# FIELD: A PROSPECTIVE CROSS-SPORT BENCHMARK FOR LANGUAGE-MODEL FORECASTING AND INFORMATION USE

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Forecasting unresolved events requires models to use evolving evidence, express uncertainty, and update beliefs, yet how observable information use and belief updating relate to forecast quality remains unclear. Sports provide a natural testbed for studying these relationships, with scheduled events, verifiable outcomes, and meaningful intermediate states. We introduce **FIELD**, a dynamic, prospective cross-sport benchmark for early and near-event forecasting of how sporting events unfold and end. FIELD freezes forecasts and observable workflow records before each event, enabling assessment of both predictive quality and information use. We evaluate seven frontier language-model systems on 232 football, basketball, and baseball events. Across sports, forecast performance is strongest for final Outcome, while richer score distributions and trajectories remain harder. We further find that better trajectory forecasts are associated with better outcome forecasts; observable workflow does not reliably track forecast quality; and later forecasts mostly refine probabilities rather than change the leading prediction. A separate *STEP* study prospectively tests whether these findings can be translated into a structured forecasting procedure, with effects varying substantially across events.

## 1 INTRODUCTION

Forecasting real-world future events is emerging as an important application of LLMs and research agents (Su et al., 2026; Lopez-Lira & Tang, 2026; Metaculus, 2026). Beyond its value for decisions under uncertainty, forecasting evaluation provides a demanding test of core model capabilities: integrating incomplete and evolving evidence, expressing uncertainty probabilistically, and revising judgments as new information arrives. Accordingly, recent work has increasingly studied language-model forecasting through forecasting agents, dynamic benchmarks, and continuously refreshed unresolved questions (Halawi et al., 2024; Karger et al., 2025; Zeng et al., 2026; Yang et al., 2026).

Rigorous forecasting evaluation increasingly follows a prospective and probabilistic paradigm: forecasts are collected at fixed prediction times to prevent access to future information (Paleka et al., 2026), and, once events resolve, the submitted probabilities are evaluated with proper scoring rules (Gneiting & Raftery, 2007). Yet predictive performance alone shows how well a system forecasts, not what characterizes strong forecasting or how information use and belief updating relate to forecast quality. Similar scores may arise from very different workflows, while well-supported evidence can still be translated into poor probabilistic judgments. Recent work has therefore begun to examine forecasting pipelines, information acquisition, evidence grounding, and belief revision alongside forecast outcomes (Yang et al., 2026; Yuan et al., 2026; Chi et al., 2026).

Yet differences in tasks and resolution criteria complicate such joint evaluation across domains. We therefore focus on sports as a structured testbed. Sports events are scheduled in advance, recur at scale, resolve against public records, and retain meaningful uncertainty until resolution. They also expose richer future structure than a winner label alone, supporting linked forecasts of final outcome, score distribution, and intermediate game states. Sports forecasting further involves time-sensitive information gathering and integration; when retrieval and workflow traces are exposed, these processes can be studied alongside the forecasts they produce. Broad forecasting benchmarks such as FutureX include sports among many domains (Zeng et al., 2026), while dedicated prospec-



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.

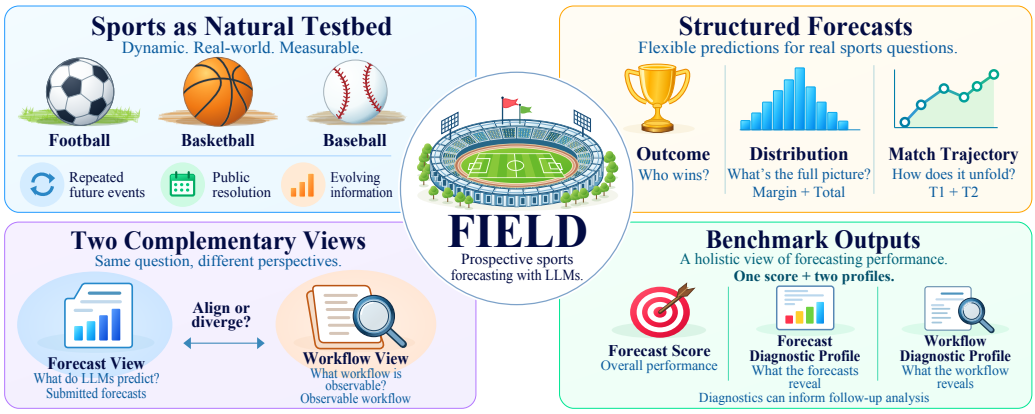


Figure 1: **Conceptual overview of FIELD.** Sports provide repeated, prospectively observable events with public outcomes and evolving information. FIELD maps each event into three forecast facets—Outcome, Distribution, and Match Trajectory—and evaluates the submitted forecast and observable workflow through complementary Forecast and Workflow views. Only forecast quality contributes to the Forecast Score; complementary diagnostic profiles support deeper analysis and future system design.

tive sports benchmarks provide fine-grained evaluation within football (Schröder et al., 2026; Wang et al., 2026). Yet broad coverage and single-sport depth leave a question: which patterns of forecasting competence recur across sports? A shared cross-sport protocol allows us to examine these patterns across different event structures while preserving sport-specific semantics.

We introduce **FIELD**, a dynamic, prospective cross-sport benchmark for measuring and studying language-model forecasting (Figure 1). FIELD maps heterogeneous sports onto three linked forecast facets—Outcome (final result), Distribution (final-score uncertainty), and Match Trajectory (intermediate game states)—evaluated at $H_1 = T - 24\text{h}$ and $H_2 = T - 2\text{h}$ under provider-Native and shared-Controlled settings (Sec. 3.2). FIELD pairs a *Forecast View* of predictive quality with a *Workflow View* of observable information use (Sec. 3.3). These views allow us to examine whether stronger forecasts are accompanied by stronger measured workflows, and where they diverge.

We instantiate FIELD on 232 sporting events across football, basketball, and baseball, evaluating seven frontier language-model systems. Across sports, forecast performance is strongest for final Outcome, while richer score distributions and match trajectories remain more challenging. FIELD's diagnostics reveal three patterns: **(1)** more accurate forecast records tend to include more accurate trajectory predictions and, in some settings, greater attention to intermediate game states; **(2)** observable workflow and forecast quality show some local associations but no stable overall correspondence; and **(3)** as events approach, forecasts usually refine existing probabilities rather than reverse the leading prediction.

These findings motivate testable hypotheses for forecasting-system design: modeling match trajectories, mapping evidence to probabilities, and revisiting earlier premises when updating forecasts. We probe them in a separate prospective 68-event *STEP* study: 20 of 59 fully paired events show positive mean gains, with at least 4 of 6 models improving. Mixed effects in the remaining fully paired events leave when and how these interventions help an open empirical question. Our contributions are threefold:

- **A dynamic, prospective cross-sport benchmark.** To our knowledge, FIELD is the first prospective LLM benchmark to unify Outcome, Distribution, and Match Trajectory forecasting across sports, with seven systems evaluated on 232 events.
- **A two-view paradigm for forecasting evaluation.** We introduce complementary Forecast and Workflow views to separately assess predictive quality and observable information use, enabling joint diagnosis beyond aggregate scores.
- **Testable directions for forecasting-system design.** FIELD's diagnostic findings motivate design hypotheses, which an exploratory prospective STEP study probes on new events.

# 2 Related Work

## 2.1 LLM Forecasting and Its Evaluation

Forecasting has become an important setting for evaluating language-model judgment under uncertainty. Modern systems do more than emit a final label: they retrieve time-bounded information, combine heterogeneous evidence, express uncertainty probabilistically, and may revise beliefs as new evidence arrives (Halawi et al., 2024; Ye et al., 2024; Zeng et al., 2026; Yang et al., 2026; Yuan et al., 2026). Correspondingly, forecasting benchmarks span several temporal protocols, from historical cutoffs and news-based questions (Jin et al., 2021; Zou et al., 2022), to repeatedly refreshed unresolved events (Karger et al., 2025; Zeng et al., 2026), continuous evaluation against incoming information (Dai et al., 2025), and controlled pastcasting within frozen historical information environments (Wildman et al., 2025). Evaluation has also expanded beyond final correctness to proper probabilistic scoring, calibration, consistency, updating, and tighter control of temporal validity (Gneiting & Raftery, 2007; Paleka et al., 2025; Yuan et al., 2026; Paleka et al., 2026). These advances make forecasting evaluation increasingly realistic, but aggregate scores still reveal little about how forecasts are formed.

## 2.2 Sports as a Testbed for Real-World Forecasting

Sports provide a particularly structured setting for real-world forecasting: events are scheduled in advance, outcomes resolve objectively, information changes rapidly before the event, and uncertainty remains meaningful until resolution (Spann & Skiera, 2009; Wunderlich & Memmert, 2020). Recent studies around the 2026 FIFA World Cup extend beyond winner prediction to compare information access, prompting and forecast horizons, detailed match and tournament targets, autonomous search, and market-based baselines (Schröder et al., 2026; Wang et al., 2026; Ding et al., 2026; Shianifar & Faiud, 2026). SPORTU and SportR further show that heterogeneous sports can share evaluation abstractions while retaining sport-specific semantics, though for multimodal understanding and reasoning rather than prospective probabilistic forecasting (Xia et al., 2025; 2026). Together, these efforts leave open a shared cross-sport setting for prospective probabilistic forecasting, making it difficult to determine which forecasting patterns are sport-specific and which generalize across sports. FIELD addresses this gap with a common task and scoring framework for Outcome, Distribution, and Match Trajectory while preserving sport-specific semantics.

## 2.3 Forecast Outcomes and Observable Forecasting Processes

A parallel line of work has begun to examine observable forecasting processes. Forecasting agents have been decomposed into research and judgment stages, observable research traces have been analyzed, and belief revision has been studied as new information arrives (Yang et al., 2026; Liptay et al., 2026; Yuan et al., 2026). WorldReasoner likewise evaluates forecast outcomes together with temporally valid evidence and explicit reasoning structure, and identifies the conversion of grounded evidence into forecast probabilities as an important remaining challenge (Chi et al., 2026). Methods for factuality and citation grounding further support the evaluation of externally verifiable evidence (Min et al., 2023; Gao et al., 2023), while work on reasoning faithfulness cautions against treating generated rationales as transparent access to latent reasoning (Lanham et al., 2023). Although recent studies explore forecasting processes, how observable information use relates to forecast quality remains insufficiently understood. FIELD examines this relationship through separate evaluation and joint analysis of the same prospective records.

# 3 FIELD Framework and Methodology

FIELD turns unresolved sporting events into time-stamped prospective forecasting records through four stages. First, we register future events and map heterogeneous sports onto a shared forecasting task (§3.1). Second, we collect forecasts prospectively under fixed temporal and information conditions (§3.2). Third, each archived run is evaluated through separate Forecast and Workflow views (§3.3). Finally, forecast quality is normalized against frozen historical references and aggregated into benchmark outputs (§3.4). Figure 2 summarizes this end-to-end pipeline; Appendices A–C detail execution protocols, truth resolution, and statistical procedures.

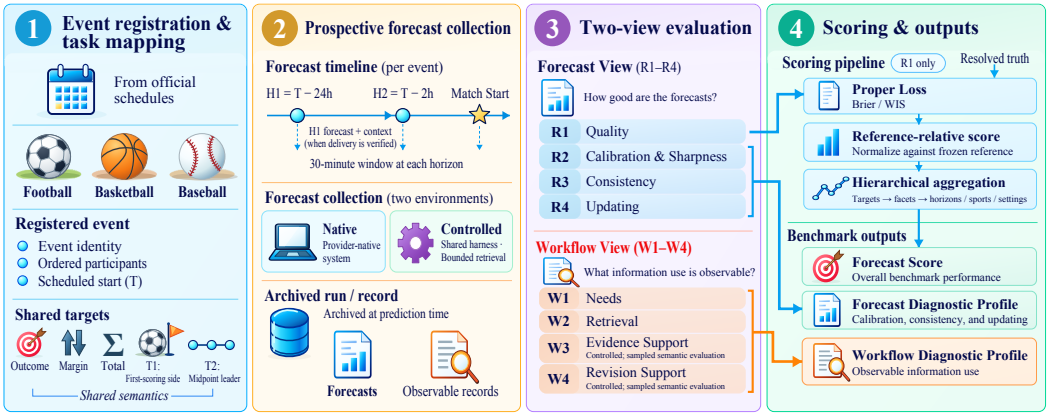


Figure 2: **FIELD operational pipeline. 1**: register future events and map shared cross-sport targets (§3.1); **2**: collect and archive forecasts and observable records under two settings and two pre-event horizons (§3.2); **3**: evaluate Forecast and Workflow views separately (§3.3); **4**: normalize forecast quality against frozen references and aggregate benchmark scores (§3.4).

### 3.1 Prospective Registration and Cross-Sport Forecasting Task

FIELD begins by registering unresolved sporting events before forecasting and mapping football, basketball, and baseball onto a shared five-target task: final Outcome, final-score Margin (score difference) and Total (combined score), and two intermediate-state targets, $T1$ and $T2$. These targets form three connected forecast facets. **Outcome** is the final categorical result; **Distribution** represents final-score uncertainty through Margin ($S_a - S_b$) and Total ($S_a + S_b$), each submitted as five predictive quantiles ($q_{.10}, q_{.25}, q_{.50}, q_{.75}, q_{.90}$); and **Match Trajectory** captures how the event unfolds through $T1$, the first scoring side, and $T2$, the leader/tie state at a sport-specific midpoint—halftime in football, the end of the second quarter in basketball, and the state after five completed innings in baseball. Football additionally permits a draw for Outcome and `no_score` for $T1$. Each event's sport, participant ordering, scheduled start time, and side-a/side-b orientation are fixed before forecasts are collected, providing a common task abstraction while preserving sport-specific semantics. These three sports constitute the present instantiation of FIELD rather than the framework's fixed scope. Additional sports can be incorporated when their events can be registered prospectively, the forecast targets can be resolved objectively, and the shared targets admit clear sport-specific realizations. Complete category spaces, target definitions, and target-level truth-resolution rules for the present instantiation are given in Appendix B.

### 3.2 Prospective Forecast Collection

With the shared cross-sport task defined, FIELD next collects forecasts prospectively at two pre-event horizons, $H_1 = T - 24\text{h}$ and $H_2 = T - 2\text{h}$, each with a 30-minute execution window. We evaluate two deployment settings that reflect two common paradigms for using LLMs on complex tasks. **Native** treats the provider's deployed end-to-end system as the forecasting system itself, including the information-access capabilities exposed by that deployment. **Controlled** instead uses each model as the backbone of a shared external agent harness with a common task format, bounded retrieval interface, and common procedure. Together, these settings evaluate forecasting in provider-native deployments and under a shared external harness. Detailed retrieval budgets and execution procedures are given in Appendix A.

The two horizons define a sequential forecasting trajectory for each event. $H_1$ provides the initial forecast. When delivery of the earlier forecast and carried context is verified, $H_2$ can begin from the accepted $H_1$ forecast and the context actually carried forward before seeking new information and submitting an updated five-target forecast. FIELD archives submitted forecasts, observable information records, and the realized $H_1 \rightarrow H_2$ lineage from the as-run record. The prospective record for each horizon is fixed when its execution window closes. This preserves the information boundary while capturing both an earlier and a near-event view of the same unresolved event.

Table 1: **FIELD two-view evaluation framework.** Only R1 contributes to the aggregate Forecast Score; R2–R4 and W1–W4 are diagnostic measurements. Complete forecast-metric and workflow-diagnostic definitions are provided in Appendices C and D.

| View | Dimension | What it asks | Main measurement |
|---|---|---|---|
| **Forecast** | R1 Forecast Quality | Did the model assign good probabilities to what eventually occurred? | Brier/WIS proper losses; reference-relative score (Appendix C). |
| | R2 Calibration & Sharpness | Does stated uncertainty match empirical outcomes without being unnecessarily diffuse? | Top-label ECE; predictive-interval coverage and width (§C.8). |
| | R3 Forecast Consistency | Is the submitted forecast numerically valid and coherent across related targets? | Probability/quantile validity; necessary coherence violations (§C.9). |
| | R4 Forecast Updating | How does the forecast change as the event approaches, and does realized loss improve? | Forecast movement; $H_1 \to H_2$ realized-loss gain (§C.10). |
| **Workflow** | W1 Information Needs | What information does the model say it needs, and does it pursue it? | Declared information topics; topic-pursuit rate (§D.2). |
| | W2 Information Retrieval | What information does the model search for, access, and receive? | Requested queries/run; opened sources; delivered evidence (§D.2). |
| | W3 Evidence Support | Are checkable factual claims supported by prediction-time evidence? | Conditional evidence-support rate (§D.3). |
| | W4 Revision Support | Are stated reasons for changing or retaining a forecast supported by prediction-time evidence? | Conditional revision-support rate (§D.3). |

### 3.3 Two-View Forecast and Workflow Evaluation

Once forecasts and their observable records have been archived prospectively, FIELD evaluates them through two complementary perspectives. The **Forecast View** evaluates the submitted predictions themselves, while the **Workflow View** characterizes the observable information-seeking and evidence-use process surrounding them. Table 1 summarizes the eight evaluation dimensions and their main measurements.

**Forecast View.** R1 measures predictive quality with proper scoring rules, so forecasts are rewarded for assigning probability mass accurately rather than only for selecting the eventual outcome. Outcome, $T1$, and $T2$ use normalized Brier loss, which measures the squared discrepancy between the submitted categorical probabilities and the realized outcome. Margin and Total use finite-quantile Weighted Interval Score (WIS), which balances the width of the submitted predictive intervals against penalties when the realized score falls outside them (Brier, 1950; Gneiting & Raftery, 2007; Bracher et al., 2021). R2 evaluates calibration and sharpness—whether expressed uncertainty agrees with empirical outcomes and how concentrated that uncertainty is. For categorical forecasts, top-label expected calibration error (ECE) measures the discrepancy between confidence in the most probable category and its empirical frequency of being correct; a well-calibrated set of forecasts assigned roughly 70% confidence, for example, should realize that category at approximately the same rate. For Margin and Total, predictive-interval coverage measures how often realized scores fall within the submitted intervals, while interval width characterizes their concentration.

R3 evaluates numerical validity and coherence across related forecast components. Categorical probabilities must form valid distributions, predictive quantiles must remain ordered, and linked targets must satisfy necessary compatibility constraints. In particular, the submitted probability that side $a$ wins should remain compatible with the range implied by the corresponding Margin quantiles. The principal R3 summary reports the rate at which evaluable forecasts violate this necessary Outcome–Margin coherence condition. R4 evaluates temporal updating on matched $H_1$–$H_2$ forecasts. For target $d$, $\Delta L_d = L_{H_1,d} - L_{H_2,d}$, so positive values indicate lower realized loss at $H_2$; the

model-level R4 gain summarizes these improvements after reference normalization. Exact formulas and additional diagnostics are provided in Appendix C, with R2–R4 detailed in §C.8–§C.10.

**Workflow View.** Workflow evaluation uses only artifacts observable at the system boundary; unexposed provider activity and private reasoning are not reconstructed. W1 measures whether information topics declared as needed are subsequently pursued, while W2 characterizes the resulting retrieval trace, including queries, opened sources, and delivered evidence. W3 evaluates whether externally checkable factual claims are supported by evidence available at prediction time, and W4 evaluates whether stated reasons for revising or retaining a forecast are similarly supported. W3/W4 semantic judgments are computed from source-bound prediction-time records and validated against two independent human references. Their reported support rates characterize support among the extracted and evaluated units. Full extraction, sampling, and validation procedures are given in §D.2–§D.4.

Lower R2 ECE and R3 violation rates indicate more favorable measurements, as do higher R4, W3, and W4 values; W1 and W2 are descriptive rather than intrinsically ordered. Only **R1 Forecast Quality** contributes to the aggregate Forecast Score. R2–R4 and W1–W4 remain diagnostic measurements, allowing predictive performance and observable forecasting process to be studied together without conflating them.

### 3.4 Reference-Normalized Scoring and Benchmark Outputs

The final step is to place heterogeneous forecast targets on a common scale. FIELD first scores each forecast directly against realized truth, using Brier loss for Outcome, T1, and T2 and WIS for Margin and Total. Because these raw losses are not directly comparable across heterogeneous targets, we normalize them against frozen sport- and target-specific historical references using the standard skill-score transformation from forecast verification (Murphy, 1988; Wilks, 2011). We call the resulting quantity a reference-relative score and use it for benchmark aggregation; underlying raw losses are reported in Table 4. For model $m$ in evaluation cell $c$ with eligible event support $I_{m,c}$,

$$\mathrm{Score}^{\mathrm{ref}}_{m,c} = 100\left(1 - \frac{\sum_{i\in I_{m,c}} L^{\mathrm{model}}_{i,c}}{\sum_{i\in I_{m,c}} L^{\mathrm{ref}}_{i,c}}\right).$$

The reference provides a fixed normalization baseline, with the model and reference evaluated on exactly the same events. A score of 0 matches the reference, positive values indicate lower aggregate loss, and negative values indicate higher aggregate loss. The frozen reference uses 120 previously resolved events in each sport–target cell, restricted by a pre-evaluation historical cutoff.

FIELD then aggregates these normalized scores hierarchically. Outcome forms one facet; Distribution averages Margin and Total; and Match Trajectory averages $T1$ and $T2$. The three facets receive equal weight, followed by equal weighting across the two forecast horizons and the three sports. Native and Controlled scores are computed separately and then equally combined into the final Forecast Score. Full reference construction and aggregation rules are provided in Appendix C.

## 4 FIELD Benchmark Results and Scientific Analysis

We first establish the main benchmark comparison and assess its robustness to analysis choices. We then compare cross-sport forecasting profiles and use FIELD's diagnostic records to examine trajectory prediction, observable workflow, and temporal updating.

### 4.1 Prospective Evaluation and Benchmark Validation

Our initial prospective evaluation covers 232 events and seven systems. Because some scheduled forecasts are missing or invalid and are never backfilled, the main Forecast Score comparison uses shared valid support across all seven systems. This shared support spans a union of 226 events, with the exact event set determined separately across forecast targets and evaluation conditions (Appendix C.4). The resulting ranking is highly concordant with the ranking obtained from each system's own available forecasts (Spearman $\rho = 0.964$), and also with a reference-free aggregation of within-cell raw-loss ranks ($\rho = 0.964$; Appendix C.11). It remains unchanged under three alterna-

Table 2: **FIELD forecasting performance and diagnostic profiles.** Overall and facet scores use the shared valid support in Section 4.1; diagnostics use their own eligible records or temporal pairs. Panel B reports Controlled W1–W4, with W3/W4 conditional on sampled, extracted units assigned substantive support labels. Only R1 contributes to Overall. Facet scores are reference-relative. Arrows and bold mark favorable observed values where intrinsically ordered; W1/W2 are descriptive. Bold does not imply significant superiority. Definitions and coverage are in Appendices C–D.4.

**Panel A: Forecast View**

| Model | R1 Forecast Quality | | | | Forecast Diagnostics | | |
|---|---|---|---|---|---|---|---|
| | **Overall** Forecast Score ↑ | **Outcome** Score ↑ | **Distrib.** Score ↑ | **Traj.** Score ↑ | **R2** ECE (%)↓ | **R3** Viol. (%)↓ | **R4** Gain (×100)↑ |
| Gemini-3.8-flash | **16.11** | **28.63** | **11.53** | 8.18 | 12.08 | 4.66 | 0.68 |
| Kimi-K3 | 15.44 | 25.81 | 11.36 | **9.16** | 10.31 | 2.44 | 0.15 |
| DeepSeek-v4-pro | 14.54 | 24.44 | 10.24 | 8.95 | **8.43** | 2.11 | 0.00 |
| GPT-5.6-sol | 14.52 | 24.33 | 10.29 | 8.94 | 12.01 | 9.24 | **2.20** |
| Grok-4.6 | 14.44 | 24.20 | 10.34 | 8.78 | 10.69 | 5.28 | 1.03 |
| Claude-Opus-5 | 12.90 | 24.49 | 7.64 | 6.55 | 11.06 | 14.07 | -1.03 |
| Qwen3.8-max | 11.80 | 22.31 | 7.23 | 5.85 | 10.08 | **1.64** | 1.01 |

**Panel B: Workflow View (Controlled)**

| Model | **W1 Information Needs** Pursuit (%) | **W2 Information Retrieval** Queries / run | **W3 Evidence Support** Cond. support (%)↑ | **W4 Revision Support** Cond. support (%)↑ |
|---|---|---|---|---|
| Gemini-3.8-flash | 59.3 | 9.27 | 85.31 | 73.6 |
| Kimi-K3 | 70.7 | 11.94 | 83.79 | 81.8 |
| DeepSeek-v4-pro | 78.7 | 10.51 | 91.43 | 91.1 |
| GPT-5.6-sol | 79.9 | 10.90 | **92.98** | **91.2** |
| Grok-4.6 | 80.5 | 11.99 | 89.99 | 83.7 |
| Claude-Opus-5 | 77.4 | 11.69 | 85.07 | 80.0 |
| Qwen3.8-max | 82.3 | 11.75 | 90.19 | 90.4 |

tive facet weightings ($\rho = 1.000$) and broadly similar when any one sport is excluded ($\rho = 0.857$–$0.964$). Because W3/W4 rely on automated semantic evaluation, we also validate the automated judgments against two independent human references. Weighted agreement is 89.5–90.6%, while agreement between the two human annotators is $\kappa = .883$ (Appendix D.4).

### 4.2 Cross-Sport Benchmark Results and Capability Trends

Table 2 summarizes each system's aggregate and facet-level forecast performance together with its Forecast and Workflow diagnostic profile.

**Forecasting strength is multidimensional.** *Gemini* has the highest Overall, Outcome, and Distribution scores, while *Kimi* performs best on Match Trajectory. The diagnostic profiles reveal a different pattern. *DeepSeek* has the lowest R2 top-label ECE, indicating the closest alignment between categorical confidence and empirical accuracy; *Qwen* has the lowest R3 necessary-coherence violation rate, indicating greater consistency between its outcome and score-margin forecasts; and *GPT* shows the largest positive R4 updating gain, corresponding to the greatest average realized-loss improvement from $H_1$ to $H_2$. *GPT* also has the highest conditional W3 and W4 support rates, meaning that, among the evaluated extracted units, its factual claims and stated revision reasons are most often fully supported by prediction-time evidence. These contrasts show that stronger aggregate forecasting performance does not imply uniformly stronger calibration, coherence, updating, or observable information use.

**The clearest common pattern is endpoint-heavy forecasting.** Across all 21 model–sport combinations, the Outcome facet score exceeds both Distribution and Match Trajectory; within Distribution, Margin also consistently exceeds Total (Table 15). Because these facet scores use target-specific reference normalization to place heterogeneous forecast tasks on a common scale, the pattern indicates that performance is consistently strongest for final Outcome, while richer score distributions and intermediate event states remain more challenging (relative to each target's frozen reference baseline; Section 3.4). Calibration shows another shared tendency: across all 14 model–setting categorical profiles, top-label accuracy exceeds mean confidence, indicating average underconfidence (Table 16).

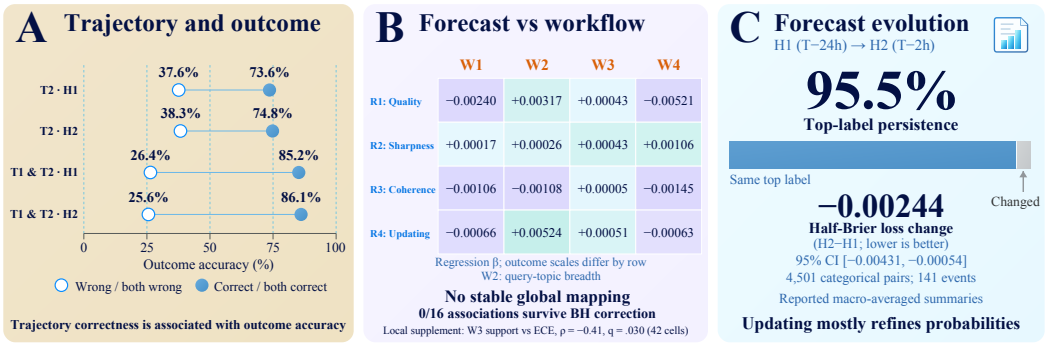


Figure 3: **Three diagnostic views of forecasting competence. A**, more accurate trajectory forecasts are associated with higher Outcome accuracy. **B**, the pre-specified Forecast–Workflow atlas shows no stable global mapping after multiplicity correction, while a separate calibration analysis identifies a local W3–ECE association that is not stable across sensitivity analyses. **C**, matched forecasts usually retain the same leading category from $H_1$ to $H_2$ while improving probabilistic loss, indicating predominantly probability refinement rather than answer reversal.

### 4.3 Diagnosing Forecasting Competence Beyond Aggregate Scores

Aggregate benchmark scores tell us which forecasts perform better, but leave important aspects of forecasting competence unresolved. We therefore use FIELD's archived prospective records to examine three complementary questions: what characterizes more accurate forecast records at a fixed prediction time; how observable forecasting processes align with or diverge from forecast quality; and how forecasts evolve as the same event approaches.

**More accurate forecast records show better representations of how the event may unfold.** Outcome accuracy is about 74% when the midpoint-state target T2 is correct, versus about 38% when T2 is wrong. The association persists after controlling for event, model, and setting: a one-standard-deviation improvement in trajectory quality is associated with a 0.17–0.20 increase in Outcome quality, measured as negative reference-normalized loss. Observable search behavior provides complementary evidence. At $H_2$, systems with the strongest Outcome forecasts within an event more often query midpoint states and first-scoring events than other systems forecasting the same event, with no detected increase in total query count. Thus, stronger records combine more accurate trajectory predictions with, in some settings, greater attention to the event's unfolding dynamics (Tables 8–9).

**Forecast quality and observable workflow are complementary, not interchangeable.** None of the 16 pre-specified R1–R4 $\times$ W1–W4 associations survives Benjamini–Hochberg correction at 5%, leaving no stable global mapping within the tested atlas. A separate calibration analysis across 42 Controlled model–sport–horizon cells nevertheless identifies a local association: stronger W3 evidence support coincides with better categorical calibration (lower ECE; Spearman $\rho = -0.41$, $q = .030$). Archived cases further show both mismatches: stronger measured workflow signals can accompany weaker forecasts, and stronger forecasts can arise with weaker measured workflow signals (Table 12). Observable workflow can therefore inform but not replace direct forecast evaluation, supporting FIELD's two-view design of separate evaluation and joint analysis. Full associations, support sets, and cases are reported in Appendix C.7.

**Later forecasts mainly refine an existing belief as search behavior narrows.** From $H_1$ to $H_2$, the macro-averaged top-label persistence is 95.5% across matched categorical forecasts, yet half-Brier loss improves by 0.00244 (95% CI [0.00054, 0.00431]). Updating therefore occurs mainly through probability refinement rather than answer reversal. Search behavior changes at the same time: Controlled $H_2$ forecasts issue queries spanning roughly 1.5 fewer topic families, with broad history, standings, and form queries receding while market-, venue/weather-, and official-source-oriented queries become relatively more prominent. Later forecasting therefore combines a narrower search regime with refinement of an already formed belief.

# 5 Outlook and Discussion

## 5.1 Implications for Sports-Forecasting Systems

FIELD's benchmark and diagnostic analyses offer three practical insights for the forecasting community. **Model the path, not only the endpoint.** Performance is strongest on final Outcome, while more accurate records also tend to contain better trajectory predictions, motivating richer representations of score distributions and intermediate event states. **Separate information acquisition from probabilistic judgment.** Observable workflow provides useful local signals but is not a stable surrogate for forecast quality, so evidence acquisition and probability assignment should be treated as distinct capabilities. **Treat repeated forecasts as belief updates.** Later forecasts usually preserve the leading category while refining probabilities, motivating explicit belief-state carryover as a design hypothesis for repeated forecasting.

## 5.2 Exploratory Prospective Evidence from STEP

We prospectively probe these insights with **STEP**, a structured forecasting procedure that operationalizes them through State/Trajectory modeling, Evidence-to-Probability projection, and Transition-aware updating. At $H_1$, STEP represents how the event may unfold and connects evidence to forecast probabilities; at $H_2$, it revisits the earlier forecast using newly available information. We compare STEP with the unchanged BASE procedure on 68 new events; full details are reported in Appendix E.

Among 59 events with complete exploratory gains, 28 have positive mean effects; 20 improve for at least four of the six analyzable models and span all three sports, while 20/28 remain positive at both horizons. These recurring responses show that positive effects are not confined to isolated model–event cases. However, the pre-specified average STEP–BASE effect is $-0.57$ Forecast Score points (95% CI $[-1.40, 0.28]$), and 31 events have negative mean effects. STEP therefore shows that FIELD-derived insights can produce recurring gains on a subset of events, but does not yet establish an average benefit or a stable rule for predicting when they will help. Continued prospective use of FIELD can accumulate the evidence needed to identify such response regimes and test more targeted strategies. We plan to release evaluation code and redistributable STEP prompts to support this.

## 5.3 Limitations and Future Directions

FIELD currently covers three sports, so observed patterns may not transfer directly to other sports or forecasting domains. As an as-run prospective benchmark, deployments can differ in provider availability, exposed tools, and delivery. Shared support improves comparability, but missing forecasts may still limit representativeness. Historical references normalize scores rather than serve as competitive forecasting baselines; future versions can add time-aligned market or statistical baselines. Workflow analyses are limited to observable, successfully extracted artifacts; for W4, the owner audit further indicates incomplete extraction coverage (Appendix D.4). The relationships studied in Section 4.3 remain observational. STEP tests one bundled operationalization and does not establish an average benefit or stable conditions under which it helps. *We therefore view FIELD as a continuing prospective benchmark: broader coverage can sharpen model comparisons, test generalization, and identify when forecasting strategies are effective.*

# 6 Conclusion

We introduced **FIELD**, a prospective cross-sport benchmark for evaluating language-model forecasts and observable forecasting processes. Across football, basketball, and baseball, FIELD reveals multidimensional forecast quality and complementary structure across forecast outcomes, observable workflow, and temporal updating. We distill these findings into practical insights for future forecasting systems and probe them prospectively through STEP; the intervention produces recurring positive responses on a subset of events, without establishing an average benefit. These mixed results highlight both what FIELD can already reveal and the open questions that require continued prospective evidence. FIELD therefore provides a shared framework for comparison, diagnosis, and cumulative study across future events, systems, and interventions.

### AI USE STATEMENT

We used generative AI tools to assist with literature and source discovery, research ideation and analysis workflows, software development, and drafting and polishing portions of the manuscript. All experimental designs, evaluation protocols, data processing procedures, statistical analyses, numerical results, and scientific claims were reviewed and verified by the authors. AI-generated outputs were not treated as ground truth. The authors take full responsibility for the content of the paper.

### ETHICS STATEMENT

FIELD is designed as a research benchmark for evaluating language-model forecasting, not as betting or financial advice. The study uses scheduled sporting events, publicly observable pre-event information, and public match outcomes. Forecasting outputs are analyzed for model evaluation and should not be interpreted as recommendations to wager or make financial decisions. Where retrieved third-party material is subject to copyright, access, or redistribution restrictions, the reproducibility record preserves provenance and integrity information without assuming that the underlying page can be republished.

The human-validation stage asked two annotators to evaluate model-generated claim or revision units against archived prediction-time evidence. Human labels were used only to validate the automated semantic evaluator and did not retroactively alter forecast submissions or select favorable model results. Annotator identities and restricted admin-side materials are kept out of the public evidence release; paper-facing analyses use aggregate statistics and de-identified validation artifacts. The validation record also preserves its documented limits, including the post-inspection dual-reference amendment and the absence of a separate attestation about annotators' use of outside tools or exposure to admin-side answers.

### REPRODUCIBILITY STATEMENT

FIELD is governed by prospective registries, time-stamped forecast records, target-level truth rules, a fixed historical reference, explicit analysis populations, and versioned analysis outputs. The authors retain the broader prospective research archive, including provider-dependent execution records and restricted third-party material where redistribution is not permitted. The accompanying portable evidence package is a processed-input, bounded offline replay package: it contains the scoring and analysis implementations, processed truth/reference inputs, result tables, source mappings, and validation artifacts needed to audit the reported analyses, but it does not claim to reproduce every original provider response or third-party web capture.

After publication, we intend to release the author-generated benchmark specifications, registries, prompts where redistributable, scoring and analysis code, processed result tables, and de-identified validation artifacts supporting the paper. Third-party or provider-originated material will be redistributed only where permitted; otherwise the release will preserve provenance pointers, integrity information, and derived auditable artifacts. Exact regeneration of live provider outputs may be affected by provider availability or model updates, so deterministic offline recomputation is distinguished from provider-dependent prospective execution.

## REFERENCES

Johannes Bracher, Evan L. Ray, Tilmann Gneiting, and Nicholas G. Reich. Evaluating epidemic forecasts in an interval format. *PLoS Computational Biology*, 2021.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 1950.

Yizhou Chi, Eric Chamoun, Zifeng Ding, and Andreas Vlachos. WorldReasoner: Evaluating whether language model agents forecast events with valid reasoning. *arXiv*, 2026.

Hui Dai, Ryan Teehan, and Mengye Ren. Are LLMs prescient? a continuous evaluation using daily news as the oracle. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Jiacheng Ding, Cong Guo, and Jason Xu. FIFA World Cup 2026 as a Contamination-Free Benchmark for LLM Forecasting Agents: Four Models, a Bookmaker, and 104 Matches. *arXiv*, 2026.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 2007.

Danny Halawi, Fred Zhang, Yueh-Han Chen, and Jacob Steinhardt. Approaching human-level forecasting with language models. In *Advances in Neural Information Processing Systems*, 2024.

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. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.

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. In *International Conference on Learning Representations*, 2025.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, et al. Measuring faithfulness in chain-of-thought reasoning. *arXiv*, 2023.

Tom Liptay, Dan Schwarz, Rafael Poyiadzi, Jack Wildman, and Nikos I. Bosse. Evaluating strategic reasoning in forecasting agents. *arXiv*, 2026.

Alejandro Lopez-Lira and Yuehua Tang. Can ChatGPT forecast stock price movements? return predictability and large language models. *Journal of Financial Economics*, 2026.

Metaculus. Forecasting tools: A framework for building AI forecasting bots. `https://github.com/Metaculus/forecasting-tools`, 2026.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Allan H. Murphy. Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, 1988.

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.

Daniel Paleka, Shashwat Goel, Jonas Geiping, and Florian Tramèr. Pitfalls in evaluating language model forecasters. In *International Conference on Learning Representations*, 2026.

Jonas Schröder, Jonas Schweisthal, Oliver Müller, Markus Weinmann, and Stefan Feuerriegel. LLM-SoccerArena: Benchmarking LLMs on real-world predictions in sports. *arXiv*, 2026.

Jonaid Shianifar and Iias Faiud. AI World Cup 2026: Benchmarking Large Language Models for End-to-End Football Tournament Prediction. *arXiv*, 2026.

Martin Spann and Bernd Skiera. Sports forecasting: A comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting*, 2009.

Shiqian Su, Sen Xing, Xuan Dong, Muyan Zhong, Bin Wang, Xizhou Zhu, Yuntao Chen, Wenhai Wang, Yue Deng, Pengxiang Zhu, Ziyuan Liu, Tiantong Li, Jiaheng Yu, Zhe Chen, Lidong Bing, and Jifeng Dai. MiroFlow: Towards high-performance and robust open-source agent framework for general deep research tasks. *arXiv*, 2026.

Zhaokai Wang, Tianlin Gui, Jiayuan Rao, Shangzhe Di, Yihong Tang, and Dingli Liang. WorldCupArena: Fine-grained evaluation of language models and deep-research agents on football forecasting. *arXiv*, 2026.

Jack Wildman, Nikos I. Bosse, Daniel Hnyk, Peter Mühlbacher, Finn Hambly, Jon Evans, Dan Schwarz, and Lawrence Phillips. Bench to the future: A pastcasting benchmark for forecasting agents. *arXiv*, 2025.

Daniel S Wilks. *Statistical methods in the atmospheric sciences*. Academic press, 2011.

Fabian Wunderlich and Daniel Memmert. Are betting returns a useful measure of accuracy in (sports) forecasting? *International Journal of Forecasting*, 2020.

Haotian Xia, Zhengbang Yang, Junbo Zou, Rhys Tracy, Yuqing Wang, Chi Lu, Christopher Lai, Yanjun He, Xun Shao, Zhuoqing Xie, Yuan-fang Wang, Weining Shen, and Hanjie Chen. SPORTU: A comprehensive sports understanding benchmark for multimodal large language models. In *International Conference on Learning Representations*, 2025.

Haotian Xia, Haonan Ge, Junbo Zou, Hyun Woo Choi, Xuebin Zhang, Danny Suradja, Botao Rui, Ethan Tran, Wendy Jin, Zhen Ye, Xiyang Lin, Christopher Lai, Shengjie Zhang, Junwen Miao, Shichao Chen, Rhys Tracy, Vicente Ordonez, Weining Shen, and Hanjie Chen. SportR: A benchmark for multimodal large language model reasoning in sports. In *International Conference on Learning Representations*, 2026.

Qingchuan Yang, Simon Mahns, Sida Li, Anri Gu, Jibang Wu, and Haifeng Xu. LLM-as-a-Prophet: Understanding predictive intelligence with prophet arena. In *International Conference on Learning Representations*, 2026.

Chenchen Ye, Ziniu Hu, Yihe Deng, Zijie Huang, Mingyu Derek Ma, Yanqiao Zhu, and Wei Wang. MIRAI: Evaluating LLM agents for event forecasting. *arXiv*, 2024.

Zhangdie Yuan, Zifeng Ding, and Andreas Vlachos. Do language models update their forecasts with new information? *arXiv*, 2026.

Zhiyuan Zeng, Jiashuo Liu, Siyuan Chen, Tianci He, Yali Liao, Yixiao Tian, Jinpeng Wang, Zaiyuan Wang, Yang Yang, Lingyue Yin, Mingren Yin, Zhenwei Zhu, et al. FutureX: An advanced live benchmark for LLM agents in future prediction. In *International Conference on Learning Representations*, 2026.

Andy Zou, Tristan Xiao, Ryan Jia, Joe Kwon, Mantas Mazeika, Richard Li, Dawn Song, Jacob Steinhardt, Owain Evans, and Dan Hendrycks. Forecasting future world events with neural networks. In *Advances in Neural Information Processing Systems*, 2022.

# A Detailed Experimental Protocols

## A.1 Prospective registration and timing

Each FIELD event is registered before forecasting with a fixed event identity, sport, participant ordering, scheduled start time, terminal resolution phase, and sport-specific category spaces for the five forecast targets. For an event scheduled at time $T$, FIELD collects forecasts at

$$H_1 = T - 24\text{h}, \qquad H_2 = T - 2\text{h}.$$

Each nominal horizon opens a 30-minute execution window. Public information retrieved during that window may be used, but the final submitted forecast must be received before the window closes. FIELD never regenerates, backfills, or replaces a forecast after its deadline.

The registry stores scheduled times in UTC together with the source-local representation and time-zone provenance used for conversion. If an official schedule changes, the event follows the frozen rescheduling policy described in Appendix B. Available execution timestamps and archive pointers support temporal auditing; a separate Primary finalization timestamp was not recorded.

The seven evaluated systems are GPT-5.6-sol, Claude-Opus-5, Gemini-3.8-flash, Grok-4.6, Qwen3.8-max, DeepSeek-v4-pro, and Kimi-K3.

Each is evaluated under both Native and Controlled settings. System names denote archived deployment identifiers, not independently verified immutable checkpoints. The supplement records requested and returned identifiers, route variants, and available budgets in its execution tables.

## A.2 Native forecasting

The **Native** setting evaluates each provider's end-to-end deployed system, including native information-access capabilities when they are exposed by that deployment. FIELD supplies the forecasting task and records the returned system output, but does not replace provider-native retrieval with the Controlled retrieval stack.

Providers differ in which search, citation, and tool-use artifacts they expose. FIELD records only externally visible material and never reconstructs opaque internal state or private reasoning. Provider-specific transport or continuation recovery is permitted only inside the original execution window and may not alter an already submitted forecast or create a post-deadline replacement.

## A.3 Controlled forecasting

The **Controlled** setting places each model backbone in the same external forecasting harness. The interaction follows

$$\textit{plan} \rightarrow \textit{initial retrieval} \rightarrow \textit{source selection} \rightarrow \textit{follow-up retrieval} \rightarrow \textit{final forecast}.$$

The planning stage may request at most eight initial search queries. After observing the returned search results, the model may open at most eight retrieved sources and issue at most four follow-up searches. Retrieved content is bounded to 6,000 visible characters per page and 72,000 visible characters in total; up to 12,000 characters of prior interaction history are retained in the bounded model context.

The harness executes authorized retrieval actions, assigns source identifiers, enforces budgets and deadlines, archives delivered observations, and applies the same frozen filtering rules for all model backbones. The model remains responsible for deciding what to search, which sources to inspect, how to use the evidence, and what probabilities and quantiles to submit. Retrieval budgets are not reset by retries or batching.

## A.4 Sequential forecasting across horizons

The protocol defines the two horizons as a sequential forecasting trajectory rather than two independent prediction attempts. When delivery is verified, the $H_2$ context may include the model's

accepted $H_1$ forecast together with archived material that was actually carried forward from the earlier run; previously delivered evidence is not re-fetched from the later web state. Realized execution is checked from the archived lineage rather than assumed from the nominal protocol.

The model may then seek additional information and submit a new five-target forecast. A valid $H_2$ forecast remains independently scoreable if $H_1$ is unavailable. Primary R4 is computed on the frozen matched numerical $H_1$–$H_2$ as-run support; membership in that numerical pair set does not by itself certify delivery of carried $H_1$ context. W4 requires lineage-verified delivered $H_1$ context, and the STEP R4 manipulation check likewise uses verified delivered own-arm $H_1 \to H_2$ lineages.

### A.5 Archiving and lossless parsing

Raw provider responses are preserved before parsing. FIELD permits only deterministic, content-preserving normalization required to recover a unique submitted payload, such as removing an unambiguous presentation wrapper. Both raw and normalized forms are retained.

Forecast parsing never uses another language model to repair numerical predictions, select among semantically different candidate answers, renormalize probability vectors, sort crossing quantiles, clip values, or impute missing targets. A forecast therefore enters evaluation only through the values that were actually submitted.

## B Truth Resolution and Cross-Sport Target Definitions

### B.1 Official-first, target-level truth resolution

FIELD follows an **official-first, auditable multi-source fallback** policy. Whenever a governing competition, league, federation, official statistics feed, official play-by-play, box score, gamebook, or official post-event report unambiguously resolves a target, that authoritative record is sufficient.

If an authoritative post-event source is unavailable or insufficient for a particular target, FIELD permits fallback resolution when at least two reasonably independent and reliable evidence families corroborate the same target value. These may include reputable professional sports databases, major sports or news organizations, or official club reports. Multiple URLs that clearly derive from the same upstream record are not counted as independent corroboration.

Truth is resolved *target by target*. An event may therefore have resolved Outcome, Margin, and Total values while $T1$ or $T2$ remains unresolved. Unresolved targets are never inferred from nearby game states or from the final score. In particular, FIELD does not infer a halftime state from the final score, infer a first-scoring side from the winner, or use model predictions to reconstruct truth.

The registry fixes `side_a`, `side_b`, competition identity, and event orientation before forecasting. All truth values are resolved against that frozen orientation. Later corrections are recorded append-only rather than silently overwriting the provenance of the original resolution.

### B.2 Outcome and final-score targets

Let $S_a$ and $S_b$ denote the official final scores of the two registered sides at the event's frozen terminal resolution phase. FIELD defines

$$\text{Margin} = S_a - S_b, \tag{1}$$

$$\text{Total} = S_a + S_b. \tag{2}$$

Football resolves at the end of regulation including stoppage time. Basketball includes officially played overtime, and baseball includes extra innings through the official final.

Outcome is

$$\{\texttt{side_a}, \texttt{draw}, \texttt{side_b}\}$$

for football and

$$\{\texttt{side_a}, \texttt{side_b}\}$$

for basketball and baseball.

Table 3: **Frozen truth semantics for the five FIELD targets.** All values use the pre-registered `side_a`/`side_b` orientation.

| Target | Resolved truth type | Truth definition |
|---|---|---|
| Outcome | Categorical | Final result at the registered terminal phase; football allows a draw. |
| Margin | Scalar | Final $S_a - S_b$. |
| Total | Scalar | Final $S_a + S_b$. |
| T1 | Categorical | First official scoring side; football additionally allows `no_score`. |
| T2 | Categorical | Leader/tie state at football halftime, basketball end-Q2, or baseball after five completed innings. |

### B.3 Match-Trajectory targets

The Match-Trajectory facet contains exactly two targets.

**$T1$: first scoring side.** For football, $T1$ is the side credited with the first official goal; a completed scoreless match resolves to `no_score`. For basketball, $T1$ is the side credited with the first official scoring play. For baseball, it is the side credited with the first official run. Thus,

$$\mathcal{Y}_{T1} = \begin{cases} \{\texttt{side_a}, \texttt{side_b}, \texttt{no_score}\}, & \text{football,} \\ \{\texttt{side_a}, \texttt{side_b}\}, & \text{basketball/baseball.} \end{cases} \tag{3}$$

**$T2$: midpoint state.** $T2$ is the categorical leader state

$$\{\texttt{side_a_leading}, \texttt{tied}, \texttt{side_b_leading}\}.$$

The checkpoint is sport specific: football uses the official halftime score, basketball uses the end of the second quarter, and baseball uses the score after five completed innings. If the corresponding midpoint state cannot be reliably resolved, $T2$ remains unresolved.

### B.4 Schedule changes and unresolved events

An ordinary late physical start does not change the registered forecast windows when the official scheduled start time itself is unchanged. If an event is officially rescheduled before forecasting, its future windows are recomputed mechanically from the corrected scheduled start and the revision is retained in provenance.

Forecasts already collected for an earlier scheduled time are never silently transferred to a materially rescheduled event. Cancelled events are not scored. A suspended contest that later resumes as the same official event retains its event identity and is resolved from the completed official record. If the event or an individual target cannot be resolved reliably, that target remains unresolved rather than being imputed.

## C Scoring, Analysis Populations, and Statistical Procedures

### C.1 Target-level proper losses

Outcome, $T1$, and $T2$ use normalized Brier loss. For a probability vector $\mathbf{p}$ and one-hot realized outcome $\mathbf{y}$,

$$L_{\text{Brier}}(\mathbf{p}, \mathbf{y}) = \frac{1}{2} \sum_{k=1}^{K} (p_k - y_k)^2. \tag{4}$$

A categorical submission is numerically valid only when every probability lies in $[0, 1]$ and $|\sum_k p_k - 1| \leq 10^{-6}$. FIELD does not renormalize invalid submissions. This validity tolerance differs from the $10^{-12}$ threshold for Outcome–Margin coherence violations.

Margin and Total use finite-quantile WIS. Let

$$q_{.10} \le q_{.25} \le q_{.50} \le q_{.75} \le q_{.90}$$

denote the submitted quantiles and $y$ the realized scalar value. For a central interval $[l, u]$ with miscoverage $\alpha$,

$$\mathrm{IS}_\alpha(l, u; y) = (u - l) + \frac{2}{\alpha}(l - y)\mathbb{I}[y < l] + \frac{2}{\alpha}(y - u)\mathbb{I}[y > u]. \tag{5}$$

The finite-quantile WIS is

$$L_{\mathrm{WIS}} = \frac{0.5|y - q_{.50}| + 0.1\,\mathrm{IS}_{0.2}(q_{.10}, q_{.90}; y) + 0.25\,\mathrm{IS}_{0.5}(q_{.25}, q_{.75}; y)}{2.5}. \tag{6}$$

Quantiles must be finite and non-crossing. They are evaluated without sorting, clipping, rounding, imputation, or semantic repair.

### C.2 Frozen historical reference and reference-relative score

Because Brier and WIS operate on different numerical scales, every target is normalized against a frozen historical reference forecast. The reference is constructed separately for each of the $3 \times 5 = 15$ sport–target cells using only events resolved before `2026-09-07T17:00:00Z`. Each frozen cell contains 120 historical observations. Events from either prospective study never enter this reference.

The reference is sport–target specific, not competition matched. In particular, the basketball reference comprises 120 WNBA regular-season games, whereas the Primary basketball evaluation contains 7 FIBA Women's World Cup, 8 BBL Pokal, and 3 Basketball Champions League qualification games. Reference and evaluation competition compositions are provided in the supplement.

For categorical targets, the reference forecast is the empirical historical category distribution. For Margin and Total, it uses the empirical $q_{.10}, q_{.25}, q_{.50}, q_{.75}, q_{.90}$ quantiles, following the empirical inverse-CDF (nearest-rank) convention, $q_\tau = \inf\{x : F_n(x) \ge \tau\}$.

For model $m$, target $d$, and eligible event support $I$, define

$$\mathrm{Score}^{\mathrm{ref}}_{m,d}(I) = 100\left(1 - \frac{\sum_{i \in I} L^{\mathrm{model}}_{i,d}}{\sum_{i \in I} L^{\mathrm{ref}}_{i,d}}\right). \tag{7}$$

The model and reference are always evaluated on exactly the same event support. A value of zero matches the frozen historical reference, positive values indicate lower aggregate loss, and negative values indicate worse performance. FIELD does not average per-event relative scores and does not apply participant-dependent min–max normalization or clipping.

This normalization has the conventional mathematical form of a forecasting skill score; we use the term *reference-relative score* to avoid ambiguity with model or agent "skills" in the LLM literature.

### C.3 Hierarchical aggregation

Aggregation is performed only after target-level reference normalization. Within each sport–horizon–setting cell,

$$\mathrm{RS}_{\mathrm{Outcome}} = \mathrm{RS}_{\mathrm{outcome}}, \tag{8}$$

$$\mathrm{RS}_{\mathrm{Distribution}} = \frac{1}{2}\left(\mathrm{RS}_{\mathrm{margin}} + \mathrm{RS}_{\mathrm{total}}\right), \tag{9}$$

$$\mathrm{RS}_{\mathrm{Trajectory}} = \frac{1}{2}\left(\mathrm{RS}_{T1} + \mathrm{RS}_{T2}\right), \tag{10}$$

$$\mathrm{RS}_{\mathrm{cell}} = \frac{1}{3}\left(\mathrm{RS}_{\mathrm{Outcome}} + \mathrm{RS}_{\mathrm{Distribution}} + \mathrm{RS}_{\mathrm{Trajectory}}\right). \tag{11}$$

Table 4: **Raw predictive losses on the frozen Balanced support.** Entries report mean raw loss, equally averaged over the four setting–horizon cells within each sport–target combination. Outcome, T1, and T2 use Brier loss; Margin and Total use WIS. The Reference row reports the corresponding frozen historical-reference loss evaluated on the same cell supports. Lower is better. Values are directly comparable within a sport–target column, but raw losses should not be compared across targets with different loss functions or scales. Exact per-cell losses and support sizes are provided in the supplementary material.

| **Model** | **Outcome** | **Margin** | **Total** | **T1** | **T2** |
|---|---|---|---|---|---|
| *Football* | | | | | |
| Gemini | 0.287 | 0.866 | 0.955 | 0.246 | 0.305 |
| Kimi | 0.283 | 0.858 | 0.950 | 0.246 | 0.304 |
| DeepSeek | 0.283 | 0.860 | 0.967 | 0.245 | 0.302 |
| GPT | 0.283 | 0.884 | 0.982 | 0.249 | 0.304 |
| Grok | 0.286 | 0.889 | 0.983 | 0.247 | 0.307 |
| Claude | 0.287 | 0.887 | 1.016 | 0.254 | 0.310 |
| Qwen | 0.291 | 0.876 | 1.001 | 0.254 | 0.313 |
| Reference | 0.326 | 1.066 | 0.983 | 0.267 | 0.334 |
| *Basketball* | | | | | |
| Gemini | 0.097 | 8.765 | 12.264 | 0.237 | 0.236 |
| Kimi | 0.120 | 8.847 | 11.979 | 0.216 | 0.245 |
| DeepSeek | 0.134 | 9.366 | 12.251 | 0.217 | 0.259 |
| GPT | 0.131 | 9.075 | 12.164 | 0.226 | 0.231 |
| Grok | 0.133 | 8.618 | 12.332 | 0.230 | 0.239 |
| Claude | 0.134 | 9.787 | 12.562 | 0.244 | 0.246 |
| Qwen | 0.140 | 10.008 | 13.271 | 0.230 | 0.248 |
| Reference | 0.279 | 13.507 | 13.524 | 0.245 | 0.334 |
| *Baseball* | | | | | |
| Gemini | 0.229 | 2.452 | 2.875 | 0.245 | 0.301 |
| Kimi | 0.232 | 2.466 | 2.959 | 0.252 | 0.294 |
| DeepSeek | 0.230 | 2.445 | 2.952 | 0.245 | 0.296 |
| GPT | 0.234 | 2.494 | 2.865 | 0.245 | 0.304 |
| Grok | 0.230 | 2.449 | 2.950 | 0.243 | 0.297 |
| Claude | 0.226 | 2.464 | 3.007 | 0.246 | 0.299 |
| Qwen | 0.234 | 2.488 | 2.923 | 0.250 | 0.320 |
| Reference | 0.250 | 2.624 | 2.780 | 0.245 | 0.303 |

Table 5: **Setting-specific Forecast Scores.** Native and Controlled scores use the same cell-specific shared-support construction as Table 2; Overall is their equal-weight average. Setting-specific differences vary in both magnitude and direction across models: Qwen scores higher under Controlled than Native (13.75 vs. 9.84), whereas GPT shows the reverse pattern (13.86 vs. 15.18).

| Model | Native | Controlled | Overall |
|---|---|---|---|
| Gemini-3.8-flash | 16.62 | 15.60 | 16.11 |
| Kimi-K3 | 14.84 | 16.04 | 15.44 |
| DeepSeek-v4-pro | 14.35 | 14.74 | 14.54 |
| GPT-5.6-sol | 15.18 | 13.86 | 14.52 |
| Grok-4.6 | 14.83 | 14.05 | 14.44 |
| Claude-Opus-5 | 12.23 | 13.56 | 12.90 |
| Qwen3.8-max | 9.84 | 13.75 | 11.80 |

The two horizons are equally weighted within each sport. Football, basketball, and baseball then receive equal macro weight regardless of their raw event counts. Native and Controlled are reported separately, and the final Forecast Score is

$$\mathrm{ForecastScore} = \frac{1}{2}\left(\mathrm{Score}_{\mathrm{Native}} + \mathrm{Score}_{\mathrm{Controlled}}\right). \tag{12}$$

R2–R4 and W1–W4 never enter these equations. Equal setting weights define a balanced benchmark summary rather than equivalent execution environments. Setting-specific Forecast Scores are reported in Table 5.

Table 6: **Numerical forecast coverage in the initial evaluation population.** Coverage is scored target records divided by truth-eligible target opportunities in the archived Primary analysis.

| Model | Coverage (%) |
|---|---|
| Gemini-3.8-flash | 83.2 |
| Kimi-K3 | 96.9 |
| GPT-5.6-sol | 95.5 |
| DeepSeek-v4-pro | 92.1 |
| Grok-4.6 | 98.0 |
| Claude-Opus-5 | 94.2 |
| Qwen3.8-max | 98.4 |

### C.4 Analysis Populations and Shared Support

The registered evaluation universe contains 236 prospective events. FIELD fixes the initial 232-event evaluation population (127 football, 18 basketball, and 87 baseball events) without using model performance; we retain the name **Primary** for its archived analysis outputs. An event is removed from this population only when a required setting–horizon opportunity is structurally empty for the entire seven-system panel.

Within the Primary analysis, each system is evaluated on its own valid target-level forecast support with resolved truth. Invalid or missing forecasts are not assigned fabricated numerical losses; instead they are excluded from the numerical loss sum and reduce the reported forecast coverage. In particular, no arbitrary finite sentinel is assigned to an invalid or missing Distribution forecast.

The main-text Forecast Score comparison uses the frozen **availability-balanced analysis**, while the larger Primary analysis is retained as a sensitivity comparison. For each sport $\times$ setting $\times$ horizon $\times$ target cell, the event support is restricted to the exact intersection on which all seven systems have valid scoreable forecasts. The frozen availability-balanced analysis contains 60 target cells, spans the union of 226 unique events, and contains 3,203 common-support event–cell positions per model, or 22,421 matched model-target records across the seven systems. The 226-event union is not an event-level complete-case cohort: membership is frozen separately within each target cell. Pairwise model contrasts use their own exact pair-specific common valid support.

Across the 232 Primary events there are 6,496 model–setting–event–horizon run opportunities and 32,480 target opportunities. The archived valid-support analysis contains 30,546 scored target records and 1,934 null losses. Null losses are retained as availability observations rather than replaced by failure penalties; valid targets from partial runs remain usable.

These fixed populations and their existing outputs are accepted for descriptive as-run comparison, not as an execution-equivalent official leaderboard. Observed score ordering is not an isolated measure of intrinsic model capability. Execution strata with supported, differing, or unverified opportunities are retained without result-based selection; event-support balancing does not remove this execution heterogeneity.

The diagnostics shown in Table 2 use different denominators. R2 ECE is aggregated over evaluable categorical cells; R3 pools evaluable Outcome–Margin runs and counts violations above $10^{-12}$. R4 summarizes matched numerical temporal support, which alone does not certify shared carried context. Controlled W1 averages observed plan-topic pursuit within runs before macro aggregation, whereas W2 reports requested queries per run. Neither the seven-model Balanced mask nor a single common denominator is imposed on all these diagnostics.

Workflow observability is independent of numerical forecast validity. A valid forecast can therefore contribute to R1 even when a Native provider does not expose enough workflow information for a particular W diagnostic.

### C.5 Pairwise comparisons and uncertainty

Direct model–model contrasts use exact pairwise common support. For each model pair, the score difference is recomputed only on event–target records for which both systems have valid losses and a valid frozen-reference loss. The reference is evaluated on that same intersection.

Table 7: **Adjacent-model uncertainty on the Balanced Forecast Score.** Differences are computed on the same frozen seven-model shared support as Table 2, with the higher-ranked system listed first. Intervals are pointwise 95% percentile event-cluster bootstrap intervals from 2,000 replicates and are not multiplicity-adjusted.

| Pair | Difference | 95% CI |
|---|---|---|
| Gemini-3.8-flash − Kimi-K3 | +0.67 | [−1.01, +2.63] |
| Kimi-K3 − DeepSeek-v4-pro | +0.90 | [−0.99, +3.33] |
| DeepSeek-v4-pro − GPT-5.6-sol | +0.02 | [−3.57, +2.77] |
| GPT-5.6-sol − Grok-4.6 | +0.08 | [−1.55, +1.67] |
| Grok-4.6 − Claude-Opus-5 | +1.55 | [−2.82, +5.59] |
| Claude-Opus-5 − Qwen3.8-max | +1.10 | [−1.88, +4.90] |

Uncertainty is estimated with 2,000 percentile bootstrap replicates. Events, rather than individual target rows, are the resampling units, and resampling is stratified by sport so that all targets, settings, and horizons belonging to a sampled event move together. The analysis uses the fixed seed `20260909`.

For the main Balanced ranking, Table 7 reports pointwise uncertainty for adjacent systems. All six intervals include zero, so the current sample does not reliably separate closely matched neighboring ranks. Ranking stability under alternative analysis choices is assessed separately in Appendix C.11. Continued prospective operation of FIELD will increase event support and sharpen these fine-grained comparisons over time.

### C.6 Trajectory Associations and Temporal Updating

The temporal analysis in Figure 3C compares forecasts for the same event and categorical target at $H_1$ and $H_2$. For Outcome, T1, and T2, the exact intersection of the frozen Balanced supports contains 643 event–setting–target positions. Each position contributes one $H_1$–$H_2$ pair per system, giving 4,501 model-level pairs across 141 events. The five-target update composite uses a different support of 1,491 matched runs from 139 events; the Controlled information-regime comparison uses 808 matched runs from 120 events.

Top-label persistence and update summaries are macro-averaged over the frozen analysis cells, rather than pooled across all records. The categorical loss change is measured in raw half-Brier loss, not reference-relative score. Intervals use 2,000 sport-stratified event-cluster bootstrap replicates with seed 20260909. The productive-update retrieval screen conditions on the corresponding $H_1$ feature and applies Benjamini–Hochberg correction within the pre-specified family.

The trajectory–Outcome analysis in Figure 3A asks whether more accurate trajectory forecasts accompany better Outcome forecasts. For sport $s$, let $\mu^{\mathrm{ref}}_{s,d}$ be the frozen reference mean loss for target $d$ on the Primary truth support. We use $Q_O = -L_O/\mu^{\mathrm{ref}}_{s,O}$ and $Q_T = -\frac{1}{2}(L_{T1}/\mu^{\mathrm{ref}}_{s,T1} + L_{T2}/\mu^{\mathrm{ref}}_{s,T2})$. Larger values indicate better forecasts; binary Outcome correctness is coded as 0/1. Continuous trajectory and behavior predictors are standardized using their population standard deviation (ddof 0), while binary T1/T2 correctness indicators remain 0/1.

The trajectory regressions include event, model, and setting fixed effects. Controlled behavior regressions include event and model fixed effects. Standard errors are clustered by event with the finite-cluster correction $G/(G-1)$, and 95% intervals use the normal critical value 1.959964. Behavioral tests receive Benjamini–Hochberg correction separately within each horizon × lens family. These models estimate associations rather than causal effects.

Table 8 reports the trajectory regressions. Separately, Table 9 compares the observable behavior of the Outcome-best system with that of other systems forecasting the same event under Controlled $H_2$.

### C.7 Forecast–Workflow Associations and Case Evidence

The atlas tests all 16 pairings between R1–R4 and W1–W4. Each pairing uses a fixed measurement of the forecast dimension and a compatible workflow measurement, listed in Table 10. These

Table 8: **Trajectory–Outcome quality associations.** The response is negative reference-normalized Outcome loss and the predictor is trajectory quality standardized by its population standard deviation. Regressions include event, model, and setting fixed effects; intervals use event-clustered standard errors.

| Horizon | Records | Events | $\beta$ per SD | 95% CI |
|---|---|---|---|---|
| $H_1$ | 2,092 | 175 | 0.172 | [0.107, 0.237] |
| $H_2$ | 2,393 | 195 | 0.197 | [0.139, 0.256] |

Table 9: **Controlled $H_2$ Outcome-best versus same-event peers.** Descriptive differences compare the Outcome-best system in each event with the remaining systems forecasting the same event.

| Feature | Groups | Difference | 95% CI |
|---|---|---|---|
| Midpoint-state query | 146 | +0.0573 | [0.0060, 0.1109] |
| First-score query | 146 | +0.0508 | [0.0046, 0.1087] |
| Requested queries / run | 146 | -0.2359 | [-0.6280, 0.1005] |

measurements are chosen for the association analysis and differ from the model-level summaries in the main table. W1/W2 come from observable Controlled-run traces; W3/W4 come from the final sampled semantic evaluations. Each pairing uses its own complete observations, without replacing missing workflow measurements by zero.

For a complete run, R1 is $-\sum_d w_d L_d/\mu^{\mathrm{ref}}_{s,d}$, with $w_O = 1/3$ and $w_d = 1/6$ for the other four targets. R2 is the mean top-label confidence across Outcome, $T1$, and $T2$, used here as a categorical sharpness proxy rather than ECE. R3 is the negative Outcome–Margin coherence gap, and R4 is the matched $H_2 - H_1$ change in the R1 primitive. The reference means follow the trajectory-quality definitions above.

We fit an unweighted fixed-effects regression for each pairing, standardizing the workflow predictor by its population standard deviation. Cross-sectional regressions include event and model fixed effects, and horizon fixed effects except for W4. Temporal regressions relate matched forecast gains to compatible workflow measurements. Standard errors are clustered by event, and the 16 associations form one pre-specified Benjamini–Hochberg family.

For W3/W4, the regressions describe runs or pairs with sampled, adjudicable measurements. They are not inverse-inclusion-probability-weighted population estimates, and their intervals do not include a separate model of semantic-judgment error.

Table 11 reports the complete 16-cell atlas. None of the associations survives BH correction at $q < .05$. R3–W1, R3–W2, and R3–W4 have nominal $p < .05$ but each has $q \approx 0.25$. We therefore use the atlas to test for a stable cross-view mapping rather than to declare local nominal associations as discoveries.

Two supplementary analyses are kept separate from the 16-cell atlas. The first asks whether workflow summaries are associated with categorical calibration. It uses Controlled forecasts on frozen Balanced categorical support, with 42 model–sport–horizon cells for W1–W3 and 21 $H_2$ cells for W4. Thus, the observations are cell summaries rather than individual runs. ECE follows the construction in Section C.8, and the four Spearman correlations form a separate Benjamini–Hochberg family.

W3 evidence support has the strongest original association with ECE ($\rho = -0.407$, $q = .030$; Table 13). The dependence checks in Table 14, however, do not establish its robustness. For W3–ECE, the sport-stratified event-bootstrap 95% interval is $[-0.551, 0.062]$, and the exact model-block permutation gives $p = .505$. The correlation remains negative after removing any single model or horizon, but changes sign when Basketball is excluded. We therefore retain it as a local descriptive association rather than a robust cross-view finding. The table reports bootstrap and permutation checks for all four workflow dimensions.

Second, same-event pairwise process-similarity analyses compare planning, query, opened-domain, and evidence overlap with forecast similarity. The archived mismatch index provides descriptive ex-

Table 10: **Frozen primitives used in the cross-view coupling atlas.** R4 cells are temporal and therefore use change/$H_2$ workflow primitives rather than purely cross-sectional quantities.

| Dim. | Forecast primitive | Workflow primitive(s) |
|---|---|---|
| R1 | broad forecast quality | W1 pursuit; W2 semantic breadth; final W3/W4 support |
| R2 | categorical sharpness | W1 pursuit; W2 semantic breadth; final W3/W4 support |
| R3 | coherence quality | W1 pursuit; W2 semantic breadth; final W3/W4 support |
| R4 | broad $H_1 \rightarrow H_2$ gain | $\Delta$W1 pursuit; $\Delta$W2 semantic breadth; $H_2$ W3/W4 support |

Table 11: **Complete Forecast–Workflow association atlas.** Coefficients correspond to the frozen analysis primitives in Table 10. Workflow predictors are standardized within their eligible support; intervals use event-clustered uncertainty, and $q$ values apply Benjamini–Hochberg correction jointly across all 16 associations.

| Pair | Records | Events | $\beta$ | 95% CI | $p$ | $q$ |
|---|---|---|---|---|---|---|
| R1–W1 | 2137 | 217 | -0.00240 | [-0.00748, +0.00269] | 0.355 | 0.745 |
| R1–W2 | 2318 | 217 | +0.00317 | [-0.00193, +0.00827] | 0.223 | 0.658 |
| R1–W3 | 2179 | 215 | +0.00043 | [-0.00363, +0.00448] | 0.836 | 0.910 |
| R1–W4 | 672 | 167 | -0.00521 | [-0.01403, +0.00361] | 0.247 | 0.658 |
| R2–W1 | 2137 | 217 | +0.00017 | [-0.00147, +0.00181] | 0.838 | 0.910 |
| R2–W2 | 2318 | 217 | +0.00026 | [-0.00126, +0.00178] | 0.738 | 0.910 |
| R2–W3 | 2179 | 215 | +0.00043 | [-0.00093, +0.00179] | 0.537 | 0.910 |
| R2–W4 | 672 | 167 | +0.00106 | [-0.00127, +0.00340] | 0.373 | 0.745 |
| R3–W1 | 2137 | 217 | -0.00106 | [-0.00206, -0.00006] | 0.038 | 0.250 |
| R3–W2 | 2318 | 217 | -0.00108 | [-0.00215, -0.00001] | 0.049 | 0.250 |
| R3–W3 | 2179 | 215 | +0.00005 | [-0.00083, +0.00094] | 0.910 | 0.910 |
| R3–W4 | 672 | 167 | -0.00145 | [-0.00288, -0.00001] | 0.049 | 0.250 |
| R4–W1 | 675 | 120 | -0.00066 | [-0.00658, +0.00526] | 0.828 | 0.910 |
| R4–W2 | 808 | 120 | +0.00524 | [-0.00027, +0.01075] | 0.063 | 0.250 |
| R4–W3 | 772 | 120 | +0.00051 | [-0.00400, +0.00501] | 0.825 | 0.910 |
| R4–W4 | 511 | 118 | -0.00063 | [-0.00823, +0.00696] | 0.870 | 0.910 |

amples of alignment and tension between the two views. Table 12 shows representative mismatches in both directions: higher forecast quality with weaker measured workflow signals, and lower forecast quality with stronger measured workflow signals. These cases are descriptive and are not used to estimate population prevalence or define the 16-cell association family.

### C.8 R2: Calibration and Sharpness

For categorical targets, top-label confidence is $c_i = \max_k p_{ik}$, and correctness indicates whether the selected top category matches the resolved target. Top-category ties in the main R2 summary follow the registered category order. The ECE reported in Table 2 uses ten equal-width bins, $[0, .1), \ldots, [.8, .9), [.9, 1]$. Within each model–sport–setting–horizon–target cell,

$$\mathrm{ECE} = \sum_b \frac{n_b}{n} \left| \mathrm{acc}_b - \mathrm{conf}_b \right|,$$

where $n_b$ is the number of forecasts in bin $b$, $\mathrm{acc}_b$ is their realized top-label accuracy, and $\mathrm{conf}_b$ is their mean top-label confidence. The model-level summary reported in Table 2 macro-averages these cell-level ECE values and multiplies the result by 100.

The separate cross-view calibration analysis in Section C.7 instead pools eligible categorical targets within Controlled model–sport–horizon cells and retains its frozen right-closed bin convention, $[0, .1], (.1, .2], \ldots, (.9, 1]$; top-category ties follow the serialized probability-key order used by the frozen implementation.

For Margin and Total, FIELD reports coverage and width for the submitted central predictive intervals. The 50% interval is $[q_{.25}, q_{.75}]$ and the 80% interval is $[q_{.10}, q_{.90}]$. Coverage is the fraction

Table 12: **Illustrative forecast–workflow mismatches from the frozen case index.** Cases are descriptive examples selected from the frozen mismatch index and are not used to estimate prevalence. "High/low forecast" follows the index's R1 broad-quality classification. W1 is topic-pursuit rate and W3 is conditional evidence support among substantive sampled semantic units; parentheses report fully supported / substantive sampled units. Registered start times in UTC uniquely identify the corresponding events.

| Direction | Event / system | Horizon | Forecast signal | Workflow signal |
|---|---|---|---|---|
| Forecast strong, workflow weak | Kreisbau Kirchheim Knights–MHP RIESEN Ludwigsburg *2026-09-13 13:00 UTC; GPT-5.6-sol* | $H_2$ | High R1 broad quality | W1 pursuit $= 0.40$; W3 support $= 0.00$ (0/1) |
| Forecast strong, workflow weak | Atlanta Braves–Tampa Bay Rays *2026-09-08 23:15 UTC; Kimi-K3* | $H_2$ | High R1 broad quality | W3 support $= 0.00$ (0/1) |
| Workflow strong, forecast weak | Miami Marlins–New York Mets *2026-09-09 22:40 UTC; Qwen3.8-max* | $H_2$ | Low R1 broad quality | W1 pursuit $= 1.00$; W3 support $= 1.00$ (3/3) |
| Workflow strong, forecast weak | Al Khaleej–Al Nassr *2026-09-12 18:00 UTC; Claude-Opus-5* | $H_1$ | Low R1 broad quality | W1 pursuit $= 1.00$; W3 support $= 1.00$ (6/6) |

Table 13: **Controlled calibration-cell correlations.** ECE is computed over eligible Balanced categorical predictions pooled within model–sport–horizon cells. W4 is available only at $H_2$. The four Spearman correlations form a separate Benjamini–Hochberg family from the 16-cell atlas.

| Workflow | Cells | $\rho$ | $p$ | $q$ |
|---|---|---|---|---|
| W1 | 42 | -0.311 | 0.045 | 0.090 |
| W2 | 42 | -0.203 | 0.197 | 0.263 |
| W3 | 42 | -0.407 | 0.008 | 0.030 |
| W4 | 21 | +0.002 | 0.993 | 0.993 |

of truth-evaluable intervals containing the realized value, including endpoints; width is the upper quantile minus the lower quantile. These diagnostics do not modify R1 losses.

### C.9 R3: Forecast Consistency

R3 records numerical validity, quantile ordering, and necessary cross-target coherence. For Margin quantiles $\{q_{\tau_j}\}$, a necessary interval for the probability that `side_a` wins is

$$\ell = \max\left(\{0\} \cup \{1 - \tau_j : q_{\tau_j} > 0\}\right), \tag{13}$$

$$u = \min\left(\{1\} \cup \{1 - \tau_j : q_{\tau_j} \leq 0\}\right). \tag{14}$$

FIELD records the Outcome–Margin coherence gap

$$\max\{\ell - p_a,\, 0,\, p_a - u\},$$

and counts a violation only when this gap exceeds $10^{-12}$. This is a necessary coherence diagnostic, not a reconstructed joint distribution, and it never repairs or rescales the submitted forecast.

### C.10 R4: Temporal Updating

For a valid matched $H_1$–$H_2$ target pair,

$$\Delta L_d = L_{H_1,d} - L_{H_2,d}, \tag{15}$$

so positive values indicate lower realized loss at $H_2$.

Table 14: **Dependence sensitivity for workflow–ECE correlations.** Intervals use 2,000 sport-stratified event-cluster bootstrap replicates. Block permutation preserves sport/horizon structure while permuting model identity globally.

| Workflow | Bootstrap 95% CI | Block $p$ |
|---|---|---|
| W1 | $[-0.476, -0.003]$ | .816 |
| W2 | $[-0.471, +0.050]$ | .936 |
| W3 | $[-0.551, +0.062]$ | .505 |
| W4 | $[-0.375, +0.369]$ | .991 |

FIELD reports forecast movement separately from realized improvement. For complete five-target temporal pairs, the temporal-pair-level normalized update gain is

$$G_r = \frac{1}{3}\frac{\Delta L_{\text{Outcome}}}{\mu^{\text{ref}}_{\text{Outcome}}} + \frac{1}{6}\sum_{d\in\{\text{Margin},\text{Total},T1,T2\}} \frac{\Delta L_d}{\mu^{\text{ref}}_d}, \tag{16}$$

where $\mu^{\text{ref}}_d$ is the frozen reference mean loss for target $d$ on the corresponding Primary truth support. The model-level R4 summary reported in the main paper is 100 times the macro-averaged $G_r$.

A valid $H_2$ forecast without a matched $H_1$ remains eligible for its own R1 score but does not contribute to an R4 update pair.

### C.11 Ranking Sensitivity and Performance Profiles

The main aggregation weights are frozen before result analysis. We examine how the model ordering changes with analysis support, facet weights, reference normalization, and sport composition, while retaining the main shared-support comparison.

Changing analysis support from Primary to availability-balanced yields highly concordant rankings (Spearman $\rho = 0.964$), with only GPT and DeepSeek exchanging positions.

For facet weighting on the frozen Balanced population, we give the emphasized facet weight 0.50 and each remaining facet weight 0.25. Outcome-heavy, Distribution-heavy, and Trajectory-heavy aggregation all preserve the complete seven-model ordering under equal facet weights (Spearman $\rho = 1.000$ in each case).

A reference-free aggregation of within-cell raw-loss ranks also agrees closely with the formal ranking ($\rho = 0.964$), with only GPT and Grok exchanging positions.

For sport composition, we omit one sport at a time from the same Balanced population. Relative to the full three-sport ranking, Spearman correlations are 0.857 without Football, 0.857 without Basketball, and 0.964 without Baseball. DeepSeek leads when Basketball is removed; Gemini leads under the other two exclusions. High overall rank concordance therefore coexists with a change in the leading model.

**Evidence for cross-sport and calibration trends.** Tables 15 and 16 report the complete model-level profiles underlying the 21/21 cross-sport pattern and the 14/14 calibration pattern summarized in Section 4.2.

## D Workflow Diagnostic Rubrics and Validation

### D.1 Observability boundary

FIELD evaluates only workflow artifacts exposed at the system boundary. Controlled runs provide public planning text, explicit search queries, source-opening actions, delivered content, source identifiers, and retrieval status. Native runs contribute only material that the provider actually exposes through its interface. Opaque internal tool use and private reasoning are never reconstructed.

Table 15: **Cross-sport reference-relative score profiles on shared valid support.** Each row equally averages the two settings and horizons. Outcome exceeds both Distribution and Trajectory, and Margin exceeds Total, in all 21 model–sport rows.

| Model | Sport | Outcome | Distrib. | Traj. | Margin | Total |
|---|---|---|---|---|---|---|
| Gemini-3.8-flash | Football | 12.27 | 10.79 | 8.20 | 18.76 | 2.82 |
| Gemini-3.8-flash | Basketball | 65.07 | 22.23 | 16.11 | 35.23 | 9.22 |
| Gemini-3.8-flash | Baseball | 8.54 | 1.56 | 0.23 | 6.52 | -3.41 |
| Kimi-K3 | Football | 13.24 | 11.37 | 8.33 | 19.41 | 3.33 |
| Kimi-K3 | Basketball | 56.91 | 22.94 | 19.13 | 34.56 | 11.31 |
| Kimi-K3 | Baseball | 7.29 | -0.22 | 0.02 | 5.95 | -6.39 |
| DeepSeek-v4-pro | Football | 13.22 | 10.39 | 8.94 | 19.20 | 1.58 |
| DeepSeek-v4-pro | Basketball | 52.16 | 19.98 | 16.73 | 30.68 | 9.28 |
| DeepSeek-v4-pro | Baseball | 7.93 | 0.35 | 1.19 | 6.79 | -6.10 |
| GPT-5.6-sol | Football | 13.50 | 8.53 | 7.82 | 17.01 | 0.05 |
| GPT-5.6-sol | Basketball | 52.95 | 21.42 | 19.07 | 32.93 | 9.91 |
| GPT-5.6-sol | Baseball | 6.53 | 0.93 | -0.08 | 4.91 | -3.04 |
| Grok-4.6 | Football | 12.29 | 8.26 | 7.81 | 16.58 | -0.05 |
| Grok-4.6 | Basketball | 52.38 | 22.49 | 17.22 | 36.31 | 8.67 |
| Grok-4.6 | Baseball | 7.94 | 0.27 | 1.31 | 6.59 | -6.05 |
| Claude-Opus-5 | Football | 12.22 | 6.66 | 6.01 | 16.64 | -3.32 |
| Claude-Opus-5 | Basketball | 51.84 | 17.32 | 13.30 | 27.64 | 6.99 |
| Claude-Opus-5 | Baseball | 9.43 | -1.06 | 0.35 | 6.06 | -8.19 |
| Qwen3.8-max | Football | 10.83 | 7.83 | 5.51 | 17.55 | -1.88 |
| Qwen3.8-max | Basketball | 49.78 | 13.82 | 15.84 | 25.86 | 1.77 |
| Qwen3.8-max | Baseball | 6.30 | 0.04 | -3.79 | 5.13 | -5.06 |

Table 16: **Categorical calibration profiles.** Entries are percentages and macro-average the eligible sport–horizon–target cells within each model and setting on Primary valid-target support. Accuracy exceeds mean confidence in all 14 rows. ECE is the average of cell-level ECE values.

| Model | Setting | Accuracy (%) | Confidence (%) | ECE (%) |
|---|---|---|---|---|
| Gemini-3.8-flash | Native | 63.20 | 57.08 | 12.66 |
| Gemini-3.8-flash | Controlled | 61.56 | 56.51 | 11.49 |
| Kimi-K3 | Native | 60.97 | 57.66 | 9.66 |
| Kimi-K3 | Controlled | 62.19 | 57.61 | 10.96 |
| DeepSeek-v4-pro | Native | 59.56 | 57.44 | 7.83 |
| DeepSeek-v4-pro | Controlled | 61.47 | 57.35 | 9.02 |
| GPT-5.6-sol | Native | 62.63 | 56.91 | 12.10 |
| GPT-5.6-sol | Controlled | 62.66 | 56.14 | 11.92 |
| Grok-4.6 | Native | 59.99 | 55.63 | 9.73 |
| Grok-4.6 | Controlled | 62.48 | 55.78 | 11.65 |
| Claude-Opus-5 | Native | 56.39 | 54.65 | 10.03 |
| Claude-Opus-5 | Controlled | 60.19 | 55.41 | 12.08 |
| Qwen3.8-max | Native | 58.13 | 53.79 | 9.20 |
| Qwen3.8-max | Controlled | 59.97 | 55.22 | 10.97 |

For every extracted workflow unit, FIELD retains the exact public candidate span and the exact archived evidence material that was visible at prediction time. An $H_1$ fragment may enter an $H_2$ W4 record only if that literal archived material was actually carried forward.

### D.2 W1 AND W2: OBSERVABLE WORKFLOW TRACES

W1 **Information Needs** measures whether publicly stated information needs are followed by observable retrieval actions. The frozen implementation maps declared needs and later actions to a topic vocabulary using lexical matching, then computes pursuit over distinct declared topics within each run. This measures observable follow-through, not the quality of the plan or the success of retrieval; neither need counts nor planning verbosity is scored as planning quality.

W2 **Information Retrieval** describes exposed retrieval activity: search queries, page opens, delivered pages, source identifiers, outcomes, failures, and truncation. The main table reports requested queries per run. The cross-view analyses use a separate query-topic breadth feature, whose topic system is distinct from the W1 vocabulary. These quantities describe different aspects of retrieval, and larger values are not automatically better.

### D.3 W3 AND W4: SAMPLED AUTOMATED SEMANTIC EVIDENCE EVALUATION

W3 **Evidence Support** evaluates externally checkable factual claims in the public forecasting record. W4 **Revision Support** evaluates the stated factual basis for changing or retaining a forecast from $H_1$ to $H_2$. W4 may use only evidence literally carried from $H_1$ and evidence newly observed at $H_2$; it does not evaluate whether the resulting probability update was numerically beneficial, which belongs to R4.

Candidate spans come from a frozen extraction frame constructed from Gemini and Luna parser outputs. Subsequent source binding, duplicate resolution, eligibility filtering, and sampling are deterministic, retaining prediction-time material with auditable archive provenance.

We evaluate a fixed 30% stratified sample of the source-verifiable observable *Controlled* frame. Sampling is frozen before semantic judging within model $\times$ sport $\times$ dimension (W3/W4) strata. Each stratum records its population size $N$, sample size $n$, inclusion probability $n/N$, and survey weight $w = N/n$.

The sample contains 20,527 semantic units: 20,481 receive technically valid judgments, and 46 remain technical failures. These failures stay explicit in coverage accounting and are neither imputed nor replaced by a new sample.

A fixed Gemini-3.8-flash evaluator judges each sampled span against its archived prediction-time evidence under a frozen prompt. Its inputs exclude realized match outcomes, forecast losses, model ranks, post-event web evidence, and private reasoning. It returns one of seven labels: SUPPORTED, PARTIALLY SUPPORTED, CONTRADICTED, NOT SUPPORTED, UNRESOLVED, UNOBSERVABLE, or OUT OF SCOPE.

A compound span remains one evaluation unit. SUPPORTED requires support for all meaningful components and qualifiers. PARTIALLY SUPPORTED requires substantive support with an explicit remaining gap and no contradicted component. CONTRADICTED requires a contradiction within the matching scope. NOT SUPPORTED means that a complete observable archive does not support the claim, not that the claim is false. When the archive is insufficient for that negative judgment, the label remains UNRESOLVED.

For W4, a substantive label also requires an explicit textual connection to changing or retaining the $H_1$ forecast.

For a Controlled model $m$, sport $s$, and dimension $d \in \{\mathrm{W3}, \mathrm{W4}\}$, define the substantive denominator set $\mathcal{S} = \{\text{SUPPORTED}, \text{PARTIALLY SUPPORTED}, \text{CONTRADICTED}, \text{NOT SUPPORTED}\}$. The final conditional support rate is

$$\widehat{\mathrm{Support}}_{m,s,d} = \frac{\sum_i w_i \, \mathbb{I}[Y_i = \text{SUPPORTED}]}{\sum_i w_i \, \mathbb{I}[Y_i \in \mathcal{S}]}. \tag{17}$$

The rate measures full support among units assigned a substantive label. PARTIALLY SUPPORTED enters the denominator but contributes nothing to the numerator; it is also reported separately. UNRESOLVED, UNOBSERVABLE, OUT OF SCOPE, and technical failures remain separate coverage categories rather than zero-valued support judgments.

The model-level values in Table 2 are equal-weight averages of the three sport-level conditional support rates. All three sports are required, and a sport does not receive more weight because it has more observable units. Sport-level intervals use the frozen conditional run/W4-pair cluster bootstrap; the model-level table reports no confidence interval.

Final W3/W4 metrics are reported only for Controlled runs. Native values remain N/A because provider interfaces do not expose traces in a standardized way, even when some intermediate material is visible.

Table 17: **Human validation of W3/W4 semantic labels.** Agreement and macro-F1 use inverse-inclusion-probability weights from the frozen validation design; intervals use 5,000 stratified owner-cluster bootstrap replicates. Alpha and Beta remain separate references under the post-inspection dual-reference amendment.

| View | Reference | Agreement (95% CI) | Macro-F1 (95% CI) |
|---|---|---|---|
| All | Alpha | .906 [.845, .959] | .909 [.821, .956] |
| All | Beta | .895 [.828, .952] | .884 [.772, .943] |
| W3 | Alpha | .905 [.836, .964] | .884 [.728, .952] |
| W3 | Beta | .901 [.825, .966] | .858 [.673, .946] |
| W4 | Alpha | .910 [.829, .976] | .908 [.828, .971] |
| W4 | Beta | .850 [.754, .932] | .833 [.724, .922] |

### D.4 HUMAN VALIDATION OF THE AUTOMATED EVALUATOR

Human validation addresses two questions: whether the automated labels agree with independent human judgments, and how completely the extraction recovers relevant content from the source records. The semantic frame contains 168 technically valid W3/W4 units across 42 model $\times$ sport $\times$ dimension strata. Its estimates use the inverse-inclusion-probability weights from the frozen validation design. A separate frame of 168 owners audits extraction completeness and atomicity at the owner-record level.

For semantic validation, two annotators receive the same de-identified items in different deterministic orders and return independent labels. We report results against both references, Alpha and Beta. This dual-reference approach replaced the original single-adjudicated-gold plan after annotation and preliminary inspection of agreement/results; the amendment was **post-inspection and not preregistered**. All disagreements are retained, and neither reference is selected on the basis of more favorable evaluator performance.

The release pipeline de-identified the annotation packets. No separate attestation was collected concerning annotators' use of outside tools or exposure to admin-side answers.

**Semantic-label reliability.** Alpha and Beta agree on 94.0% of the 168 semantic units, with unweighted Cohen's $\kappa = .883$ (95% CI $[0.822, 0.939]$). Table 17 compares the automated evaluator with each reference using inverse-inclusion-probability weights from the frozen validation design. Agreement is 90.6% against Alpha and 89.5% against Beta; weighted macro-F1 is .909 and .884, respectively. Macro-F1 averages classwise F1 over the union of classes present in the reference or automated predictions; classes absent from both are omitted. Both references exceed the unchanged semantic validation thresholds (agreement $\geq .80$, macro-F1 $\geq .70$, and recall $\geq .50$ for every reference class with at least 10 sampled examples). These results validate semantic labeling on the sampled frame. Automated atomicity accuracy is not estimable because the direct-judge pipeline does not produce a frozen atomicity prediction.

The class-level audit is necessarily uneven. SUPPORTED, PARTIALLY SUPPORTED, UNRESOLVED, and OUT OF SCOPE have enough representation for the planned reference-class checks, and both references meet the recall threshold for classes with $n \geq 10$. CONTRADICTED appears only once in each human reference, while NOT SUPPORTED and UNOBSERVABLE do not appear in the semantic sample. We therefore do not infer reliable class-specific performance for those rare or absent labels. Abstention-like labels are retained in the seven-label confusion matrices rather than dropped to inflate agreement.

**Dependence-aware uncertainty.** Confidence intervals in this human-validation section use 5,000 percentile bootstrap replicates with seed 20260921. Resampling is stratified by the frozen semantic strata and performed at the owner-cluster level, carrying together all sampled units belonging to the same owner; inverse-inclusion-probability weights from the frozen validation design are retained. Owner-audit intervals analogously resample owner rows within their frozen strata. These are design-aware bootstrap intervals conditional on the validation sample, not IID-row or exact finite-population intervals.

Table 18: **STEP components and their FIELD motivation.** The table summarizes the intervention logic; detailed instructions follow below.

| Comp. | Operation | What it changes | FIELD motivation |
|---|---|---|---|
| S | State/Trajectory modeling | Represents plausible event states and trajectories before endpoint probabilities are finalized. | Endpoint quality is associated with trajectory quality. |
| T | Transition-aware updating | At $H_2$, revisits earlier premises before deciding what to retrieve and revise. | Short-horizon forecasting is a sequential refinement problem. |
| EP | Evidence-to-Probability projection | Connects evidence strength, conflict, redundancy, and uncertainty to probabilities/quantiles. | Observable evidence quality is not itself a surrogate for forecast quality. |

**Extraction completeness and atomicity.** The owner audit examines completeness of the extracted material at the owner-record level. Weighted by inverse inclusion probability, W3 completeness is 83.3% YES (95% CI $[77.7, 88.9]$), 4.2% NO, 10.9% UNCERTAIN, and 1.7% NO-SOURCE. W4 is less complete: 28.9% YES ($[19.6, 38.6]$), 52.4% NO ($[41.7, 63.1]$), and 18.7% UNCERTAIN ($[11.1, 27.3]$).

These percentages describe owner-level judgments, not the fraction of individual semantic units omitted. They show that W4 extraction does not recover all observable revision-support objects. Accordingly, Table 2 measures support conditional on an extracted revision unit receiving a substantive label in $\mathcal{S}$, rather than coverage of the full revision-support record.

The auxiliary atomicity audit reports weighted YES/UNCERTAIN proportions of 57.9%/40.4% for W3 and 87.0%/12.6% for W4, with essentially no explicit NO judgments. These are human characterizations of the sampled owner records, not accuracy estimates for an automated atomicity predictor.

# E Exploratory Prospective STEP Study

## E.1 From FIELD analysis to STEP

The intervention study translates three observations from the primary evaluation into a structured forecasting procedure: (i) endpoint quality is associated with trajectory quality; (ii) $H_2$ is a sequential updating problem; and (iii) evidence gathering does not automatically translate into calibrated probabilistic judgment.

STEP consists of three components: **S**tate/Trajectory modeling, **T**ransition-aware updating, and **E**vidence-to-**P**robability projection.

## E.2 State/Trajectory modeling

Before committing to the final endpoint forecast, S requires a compact event-state representation covering relative competitive conditions, first-score dynamics, the midpoint state, a main plausible trajectory, and a credible alternative trajectory. The trajectory is used to inform Outcome, Margin, and uncertainty rather than as a deterministic rule.

## E.3 Transition-aware updating

T is applied only at $H_2$. The forecaster revisits the important premises carried from $H_1$ and classifies them as still supported, newly resolved or changed, contradicted, or unresolved. It then decides what additional information, if any, should be retrieved and whether each forecast target should be retained, increased, decreased, widened, narrowed, or materially revised.

T does not require a categorical reversal and does not prescribe a fixed set of "late" information categories.

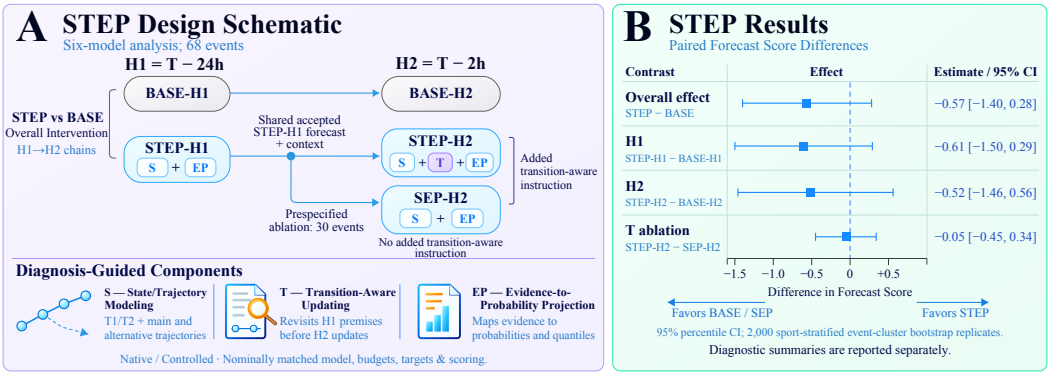


Figure 4: **Prospective STEP study. A**, STEP combines State/Trajectory modeling (S), Evidence-to-Probability projection (EP), and transition-aware updating (T); the pre-specified SEP ablation removes the added transition-aware instruction at $H_2$ while retaining the accepted STEP-$H_1$ history. **B**, paired Forecast Score differences for the overall STEP–BASE effect, the two forecast horizons, and the pre-specified T ablation. Intervals are 95% percentile bootstrap confidence intervals.

### E.4 Evidence-to-Probability Projection

EP asks the forecaster to connect evidence explicitly to the relevant state or target, considering direction, strength, redundancy, independence, conflicting evidence, and residual uncertainty. Submitted probabilities and quantiles should reflect this evidence strength and are followed by the same necessary coherence checks used elsewhere in FIELD.

### E.5 Experimental Arms and Temporal Lineage

The intervention is evaluated prospectively on 68 newly registered events: 30 football, 11 basketball, and 27 baseball events.

Every event contains parallel BASE and STEP trajectories:

$$\mathrm{BASE}_{H_1} \rightarrow \mathrm{BASE}_{H_2},$$

and

$$\mathrm{STEP}_{H_1}^{S+EP} \rightarrow \mathrm{STEP}_{H_2}^{S+T+EP}.$$

A pre-specified subset of 30 events additionally includes

$$\mathrm{SEP}_{H_2}^{S+EP},$$

which is required to receive the same accepted STEP-$H_1$ forecast and carried context as the full STEP-$H_2$ arm when that shared prior is verified; otherwise the corresponding ablation comparison is not treated as lineage-matched. On verified shared-prior, common-valid support, STEP-$H_2$ versus SEP-$H_2$ targets the incremental contribution of T after an already treated $S + EP$ history. It does not separately identify the effects of S and EP.

### E.6 Opportunity Parity

The intervention design nominally matches forecasting resources across BASE, STEP, and SEP: model backbone, Native/Controlled definition, five targets, deadlines, output schema, truth definitions, retrieval/context budgets, model-call opportunities, provider routes, retry policy, and scoring procedure. This design does not imply identical realized provider availability, retrieval delivery, or prior-context delivery across every run; those quantities are checked from the archived execution record.

In Controlled runs, STEP changes the instructions used at the existing planning, source-selection, and finalization stages without adding retrieval rounds. In Native runs, the same forecasting principles are incorporated into the provider-native task request without recreating the Controlled agent graph.

### E.7 INTERVENTION ANALYSIS

The primary intervention endpoint is the paired STEP–BASE difference in Forecast Score on exact common valid support. The analysis also reports:

- the $H_1$ effect of $S + EP$;
- the full $H_2$ effect of $S + T + EP$;
- target- and facet-level effects;
- Native/Controlled and model-level heterogeneity;
- forecast coverage and execution validity; and
- the STEP-$H_2$ versus SEP-$H_2$ contrast on the fixed ablation subset.

R2–R4 and W1–W4 are secondary diagnostic and manipulation-check outcomes rather than additional success criteria. The pre-specified primary endpoint remains the paired Forecast Score difference. Increases in search volume, plan length, trajectory terminology, or forecast movement are interpreted only as descriptive consequences of the intervention.

For each contrast, we first score both arms on their exact common-valid target support using the frozen target–facet–horizon–sport–setting aggregation. We then compute a score difference for each model and average those differences equally across the six analyzable models to obtain the panel estimate.

Intervals use 2,000 sport-stratified event-cluster bootstrap replicates with seed `20260909`; all paired records from a sampled event move together. Only the overall STEP–BASE endpoint is confirmatory. Intervals for secondary contrasts are pointwise and are not adjusted for multiple comparisons.

### E.8 STEP RESULT SCOPE, HETEROGENEITY, AND MANIPULATION CHECKS

**Analysis population and execution scope.** The six models included in the formal STEP effect analysis are GPT-5.6-sol, Gemini-3.8-flash, Grok-4.6, Qwen3.8-max, DeepSeek-v4-pro, and Claude-Opus-5. Kimi-K3 was excluded from STEP effect estimation under a scope decision fixed before effect analysis: provider/route failures prevented the intended two-horizon intervention from being realized on comparable support. No $H_1$-only, Controlled-only, pooled, or model-specific Kimi estimate is included. This restriction applies only to the intervention study and does not modify the primary FIELD evaluation.

The STEP–BASE primary analysis uses exact event–model–setting–horizon– target common support and the same frozen scoring/reference hierarchy as the primary FIELD evaluation. The pre-specified overall estimate is $-0.5676$ Forecast Score points with 95% CI $[-1.3953, 0.2777]$ across 68 events and 8,095 paired target positions. $H_1$ and $H_2$ estimates are $-0.6140$ ($[-1.5016, 0.2900]$) and $-0.5212$ ($[-1.4591, 0.5574]$), respectively. The fixed 30-event STEP–SEP $H_2$ ablation is $-0.0456$ ($[-0.4500, 0.3373]$).

**Exploratory responder stability.** Event-level gains are exploratory summaries derived from the frozen paired support and are not additive components of the primary estimator. The 59 events with complete event-level gains each contain all 120 paired positions across six models, two settings, two horizons, and five targets; the remaining nine events retain partial paired support and remain included in the 68-event primary analysis. Among the 59 complete events, 28 gains are positive and 31 are negative. $H_1$ and $H_2$ gains have Spearman $\rho \approx 0.63$; 42/59 events retain the same sign, and 20/28 positive-gain events are positive at both horizons. Leave-one-event-out primary estimates range from approximately $-0.66$ to $-0.34$, so removing any single event does not reverse the sign of the primary estimate.

The setting comparison is exploratory. On each setting's own evaluable support, the median event-level gain is about $+0.53$ for Native ($n = 64$) and $-0.77$ for Controlled ($n = 63$). On the 59 events jointly evaluable in both settings, the Native median is $+0.21$, and Native exceeds Controlled for 38/59 events. These setting-specific summaries use different evaluable supports and are reported separately; no ex-ante responder model is estimated.

Table 19: **Prospective STEP effect decomposition.** Panel A reports aggregate contrasts in Forecast Score. Panel B reports the complete panel-level facet and target decomposition using reference-relative score differences. Positive values favor STEP; negative values favor BASE or SEP in the corresponding contrast. Only the overall STEP–BASE contrast is the pre-specified primary endpoint. Secondary intervals are pointwise 95% percentile bootstrap intervals without multiplicity adjustment.

**Panel A: Aggregate contrasts**

| Contrast | $\Delta$ Forecast Score | 95% CI |
|---|---|---|
| STEP–BASE overall | -0.57 | [-1.40, +0.28] |
| $H_1$ | -0.61 | [-1.50, +0.29] |
| $H_2$ | -0.52 | [-1.46, +0.56] |
| Native | -0.22 | [-1.43, +0.93] |
| Controlled | -0.91 | [-2.44, +0.82] |
| STEP–SEP $H_2$ overall | -0.05 | [-0.45, +0.34] |

**Panel B: Facet and target decomposition**

| Metric | STEP–BASE | 95% CI | STEP–SEP $H_2$ | 95% CI |
|---|---|---|---|---|
| Outcome | -0.11 | [-1.52, +1.32] | -0.76 | [-1.49, -0.11] |
| Distribution | -0.72 | [-1.81, +0.58] | +0.56 | [-0.18, +1.24] |
| Trajectory | -0.87 | [-1.87, +0.06] | +0.06 | [-0.35, +0.43] |
| Margin | -0.50 | [-1.76, +1.01] | -0.35 | [-1.45, +0.60] |
| Total | -0.95 | [-2.58, +0.84] | +1.46 | [+0.51, +2.48] |
| T1 | +0.03 | [-1.51, +1.43] | -0.22 | [-0.72, +0.35] |
| T2 | -1.76 | [-3.05, -0.57] | +0.34 | [-0.32, +0.92] |

**Frozen diagnostic checks.** The forecast diagnostics capture different changes under STEP. R2 uses the primary FIELD evaluation's cell-level calibration definition. Averaged across six models and both settings, categorical accuracy rises from 0.5729 to 0.5778, but ECE also rises from 0.1497 to 0.1578 and mean Brier loss from 0.2540 to 0.2557. Thus, the small accuracy increase is not accompanied by improved calibration.

R3 shows modestly better coherence: the violation rate falls from 0.0562 to 0.0525 and the coherence gap from 0.00245 to 0.00183. R4 uses verified delivered own-arm $H_1 \rightarrow H_2$ lineages with the unchanged reference normalization. Its pooled descriptive mean difference in five-target update gain is near zero ($4.2 \times 10^{-5}$), combining opposite setting directions (Native $+0.00429$, Controlled $-0.00422$). These R3/R4 summaries are descriptive diagnostics, not causal mediation estimates.

For Controlled traces, mean W1 topic pursuit falls from 0.7753 to 0.7444 and declared W2 search requests from 10.69 to 10.52. These are changes in observable behavior, not improvements or failures by definition: neither measure is intrinsically ordered. In particular, the intervention does not increase average declared search volume on this support.

**W3/W4 applicability.** STEP shares the primary evaluation's fixed Gemini semantic judge, prompt, and schema validator. Candidate extraction has a different provenance: STEP uses Gemini only, whereas the primary frame includes Gemini and Luna outputs. STEP contains 10,540 sampled units, with 10,528 technically valid judgments and 12 residual technical failures.

These judgments support descriptive W3/W4 summaries for each arm. BASE, STEP, and SEP were sampled independently, and no paired semantic effect estimator with a fixed aggregation and weighting rule was specified. BASE–STEP and STEP–SEP W3/W4 effects therefore remain NOT ESTIMABLE.

### E.9 RESPONSE PATTERNS AND ILLUSTRATIVE INTERVENTION CASES

The population-level results above average over substantial event-level heterogeneity. We therefore provide two complementary descriptive views. First, we summarize recurrent response patterns on the frozen event-level responder table. Second, we inspect a small set of positive and negative cases to illustrate how the same structured intervention can either improve or degrade forecasts. These

Table 20: **Illustrative positive and negative STEP cases.** Event gains aggregate the frozen event-level STEP–BASE response across models, settings, horizons, and targets. Cases are descriptive examples selected for complete support and cross-model consistency. Registered start times in UTC uniquely identify the corresponding events.

| Event / registered start | Gain | Models + | Response pattern | Illustrative forecast/process change |
|---|---|---|---|---|
| ADO Den Haag–SC Cambuur *Football; 2026-09-19 14:30 UTC* | $+3.98$ | 5/6 | Positive at both horizons and both settings; all five target gains are positive. | Controlled runs show slightly more opened/successfully observed evidence; mean $q_{.50}$ absolute error improves for Margin ($0.33 \rightarrow 0.21$) and Total ($1.04 \rightarrow 0.96$). |
| Cleveland Guardians–Athletics *Baseball; 2026-09-19 22:10 UTC* | $+2.67$ | 6/6 | Positive at both horizons and both settings; all five target gains are positive. | Mean probability on the realized Outcome rises from 0.618 to 0.632; Controlled runs add 1.42 successful observations per run on average. |
| FC Bayern München–1. FC Union Berlin *Football; 2026-09-18 18:30 UTC* | $-5.53$ | 0/6 | Negative at both horizons and both settings; all five target gains are negative. | BASE assigns 0.848 probability to the realized Outcome; STEP lowers it to 0.835, while mean $q_{.50}$ absolute error worsens for Margin ($4.42 \rightarrow 4.88$) and Total ($3.08 \rightarrow 3.42$). |
| Valencia Basket–FC Barcelona *Basketball; 2026-09-19 19:00 UTC* | $-4.66$ | 1/6 | Negative at both horizons and both settings; strongest losses appear in Outcome and T2. | Mean probability on the realized Outcome falls from 0.540 to 0.509 and top-label accuracy from 0.667 to 0.500; Controlled runs open 1.67 fewer sources per run on average. |

cases are selected post hoc for complete support and strong cross-model sign consistency, not for population inference, and they do not define a deployable responder rule.

**Recurring positive responses.** Positive responses recur across models as well as horizons. Of the 28 positive-gain events described above, 20 improve for at least four of the six analyzable models. These events span all three sports: 11 football, 3 basketball, and 6 baseball events. Together with the temporal patterns reported above, this shows that observed gains are not confined to one model, sport, or forecasting cutoff. The pattern remains exploratory and does not identify in advance which events will benefit.

**Negative-response forensics.** The 31 negative-gain events also show recurrent degradation. Twenty-two are negative at both horizons; Controlled response is negative at $H_1$ for 28/31 and at $H_2$ for 27/31. In the pooled event-weighted summary, BASE Outcome top-label accuracy is 0.620 for negative-gain events versus 0.485 for positive-gain events. After STEP, negative-gain events more often change a correct Outcome label to an incorrect one than repair an incorrect label (4.44% vs. 1.88%), whereas positive-gain events show the reverse pattern (1.94% vs. 4.32%). The 50% predictive-interval coverage changes by $-1.68$ and $+2.98$ percentage points, respectively. In Controlled runs, negative-gain events also lose more opened sources ($-0.39$ per run on average) and successful observations ($-0.54$) than positive-gain events ($-0.13$ and $-0.16$). Because these groups are defined by observed STEP–BASE gains, these post-hoc profiles are descriptive. They are compatible with misdirected refinement or scaffold interference, but do not establish a causal mechanism or a prospective rule for identifying which events will benefit from STEP.

**Interpretation.** The positive cases illustrate that the FIELD-derived intervention can improve forecasts across different sports, horizons, and models, sometimes alongside better quantitative placement or richer realized evidence use. The negative cases show why the same structure should not be interpreted as a universal recipe: explicit re-representation and re-projection can perturb an already strong forecast, concentrate probability around a worse center, or interact poorly with an existing forecasting scaffold. The case evidence therefore supports a conservative interpretation of STEP: FIELD-derived insights can be operationally useful, but their application should remain modular, prospectively evaluated, and sensitive to the surrounding forecasting process.