaster. This makes the result map machine-checkable while preserving the restriction that scores from different contracts must not be pooled.

The local evidence ledger records ForecastQA Table 3 on PDF page 7; Autocast Table 2 on page 6; Halawi et al. Table 3 on page 6; the ICLR 2025 ForecastBench Table 2 on page 9; Prophet Arena v2 Table 2 on page 9; and OpenForecaster Table 2 on page 21. Locators refer to the archived PDF files rather than a changing web rendering. Source hashes and retrieval records accompany the corpus.

Table 5: Examples of result contracts that must remain distinct.

Record	Critical field	Invalid extrapolation
ForecastQA humans (Jin et al., 2021)	Sampled questions, skipping, information access	Exact paired full-test human–model gap
Halawi combined (Halawi et al., 2024)	Crowd input included	Independent model human parity
ForecastBench Table 2 (Karger et al., 2025)	Equal source-group weighting; unresolved rules	Ordinary all-resolved raw-event Brier mean
Prophet v2 (Yang et al., 2025)	Fixed 2025 resolved cohort	Current live-model ranking
OpenForecaster consistency (Chandak et al., 2025)	Logical violation metric	General prediction accuracy improvement

J Reproducing the secondary analysis

The original plotting script reads the retained component values and evaluates Equation 4 on 101 equally spaced weights between zero and one. A CSV preserves the resulting curves. The crossing is approximate because the source components are rounded; no uncertainty interval for the crossing is estimated from unavailable raw forecasts. The method is sufficient to illustrate aggregate sensitivity, but it cannot support a significance claim or identify a universally appropriate weight.

The expanded analysis uses `survey_audit.py` and the pinned official CSV. The artifact includes the full commit identifier, downloaded file hash, 130 retained rows and the pairwise reversal summary. Rounded component reconstruction may differ slightly from the official rounded overall column, so all comparisons use one consistently reconstructed baseline. The source CSV contains aggregate scores; this is not an event-level rescore or paired bootstrap replication.

The search log is a snapshot of the targeted follow-up, not a fabricated count of all papers in the field. A full review should add explicit duplicate resolution, inclusion decisions, backward and forward citation tracing and independent extraction checks. Adjacent numerical forecasting and temporal reasoning studies should be retained in the screening log even when excluded from the event-focused main synthesis.

J.1 Why literal source matching is insufficient

Our extraction checks distinguish literal provenance from experimental attribution. In the retained MIRAI PDF, the introduction on page 2 attributes 29.6 second-level relation F1 to GPT-4o. However, Section 3.1 and Table 1 on page 6 identify GPT-3.5-turbo-0125 and 705 queries for that result; Table 2 reports GPT-4o scores of 32.6 and 32.8 on a separate 100-query subset. A model-assisted extraction reproduced the introduction's attribution with a literally matching excerpt. We retain this source inconsistency and use the explicit experimental setting for the performance comparison. A matching quotation therefore establishes where a claim occurs, but does not establish consistency with the experiment that produced the number.

This example motivates linking each extracted result to its table, experimental paragraph, model, and cohort, rather than accepting a paper-level summary as sufficient evidence. Machine-assisted records remain provisional until those fields are checked. We do not treat this single case as an estimate of extraction error across the literature.

K Lifecycle comparisons require a moving-reference interpretation

TimeSeek evaluates forecasts at five lifecycle checkpoints for 150 Kalshi markets, with and without search (Lee et al., 2026a). The inspected workshop version uses date-filtered web retrieval and withholds market prices from the forecaster. Its Table 2 reports Claude Opus 4.5 Brier skill score (BSS) of $+0.167$ at the first checkpoint and -0.829 at the last; pooled BSS is -0.068. These are

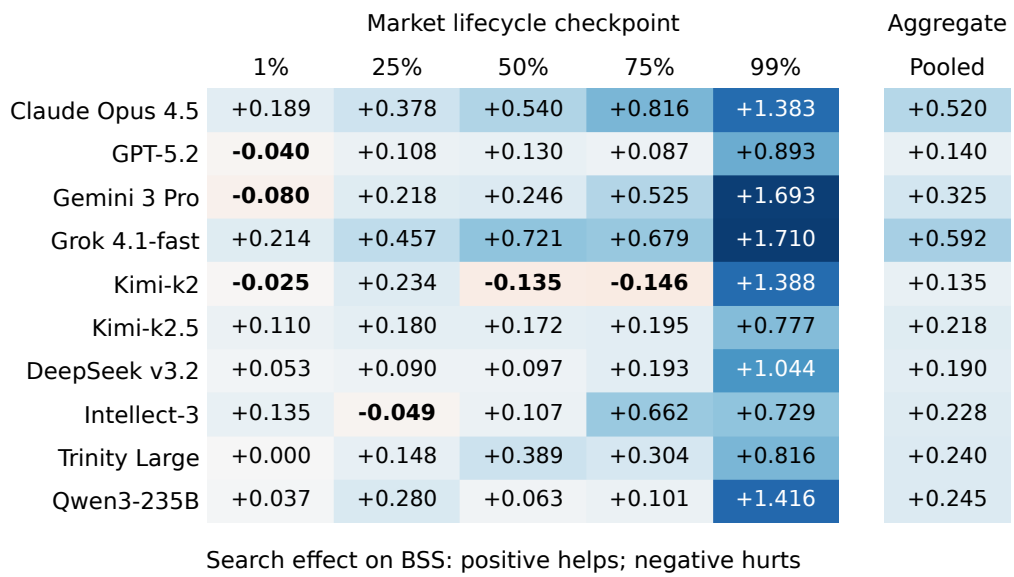


Figure 5: Positive pooled search effects coexist with six negative model–checkpoint effects in TimeSeek. Values reproduce Table 3 of the inspected version (Lee et al., 2026a): search-enabled minus no-search Brier skill score. Bold negative entries identify worse performance with search; color uses a symmetric scale around zero. Pooled contrasts use pooled losses, not the arithmetic mean of checkpoint contrasts. Results are descriptive, with no subgroup confidence intervals reported; checkpoints repeat the same markets.

descriptive point estimates without reported subgroup confidence intervals. Repeated checkpoints are observations of the same markets, not independent event samples.

The definition $\mathrm{BSS} = 1 - \mathrm{BS}_{\mathrm{model}}/\mathrm{BS}_{\mathrm{market}}$ makes the reference explicit. A lower BSS later in a market's life establishes worse relative performance at that checkpoint; it does not by itself establish that the model's absolute Brier loss increased. Market loss may also change. Accordingly, a lifecycle comparison should retain both absolute losses and the relative score, with uncertainty accounting for repeated markets. Table 3 additionally contains six negative search effects among fifty model–checkpoint contrasts despite positive pooled effects for every model. This supports reporting heterogeneous tool effects, but not a validated rule for choosing when to search. Date-filtered retrieval and post-cutoff resolutions also do not independently certify the historical availability of every retrieved document version or the contents of model training data.

Search effort is another selection variable. For Claude, TimeSeek reports non-monotonic Brier skill scores across the 750 market–checkpoint predictions grouped by the number of searches chosen: -0.811 for 1–3 searches ($n = 48$), $+0.109$ for 4–7 ($n = 152$), $+0.007$ for 8–12 ($n = 238$), and -0.115 for 13 or more ($n = 312$) (Lee et al., 2026a). Because query count is endogenous and may reflect question difficulty, these groups do not identify an optimal budget; assigning one budget to every question would change the cases and decision process being measured. A policy comparison should assign budgets or stopping rules before outcomes on a common cohort, retain failed and unanswered forecasts, and score forecast paths under identical timing and scoring rules. The source's descriptive grouping motivates, but does not test, that intervention.

K.1 A FROZEN RETRIEVAL ARCHIVE DOES NOT FREEZE MODEL KNOWLEDGE

Hindcast restricts Reddit retrieval by document creation time and compares a zero-shot model with an archive-enabled agent (Ye et al., 2026b). Its 216-market cohort is selected after resolution: a Qwen3-32B probe queries the pre-resolution archive for evidence about both sides, and the authors retain markets rated strongly covered for the realized side before balancing YES/NO outcomes.

Table 1 reports Qwen3-32B Brier loss of 0.234 without retrieval and 0.179 with retrieval, with run-to-run standard deviations of 0.007 and 0.008 across five runs. This contrast estimates performance on an outcome-conditioned, strongly covered cohort; it does not estimate retrieval's benefit over all naturally occurring markets.

Two distinctions constrain interpretation. First, Section 5.1 instructs the bare model to use knowledge available at the historical cutoff, but this instruction does not remove later knowledge from its weights. The paper's Section 2 itself distinguishes archive restriction from temporally restricted model training. We therefore treat the implementation as a retrieval-time control rather than evidence that both leakage channels are eliminated. Second, Table 1's coverage bins depend on each run's own retrieval tally, so bin membership changes across runs. A strong-versus-weak-bin contrast is not a randomized intervention on evidence quality. It combines retrieval behavior and case selection; even its reported dispersion includes membership changes. These details support retaining selection and grouping rules alongside the metric.

K.2 A BUNDLED ORCHESTRATION COMPARISON

Agentic Time Machine combines a filter over live web results with a planner–solver–aggregator forecaster (Chai et al., 2026). Its Table 2 compares the latter with GPT-5.4-medium self-consistency under the same Time Machine tools. On eighty Polymarket questions, the proposed framework scores 64.99, compared with 62.26 for five-sample self-consistency; these are scaled set-level F1 scores, not probability losses. Figure 6 reports fewer average tool-call rounds for the framework (47.1 versus 81.0). However, non-ensemble results average three runs, whereas self-consistency is reported from one ensemble run. Tool-call counts also do not establish matched token or monetary budgets.

Because planning, evidence paths and aggregation all change together, this comparison evaluates a bundled orchestration design rather than isolating the effect of decomposition. It complements the retrieval-ablation evidence above but cannot attribute the score or tool-call difference to any one component. The Time Machine itself approximates a historical environment by filtering current search results. Its comparison with unrestricted search tests the consequences of that filtering; a score reduction alone cannot certify complete leakage removal or preservation of all legitimate historical evidence. These remain separate validation questions from whether the downstream multi-agent system outperforms self-consistency on the retained questions.

K.3 LIVE TRADING AND PAPER TRADING ANSWER DIFFERENT QUESTIONS

Prediction Arena evaluates autonomous market participation rather than forecasts on a shared question set (Zhang et al., 2026). Its primary cohort contains six models trading live on Kalshi from January 12 to March 9, 2026; four newer models entered simulated paper trading from March 6 to 9. In paper mode, orders execute without the real-counterparty rejection that can prevent live orders from filling. Gemini-3.1-Pro-Preview made no Kalshi trades and was excluded from that platform's quantitative comparison. The authors describe the three-day paper cohort as preliminary directional evidence.

These cohorts differ in duration, trade opportunities and execution conditions, while the paper also reports cross-platform results over different windows. Pooling them would confound model, period, opportunity set and execution regime. Settlement win rates apply only to resolved positions and do not replace probability scores on a matched question set. We therefore treat the study as evidence about deployed decision systems under their respective cohort and platform conditions, not as a controlled ranking of predictive accuracy.

K.4 REASONING ALIGNMENT AND OUTCOME ACCURACY ARE DISTINCT ENDPOINTS

WorldReasoner reports outcome, evidence and reasoning quality separately (Chi et al., 2026). Across 525 contamination-filtered model–question pairs, weighted accuracy is 68.8% for Search-Enabled and 64.4% for Search-Enabled Graph; the latter condition reports 9.8% key-event recall (KER). The 525 records are not unique questions: filtering is applied by model, so cross-model rows need not share the same population. KER requires recall of the top five hindsight events under textual similarity and a seven-day date gate. It measures alignment to a post-resolution reference

whose causal links are approximate, not exhaustive causal validity. The accuracy contrast and KER value describe different endpoints and do not establish a causal accuracy effect from graph construction.

Appendix F defines a reduced multiple-choice Brier score from confidence c in the selected option: the loss is $(1 - c)^2$ when that option is correct and c^2 otherwise (Chi et al., 2026). Because the system records no probability vector over all answer options, this contract should remain separate from full-vector multiclass Brier scoring. Together, these results show why an evidence map must retain the target, outcome space and scoring rule alongside each metric name.

K.5 Ranking robustness and probability incentives are different claims

SocietyBench anonymizes entities and shifts dates in retrospective social-event timelines, then retains an item only when none of three adversarial LLM auditors identifies its original event and time period with high confidence (Wang et al., 2026). This checks direct recovery; it does not establish that structural event memory is absent from the evaluated models. The calibration axis maps a weighted mean absolute probability error to a 0–100 score, with weights for time to resolution and forecast-window length, rather than using a proper scoring rule. Table 5 re-scores the same answers under eight alternatives: the GPT–Doubao gap remains positive under every rule, while Brier and log-loss reduce Doubao's score by 6.9 and 18.9 points, respectively. This supports pairwise ranking robustness on this five-event set, not a general ranking or a claim that weighted absolute error elicits calibrated probabilities. Answers still unparseable after three retries are excluded, and per-question predictions are not released, limiting independent reconstruction of the scored coverage.

The incentive distinction follows directly from the loss. For a binary outcome with true probability q and reported probability p,

$$\mathbb{E}|p - Y| = q(1 - p) + (1 - q)p = q + p(1 - 2q).$$

For $p \in [0, 1]$, expected absolute error is minimized at $p = 1$ when $q > 1/2$, at $p = 0$ when $q < 1/2$, and by every report when $q = 1/2$. By contrast, expected squared error, $q(1 - p)^2 + (1 - q)p^2$, is minimized at $p = q$. This analytic distinction does not rerank SocietyBench; it separates robustness of the reported model ordering from whether the headline loss elicits truthful probabilities. A review should record both ranking sensitivity and the incentive properties of each scoring rule.

K.6 Updating forecasts requires more than a strong initial prediction

FutureSim places forecasting agents in a chronological replay of news and resolved questions, allowing them to decide which forecasts to revise and when to advance the simulated date (Goel et al., 2026). The evaluated object therefore includes an update policy, memory and information search, rather than only a probability assignment made once. Its original paper restricts access to documents available at the current simulation date and lets agents propose their own candidate outcomes. Such replay controls document availability; claims about model knowledge still depend on the evaluated model's training cutoff and the question construction process.

The study separates starting quality from adaptation by giving different agents the same initially poor forecast set. The tested agents improve to different degrees but remain below its abstention reference in that comparison. This supports examining sensitivity to initial forecasts, rather than attributing final performance entirely to adaptation. A separate comparison gives each question individual attention shortly before resolution, with more accurate forecasts than sequential updating. Because this intervention changes timing, attention allocation and exposure to earlier predictions together, its gap does not isolate anchoring alone. A mechanism study would need to vary those components separately.

FutureSim also illustrates why equally named scores cannot be pooled without checking their definitions. Its Brier skill score subtracts a multicategory squared loss from one, with unreported outcomes assigned zero probability and total submitted probability mass bounded above by one. Its zero reference is abstention. TimeSeek instead normalizes a model's Brier loss by a market reference (Lee et al., 2026a). These reference choices answer different questions. We therefore retain each score's outcome space and reference alongside results, and treat matched starting forecasts and update opportunities as additional requirements for comparing adaptation policies.

Table 6: FutureSim's reported search and updating contrasts for GPT 5.5 at maximum reasoning effort, from Figure 6 of the original preprint (Goel et al., 2026). Values are top-1 accuracy percentages. Rows describe different interventions and must not be combined into a single component attribution.

Evaluation setting	Intervention	Accuracy
Sequential simulation	Daily news updates	24.8
Sequential simulation	No context updates	17.9
One day before resolution	Agentic search	31.2
One day before resolution	Single semantic query	14.2

L Open problems and a testable research agenda

The following agenda distinguishes proposed experiments from established findings. Its organizing question is which missing comparison would change our interpretation of forecasting progress. Each direction therefore specifies an intervention, an evaluation target and a possible failure, rather than merely recommending larger models or more data.

The mechanism taxonomy becomes useful when competing explanations imply different results. Table 7 translates the synthesis into proposed comparisons. The unit of assignment should be the event or a prespecified event family, with shared evidence snapshots and paired forecasts where possible. Prompt selection, calibration and update-policy tuning belong to a development cohort; the complete procedure is frozen before the evaluation cohort is observed. These controls make a null or negative result informative: it can reject a particular explanation of improvement without claiming that a whole class of forecasting methods is ineffective.

Table 7: Proposed tests derived from the evidence synthesis. These are research designs, not completed experiments or claims of newly established performance. Each comparison requires a common event cohort and information cutoff.

Question	Comparison that distinguishes explanations	Result that would limit the claim
Information access or evidence use?	Cross retrieval procedures with inference procedures; reuse archived document bundles and separate conditions with external probabilities.	A gain disappears on shared evidence: the study supports acquisition benefit, not improved use of identical information.
Reasoning or probability adjustment?	Compare raw, development-calibrated and full-system forecasts; fit calibration without evaluation labels.	The full system does not improve over the calibrated reference: extra reasoning has no demonstrated incremental benefit in this cohort.
Useful adaptation or favorable scoring weights?	Compare fixed and adaptive update schedules under common external scoring checkpoints, including a frozen-forecast control.	A gain exists only under averaging issued messages: it does not demonstrate better forecasts at common times.
Probability skill or scenario coverage?	Hold generation budget fixed; retain atomic predictions, resolved correctness and probability scores separately.	More valid scenarios accompany worse probability loss or more unsupported additions: coverage alone is insufficient.
Transfer or recurring event templates?	Freeze the full method, then evaluate new event families with stratification by horizon and evidence availability.	Benefits occur only in familiar, well-covered families: broad transfer remains unsupported.

L.1 Cohort commitment and transfer

A prospective study should publish its eligibility rule and record all eligible questions before prediction. Resolution delay, ambiguity and adjudication disagreement are outcomes of the evaluation process, not reasons to silently replace a question. Report resolved-score estimates together with cohort coverage and the distribution of pending or excluded cases. A sensitivity analysis can bound the effect of pending outcomes; it should not assign them observed labels.

Missing forecasts require a separate rule. FutureX allows slight cross-model cohort differences because strict alignment would reduce its sample, and estimates added variability by randomly subsampling events at different missing rates (Zeng et al., 2025). That simulation assesses random omissions; it does not test whether refusals or API failures cluster on difficult events or particular domains. If missingness is selective, the random-subsample analysis may not capture its effect on model-specific scores. A prospective comparison should preserve the attempted cohort, predeclare how missing forecasts are scored, and report both coverage and conditional performance. Otherwise, models may be compared on the questions they happened to answer.

A changing leaderboard can reflect a changing evaluator even when every forecast is frozen. ForecastBench scores unresolved market questions against updated community estimates and replaces those targets with outcomes after resolution (Karger et al., 2025). A longitudinal comparison should therefore archive three separate timestamps: forecast commitment, evidence cutoff and scoring snapshot. First rescore an unchanged forecast archive on an unchanged cohort across resolution snapshots; this measures sensitivity to target updates, not forecaster learning. Then compare forecasting procedures using the same cohort, snapshot, resolution rules and aggregation weights. Report resolved-only and provisional components separately, together with pending-case coverage. A gain that disappears under a shared snapshot cannot establish improved forecasting. This design does not make provisional community estimates ground truth or remove selection bias from the subset that has resolved.

Live evaluation addresses commitment time but does not ensure representative coverage (Karger et al., 2025; Zeng et al., 2025). A useful next experiment freezes a forecasting procedure before collecting new events, then stratifies evaluation by topic, horizon and evidence availability. Report both within-population and shifted-population performance, with event families kept together when estimating uncertainty. The unresolved question is whether a method learns reusable probability judgments or exploits recurring templates and rich coverage. Failure on low-coverage events should remain visible rather than disappearing through resolution or retrieval filters.

L.2 Updating forecasts under a fixed scoring contract

An endpoint score can hide the forecast trajectory that produced it. TimeSeek measures performance at five market-lifecycle checkpoints, with and without search, but does not compare policies that update one committed forecast as evidence arrives (Lee et al., 2026a). It therefore motivates time-aware update studies without identifying an optimal policy. Compare fixed schedules, evidence-triggered updates and budget-matched adaptive schedules on the same prospectively committed questions. Record every issued probability, evidence arrival and tool expenditure. Score the latest committed forecast on a common, externally fixed checkpoint schedule, alongside terminal score, calibration and coverage. For abstention, preserve a fixed question denominator and report risk–coverage curves; otherwise skipping difficult questions can masquerade as improved forecasting. An update that merely repeats previously observed information is an important failure case.

An average over issued messages lets a model change its evaluation weights by choosing when to speak. Consider a constructed binary example with outcome one and forecasts 0.5 and 0.9 at two equally weighted checkpoints. Both systems receive Brier loss 0.13. If one system emits eight additional copies of the second forecast, its average loss per issued message falls to 0.034 without changing either checkpoint forecast. These numbers illustrate an accounting effect, not a measured model improvement. Instead, use

$$L_i = \sum_{k=1}^{K} w_{ik}\big(p_i(t_{ik}^-) - y_i\big)^2, \qquad w_{ik} \geq 0, \quad \sum_k w_{ik} = 1, \tag{6}$$

where the schedule and weights are fixed before outcomes and independently of issued updates, and $p_i(t_{ik}^-)$ is the latest forecast committed before the checkpoint. Specify an initial forecast or a common missing-forecast rule. Fixed nonnegative weights preserve the proper-score incentive at the scored checkpoints; they do not identify an optimal update policy or eliminate dependence between repeated scores for the same event. Report update costs separately and estimate uncertainty at the event or event-family level.

L.3 Consistency constraints and their limits

For an event's first occurrence time T, forecasts issued at the same time from the same information must satisfy

$$P(T \le \tau_1 \mid \mathcal{I}_t) \le P(T \le \tau_2 \mid \mathcal{I}_t), \qquad \tau_1 < \tau_2. \tag{7}$$

The events must be nested under identical resolution criteria. This inequality does not require probabilities issued on successive days to increase: new evidence changes the information set. Nor does it apply automatically to an event being true *on* a date rather than occurring *by* it. A useful experiment compares independently elicited horizon probabilities with a constrained joint forecast, retaining the same evidence and testing both consistency and resolved-outcome loss. Following the intervention limits of Paleka et al. (2025), a lower violation rate alone would not establish better prediction. Questions still awaiting resolution must remain explicitly unresolved rather than being relabeled as negative outcomes.

For the nested horizons above, Euclidean projection onto $C = \{q \in [0,1]^K : q_1 \le \cdots \le q_K\}$ gives a useful classical baseline; isotonic projection is established statistical machinery (Yang & Barber, 2018). Let $q = \Pi_C(p)$ and let the fully resolved outcome vector be $y_k = \mathbf{1}\{T \le \tau_k\}$. Since $y \in C$, the projection inequality $(p-q)^\top (y-q) \le 0$ implies

$$\|q - y\|_2^2 \le \|p - y\|_2^2 - \|p - q\|_2^2. \tag{8}$$

Thus equal-weight mean Brier loss cannot increase on that complete nested family. This is a direct application of convex projection, not a new forecasting theorem. A constructed example projects $(0.8, 0.2)$ to $(0.5, 0.5)$: for outcome $(0, 1)$, mean Brier loss falls from 0.64 to 0.25. Our retained deterministic grid check verifies the inequality numerically; it supplies no model-performance evidence. The guarantee concerns the family aggregate, not every horizon separately, and need not hold for differently weighted evaluation, nonnested questions, changed information or an outcome-selected subset of resolved horizons. It does not extend to arbitrary consistency training. Future experiments should include this baseline before attributing gains to a learned consistency procedure.

L.4 Attributing gains and evaluating practical value

The retrieval and orchestration comparisons reviewed here motivate a controlled retrieval-by-inference design: fix retrieved documents while varying the forecaster, then fix the forecaster while varying retrieval (Halawi et al., 2024; Chai et al., 2026). Keep crowd probabilities out of the model-only condition and report a separate, explicitly labeled crowd-assisted condition. Match event cohorts and report evidence-acquisition and inference costs separately. These comparisons could distinguish gains from better information access, better use of the same evidence, and an external crowd signal; interactions may still prevent a universal ranking of components.

OpenForecast addresses incomplete reference outcomes using evidence-assisted verification (Wang et al., 2025). An unresolved challenge is to score specificity, coverage and false additions together. A candidate protocol commits atomic propositions and their probabilities before resolution, then uses blinded adjudication against archived evidence, preserving disagreements and unresolved cases. Stress tests should include vague paraphrases, redundant predictions and numerous low-confidence guesses. Report adjudicator sensitivity separately from forecaster quality. The research problem is not solved by a more persuasive language-model judge: the scoring incentive itself must discourage strategically uninformative forecasts.

Prediction Arena illustrates that execution changes between live and paper trading: the preliminary paper cohort faced no real-counterparty rejection, unlike the live cohort (Zhang et al., 2026). A controlled follow-up should first elicit probability forecasts on a common unresolved-event cohort under the same information cutoff, timestamp them before resolution, and score them separately. It should then pass those forecasts through a shared frozen decision policy, market snapshot, cost schedule and execution model; alternative policies belong in a separate experiment that holds forecasts fixed. This separates predictive quality, allocation and execution. PnL alone cannot identify which component changed, and better proper-score performance need not improve every decision objective.

Public tests invite adaptation: archive versions, searches and exposure, then evaluate frozen systems on new cohorts. Report retrieval and model-weight exposure separately; replay cannot constrain model weights (Ye et al., 2026b).