CupArena: Can Large Language Models Beat the Market?
A Live 2026 FIFA World Cup Benchmark

Anonymous authors
Paper under double-blind review

Abstract

Can large language models beat betting markets in live football forecasting, and improve as match outcomes accumulate? We investigate this research question with **CupArenaBench**, a benchmark built from the 2026 FIFA World Cup. Across 103 matches, we require 10 frontier LLMs to submit real-time scoreline and win–draw–loss forecasts before kickoff. Each resolved match provides new evidence for subsequent forecasts. We find that most frontier LLMs outperform conventional forecasting baselines. This pattern persists when market odds are withheld. We then study whether LLM forecasters can improve as new match outcomes become available. Dynamic memory allows agents to retain lessons from resolved matches. Our Forecast Harness enables them to execute forecasting code. As matches resolve, the forecasting code evolves from the newly observed outcomes. This continual adaptation leads to progressively better forecasts as the tournament unfolds. We further introduce **Football-RL**, which updates model parameters from the chronological match stream. The trained policy also transfers zero-shot to the Premier League, improving over the base policy. Together, these results show that LLM forecasters can continually improve from one-shot real-world feedback as new matches resolve. We release the benchmark, data, and code.

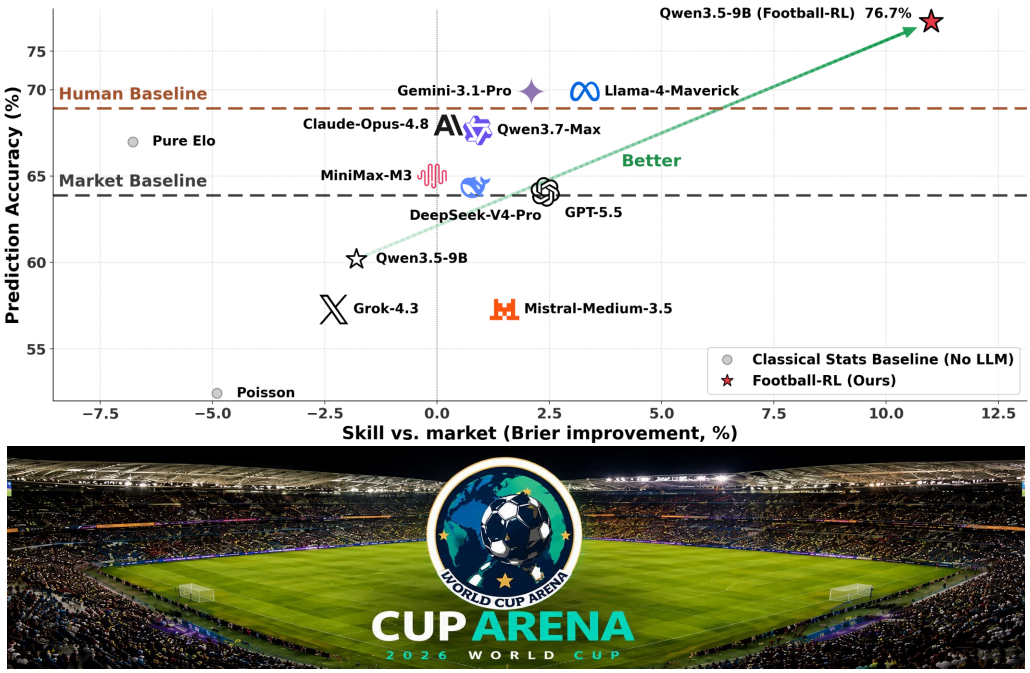


Figure 1: **LLMs show forecasting skill beyond betting markets and aggregated crowd forecasts.** The x-axis measures Brier-score improvement over de-vigged market probabilities, while the y-axis reports win/draw/loss accuracy. Most frontier LLMs outperform the market, and the strongest also surpass aggregated crowd forecasts from millions of FIFA Predictor participants.

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.

1 INTRODUCTION

From the earliest attempts to read the stars, anticipate the seasons, and prepare for war or disease, forecasting has shaped human survival and progress. It remains one of the clearest tests of intelligence: identifying relevant evidence, reasoning under uncertainty, and committing to predictions before outcomes are known. Recent work has introduced temporally controlled benchmarks and shown that LLMs can approach skilled human forecasters (Zou et al., 2022; Halawi et al., 2024; Schoenegger et al., 2024; Karger et al., 2025). Real forecasting adds a simple constraint: an agent must use the evidence available before a decision, make one prediction, and wait for the result. Because results arrive one by one, the next question is whether each result can make later forecasts better. These advances make it possible to ask whether LLMs can produce useful predictive signals in a live competitive environment and improve as outcomes accumulate.

Few domains provide a stronger setting for this question than football. Over more than 160 years, modern football has grown into one of the world's largest sports, giving match prediction substantial competitive, economic, and public significance. Football makes this question especially clear. No single record explains a match, and a forecaster must judge team condition, prior results, and the current match context before assigning probabilities. Outcomes emerge from interacting factors such as team strength, injuries, form, lineups, tactics, and tournament incentives. Betting markets continuously aggregate these signals, and their forecasts have proven difficult for feature-rich statistical models to outperform (Forrest et al., 2005).

The FIFA World Cup is football's most influential tournament, making its live matches a particularly valuable setting for this study. The 2026 tournament provides 103 matches in which LLMs, betting markets, statistical models, and millions of fans forecast the same win, draw, or loss outcomes. Every prediction can be recorded before kickoff, every outcome resolves within hours, and the sequence of matches allows models to learn from accumulated experience. **Can LLMs beat the market on 2026 live World Cup matches without seeing their odds, and can they improve throughout the tournament by learning from experience?**

We therefore build a four stage forecasting agent, S1 to S4, to measure how each stage affects LLM forecasting under the same live match stream. All stages see the same matches and commit forecasts before kickoff.

Overall, our contributions can be summarized as follows:

1. A live World Cup benchmark, **CupArenaBench**, that evaluates 10 frontier LLMs against betting market, statistical, and aggregated crowd forecasts under a protocol that fixes every forecast before kickoff, revealing that a majority of frontier LLMs outperform these forecasts even when market odds are withheld.
2. A self-improvement forecasting algorithm, **Forecast Harness Evolution (FHE)**, that enables the Forecast Harness to improve its forecasting program after every resolved match. FHE searches match evidence, evaluates alternative programs before kickoff, and carries the selected program into the next match.
3. For football forecasting, **Football-RL** is a reinforcement learning algorithm that learns from each match's one new result. Experiments show that Football-RL surpasses the strongest frontier LLM reference in World Cup forecasting by **+6.0%** with a 9B model and improves over its base policy after zero shot transfer to the Premier League by **+20.0%**, showing that match by match learning can transfer across competitions.

2 RELATED WORK

2.1 LARGE LANGUAGE MODELS FOR FORECASTING

LLM forecasting has moved beyond static prompting toward retrieval and temporal reasoning. AUTOCAST++ improves world-event prediction through zero-shot ranking-based retrieval (Yan et al., 2024). Reliability studies evaluate logical consistency through arbitrage (Paleka et al., 2025) and show that date-filtered retrieval can leak future information and inflate retrospective accuracy (El Lahib et al., 2026). Separately, SELF-REFINE uses model-generated feedback to revise outputs without parameter updates (Madaan et al., 2023), while REFLEXION stores verbal reflections in episodic

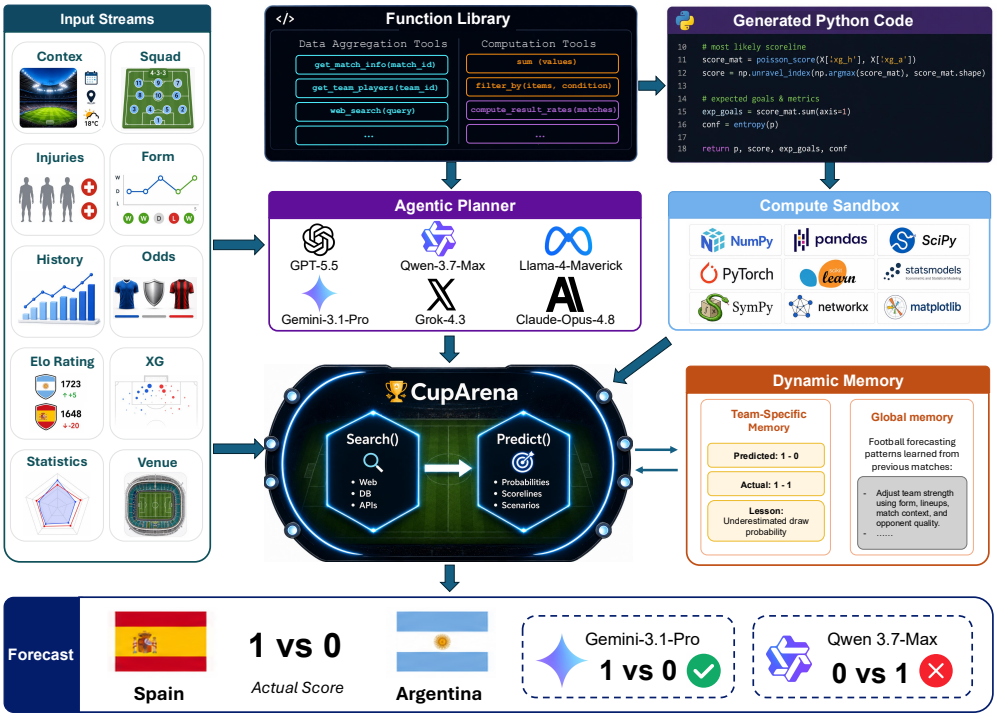


Figure 2: **Overview of CupArena.** The forecasting agent integrates structured match information, tool-based evidence retrieval and computation, executable Python code, and dynamic memory to produce pre kickoff forecasts. After each match, the observed outcome updates team-specific and global memory, enabling progressively informed predictions throughout the tournament.

memory across trials (Shinn et al., 2023). CupArena carries persistent lessons from resolved live events into subsequent forecasts.

2.2 Forecasting Football Matches

Football forecasts commonly draw on statistical models, betting markets, and crowds. Poisson models estimate team-specific attacking and defensive strengths (Maher, 1982), while Elo systems update team ratings after each result (Hvattum & Arntzen, 2010). Betting markets continuously aggregate quantitative evidence and dispersed judgment: bookmaker odds capture information missed by feature-rich models (Forrest et al., 2005), and betting odds and prediction markets outperform professional tipsters (Spann & Skiera, 2009). Aggregated crowd forecasts can outperform individuals but remain biased toward prominent teams (Madsen, 2025). Prior work has not systematically compared LLM football forecasts with statistical, betting market, and aggregated crowd baselines.

2.3 Continual Learning from Sequential Feedback

Continual adaptation studies how language agents improve from feedback that arrives during deployment. Suhr & Artzi (2023) convert real-time binary user feedback into contextual-bandit rewards for continual instruction following. Online iterative RLHF alternates policy-generated data, preference evaluation, and model updates across successive training rounds (Dong et al., 2024). In forecasting, outcome-based reinforcement learning trains language models from noisy and delayed resolutions of future events (Turtel et al., 2025), while outcome-based online RL studies learning when rewards become observable only at the end of a trajectory (Chen et al., 2025). Together, these works motivate sequential learning settings in which sparse outcomes must improve later decisions; CupArena studies this setting in a live tournament where each match occurs once and feedback becomes available only after resolution.

Table 1: Comparison of forecasting tiers.

Tier	Memory	Harness	Market
S1: Direct Prediction	–	–	–
S2: Dynamic Memory	✓	–	–
S3: Forecast Harness Agent	✓	✓	–
S4: Market Aware Agent	✓	✓	✓

3 CUPGYM

CupGym is a sequential forecasting environment built on the standard **Gymnasium** interface (Towers et al., 2025). It connects a chronological stream of World Cup matches with a stateful forecasting loop. Before each match, the environment exposes only information available before kickoff and requires the model to produce a probability distribution over win, draw, and loss. Once the match concludes, **CupGym** reveals the outcome and returns a Brier-based reward. A single environment step therefore covers the full lifecycle of one match, after which the environment advances to the next fixture while preserving the model's accumulated experience. This unified interface supports direct LLM evaluation and post-training methods such as supervised fine-tuning and reinforcement learning.

4 CUPARENABENCH: EVALUATING FRONTIER LLMS

Models. We benchmark 10 LLMs: Gemini-3.1-Pro (Google DeepMind, 2026), Llama-4-Maverick (Meta AI, 2025), Claude-Opus-4.8 (Anthropic, 2026), Qwen3.7-Max (Qwen Team, 2026b), Qwen3.5-9B (Qwen Team, 2026a), GPT-5.5 (OpenAI, 2026), DeepSeek-V4-Pro (DeepSeek-AI, 2026), MiniMax-M3 (MiniMax, 2026), Grok-4.3 (xAI, 2026), and Mistral-Medium-3.5 (Mistral AI, 2026).

Match Suite. **CupArenaBench** covers 103 matches from the 2026 FIFA World Cup. For each match, models predict a probability distribution over win, draw, and loss using only information available before kickoff. We compare these forecasts with de-vigged betting market probabilities, an Elo forecaster (Hvattum & Arntzen, 2010), a Poisson attack-and-defense model (Maher, 1982), and aggregated crowd forecasts collected from FIFA's official Match Predictor (FIFA, 2026).

Protocol. We evaluate all systems under the same live forecasting protocol. Before each match, each system submits a single irrevocable forecast consisting of a predicted scoreline and win, draw and loss probabilities. The scoreline serves as the match prediction, while the probabilities capture predictive confidence and are evaluated using the multiclass Bier score (Brier, 1950). The method follows a simple hierarchy. **S1** provides a static forecasting baseline; **S2** determines *what to remember* from resolved matches; **S3** determines *how to forecast* from the retained evidence via the Forecast Harness and Forecast Harness Evolution; **Football-RL** determines *how to update model parameters* from the chronological match stream; and **S4** augments the Forecast Harness with betting market data. Table 1 summarizes the forecasting tiers.

4.1 SINGLE TURN BENCHMARKS (S1 TO S2)

S1: Direct Prediction. **(S1)** establishes the static forecasting starting point with a fixed pre kickoff information package assembled from API-Football (API-Football, 2026). It covers match context, Elo ratings (Hvattum & Arntzen, 2010), recent form, key players and confirmed lineups, suspensions, tournament xG statistics, and head-to-head records. The complete input fields and data sources are provided in Appendix B. Given this information, the model directly submits one predicted scoreline and its win, draw, and loss probabilities.

S2: Dynamic Memory. **(S2)** answers *what to remember* by adding two persistent memory stores to the S1 forecasting setup. **Team Specific Memory** records experience associated with each team, while **Global Memory** captures patterns that apply across the tournament. Memory is updated only after a match resolves, allowing the agent to compare its predicted scoreline and probabilities with

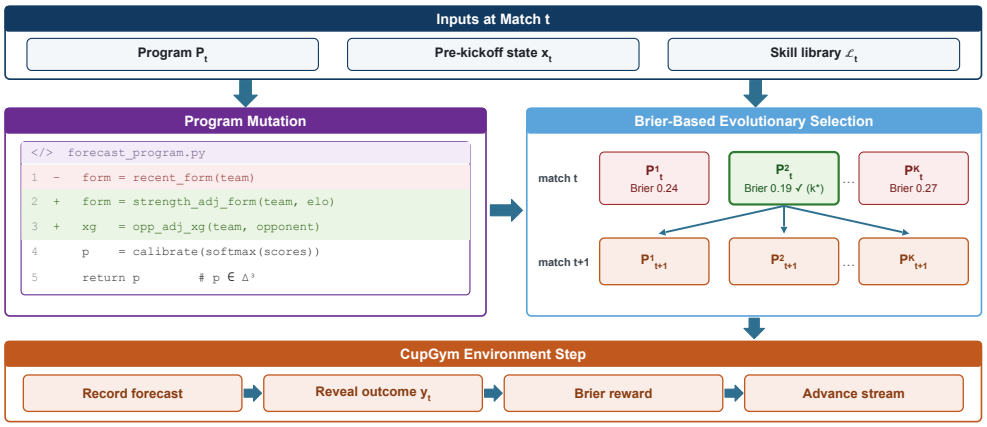


Figure 3: **Forecast Harness Evolution (FHE).** Before kickoff, FHE edits P_t with the time aligned state x_t and current Skill Library $\mathcal{L}_t$ to generate executable variants. The Forecast Harness executes every variant and fixes every forecast before the result is available. After the match resolves, FHE selects the executable program that minimizes the realized probabilistic forecasting error and carries it into the next match. CupGym records the forecast, reveals the completed match outcome, returns the Brier-based reward, and advances the stream.

the actual result and extract lessons for subsequent forecasts. Inspired by probabilistic memory (Liao et al., 2026), we assign each memory a Beta based confidence. Supporting evidence updates $\alpha \leftarrow \alpha + 1$, while contradicting evidence updates $\beta \leftarrow \beta + 1$. Memory confidence is computed as $C = \alpha/(\alpha + \beta)$. Before each prediction, a RAG module retrieves memories from both stores according to their confidence and relevance. All parameters are provided in Appendix C.

4.2 Multi Turn Benchmarks (S3 to S4)

S3: Forecast Harness Agent. **(S3)** answers *how to forecast* from the evidence retained by S2. As shown in Figure 3, S3 first directs the Forecast Harness to search for match evidence and then to issue a forecast from registered function library execution. S3 equips the agent with a search then predict Forecast Harness and **Forecast Harness Evolution (FHE)**. S3 uses screened inputs (Appendix E). FHE is a code evolution algorithm that updates executable forecasting programs from match outcomes. At match t, the current program P_t maps a pre kickoff information state x_t to a three way forecast. P_t uses registered data aggregation and computation tools to retrieve match records, team form, Elo ratings, player statistics, xG, and lineups, then implements numerical transformations and probability calibration. The program specifies concrete functions, their parameters, and the tool call sequence that produces a forecast.

Before kickoff, the Forecast Harness submits the official forecast $\hat{p}_t = \mathrm{Execute}(P_t, x_t)$, where P_t is determined before the current match outcome is available. Candidate forecasts p_t^k are recorded separately for program selection. FHE edits P_t with the time aligned state x_t and the current Skill Library $\mathcal{L}_t$ to generate K executable variants, $\{P_t^k\}_{k=1}^K = \mathrm{Mutate}(P_t, x_t, \mathcal{L}_t)$. FHE edits functions, parameters, or registered tool calls, such as adding opponent strength xG adjustment, replacing raw recent form with strength adjusted form, changing lineup adjustment, or introducing probability calibration. The Forecast Harness executes every variant before kickoff, $p_t^k = \mathrm{Execute}(P_t^k, x_t)$ with $p_t^k \in \Delta^3$, where Δ^3 is the home win, draw, and away win probability simplex. The probability validator and execution record commit every forecast before the result is available.

After the match outcome y_t arrives, FHE selects the executable program that minimizes the realized probabilistic forecasting error. With $k_t^\star$ denoting the selected variant index, the next match program is:

$$k_t^\star = \arg\max_{k \in \{1,\dots,K\}} \left[-\mathrm{Brier}\big(\mathrm{Execute}(P_t^k, x_t), y_t\big) \right], \qquad P_{t+1} \leftarrow P_t^{k_t^\star}. \tag{1}$$

The selected program becomes P_{t+1} for the next match. All reported metrics for match t use the previously submitted official forecast. **CupGym** supplies this chronological transition because one

environment step records the forecast, reveals the completed match outcome, returns the Brier-based reward, and advances the stream.

Skill Library as Program Distillation. FHE uses the Skill Library as a curated store of selected program fragments. FHE extracts atomic edits in the selected program, such as `adjust_xg_by_opponent_strength`, and records selected-match evidence. For a fragment s, $\mathcal{E}_t(s)$ is the resolved-match set whose selected program contains s, $n_t(s) = |\mathcal{E}_t(s)|$, and $\bar{r}_t(s)$ is the mean selection reward. FHE promotes s when $n_t(s) \geq n_{\min}$ and $\bar{r}_t(s) \geq q_{\min}$, yielding $\mathcal{L}_{t+1} = \mathcal{L}_t \cup \{s : n_t(s) \geq n_{\min}, \bar{r}_t(s) \geq q_{\min}\}$. Thus, each reusable skill has a concrete provenance in successful code with repeated match evidence. On later matches, FHE retrieves promoted skills while generating fresh program variants. Appendix A gives the complete algorithm.

S4: Market Aware Agent. **(S4)** extends S3 by providing sports betting odds from William Hill and bet365 (William Hill, n.d.; bet365, 2026), together with prediction market probabilities from Polymarket (Polymarket, 2026). S4 provides these market signals as additional evidence for the Forecast Harness.

4.3 Football-RL: Reinforcement Learning from Match to Match Feedback

Football-RL answers *how to update model parameters* from the chronological match stream. Football forecasting provides a sparse stream in which each match occurs once and supervision arrives after the result is known. We introduce **Football-RL**, a reinforcement learning algorithm that turns every resolved match into an immediate learning signal with **Temporal Advantage Calibration** and **Adaptive Rollout Allocation**.

Football-RL optimizes current match rollouts using Temporal Advantage Calibration, which evaluates each forecast against current alternatives and recent forecasting performance. For token $o_{i,k}$, the policy ratio is $\rho_{i,k}(\theta) = \pi_\theta(o_{i,k} \mid x, o_{i,<k}) / \pi_{\theta_{\text{old}}}(o_{i,k} \mid x, o_{i,<k})$. Football-RL clips this ratio as $\bar{\rho}_{i,k}(\theta) = \text{clip}(\rho_{i,k}(\theta), 1-\epsilon, 1+\epsilon)$ and uses $\ell_{i,k}^{\text{clip}}(\theta) = \min(\rho_{i,k}(\theta)\widehat{A}_{t,m}^{(i)}, \bar{\rho}_{i,k}(\theta)\widehat{A}_{t,m}^{(i)})$. The objective is:

$$\mathcal{J}_{\text{FRL}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{k=1}^{|o_i|}\ell_{i,k}^{\text{clip}}(\theta) - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})\right] \tag{2}$$

Let $\mathcal{R}_{t,m} = \{r_{t,m}^{(i)}\}_{i=1}^{G}$ denote current rollout rewards and $\mathcal{H}_{<t,m}^{\text{rec}}$ the recency-weighted rewards from earlier resolved matchdays. Temporal Advantage Calibration defines the following temporal advantage reference:

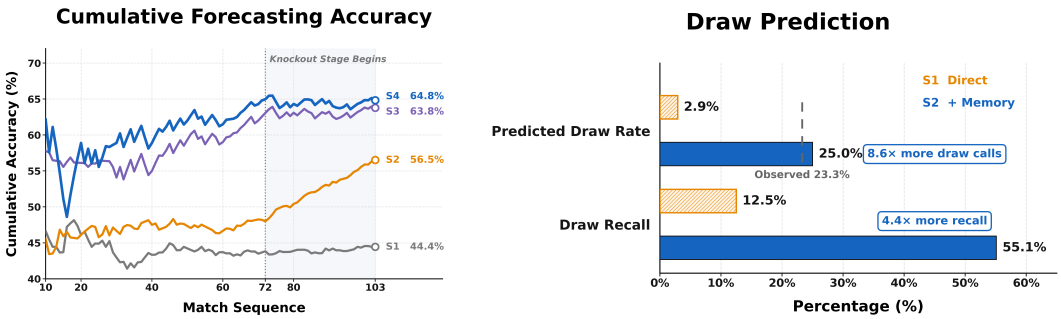
$$\widehat{A}_{t,m}^{(i)} = \frac{r_{t,m}^{(i)} - \mu\big(\mathcal{R}_{t,m} \cup \mathcal{H}_{<t,m}^{\text{rec}}\big)}{\sigma\big(\mathcal{R}_{t,m} \cup \mathcal{H}_{<t,m}^{\text{rec}}\big) + \epsilon}. \tag{3}$$

Historical rewards therefore define an evolving reference, while gradients are computed only from current on-policy rollouts. Adaptive Rollout Allocation sets the exploration budget from the initial rollout dispersion $s_{t,m} = \text{Std}(\mathcal{R}_{t,m})$. When $s_{t,m} < \tau$, Football-RL sets $G \leftarrow G_{\text{high}}$ and $\mathcal{T} \leftarrow \mathcal{T}_{\text{high}}$.

Football-RL retains the underlying GRPO optimization objective while applying Temporal Advantage Calibration and Adaptive Rollout Allocation to chronological football forecasting. Complete reward definitions and implementation details are provided in Appendix D.

5 Experiment

We structure our experimental evaluation around four central research questions: **(RQ1)** ***Forecast Harness:*** *Does the Forecast Harness improve over direct prediction?* **(RQ2)** ***LLMs vs. the Crowd:*** *can LLMs outperform aggregated crowd forecasts?* **(RQ3)** ***LLMs vs. the Market:*** *can LLMs outperform betting market forecasts on real World Cup matches?* **(RQ4)** ***Learning from Sparse Feedback:*** *can Football-RL improve a moderately performing model from sparse and noisy match level feedback, enabling better forecasts on future matches?* To answer these questions, we evaluate 10 frontier LLMs throughout the 2026 FIFA World Cup, analyze the progressive S1 to S4 agent configurations, and train models chronologically with Football-RL.



(a) Forecast Harness Configurations. **(b)** Dynamic Memory and Draw Prediction.

Figure 4: Panel (a) reports cumulative forecasting accuracy for the S1 to S4 agent configurations, while panel (b) reports the effect of dynamic memory on draw prediction.

Experiment Setting: We evaluate all forecasting models on 103 World Cup matches, excluding the third-place playoff because it is conventionally treated as a lower-stakes consolation fixture with limited relevance to the tournament's championship outcome. Predictions are generated with a temperature of 0.3 and a maximum output length of 2,500 tokens, and include a 3-way probability distribution, a 90-minute scoreline, and a short rationale. All forecasting metrics and wagering decisions use the official forecasts submitted before kickoff. For Football-RL, we conduct full-parameter BF16 training on Qwen3.5-9B using 2*H100 GPUs throughout the 39-day tournament. The 103 chronologically ordered match prompts are grouped into 34 matchdays, with 8 on-policy rollouts per match, a learning rate of 1×10^{-6}, a KL coefficient of 0.001, a maximum prompt length of 4,096 tokens, and a maximum response length of 6,144 tokens. As new match outcomes become available, the model weights are updated at the end of each matchday and evaluated on the following matchday under a rolling protocol. Full Football-RL configurations, including Qwen3.5-27B, are reported in Table 5.

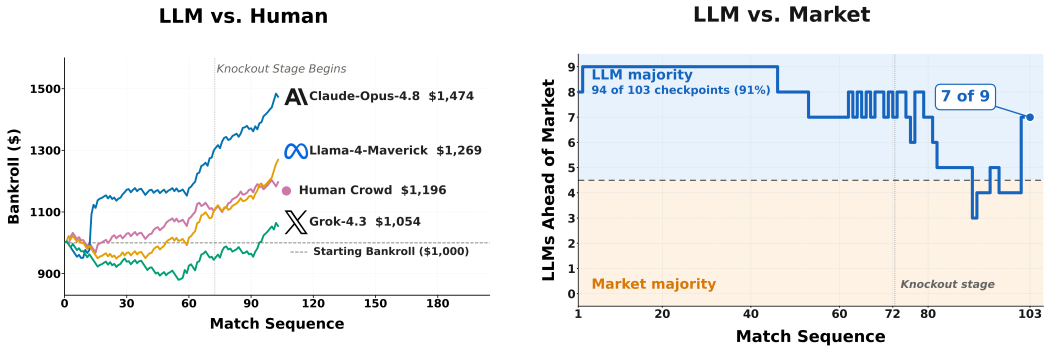
5.1 Comparison of Forecast Harness Configurations

(RQ1) ***Forecast Harness:*** *Does the Forecast Harness improve over direct prediction?* Figure 4(a) traces the cumulative accuracy of the S1 to S4 agent configurations. S1 remains nearly flat, indicating that direct prediction repeatedly applies the same forecasting bias without adapting to newly observed outcomes. Introducing dynamic memory in S2 changes this trajectory: its accuracy rises steadily as additional matches are resolved, showing that the agent progressively learns tournament-specific outcome patterns. S3 produces a substantially stronger performance regime by allowing the Forecast Harness to actively collect evidence, perform explicit computations, and construct a match specific forecasting process. S4 preserves this advantage with market signals, while the S3 Forecast Harness captures the principal predictive signal without market access. Together, these results answer **RQ1**: the Forecast Harness improves forecasting over direct prediction.

To understand why memory produces such a large improvement, we further examine draw prediction in Figure 4(b). S1 predicts draws in only 2.9% of matches despite an observed draw rate of 23.3%, revealing a strong systematic preference for decisive outcomes. Dynamic memory corrects this bias: S2 increases the predicted draw rate to 25.0%, closely matching the tournament distribution, and raises draw recall from 12.5% to 55.1%. The key role of memory is therefore not simply to store previous matches, but to recalibrate the agent's outcome prior from sequential feedback. By learning that draws occur far more frequently than direct prediction assumes, the agent corrects a persistent class-level error and converts accumulated experience into progressively higher forecasting accuracy.

5.2 LLMs vs. Human Crowd and Market

(RQ2) ***LLMs vs. the Crowd:*** *Can LLMs outperform aggregated crowd forecasts?* We conduct a real money wagering experiment on Betfair Exchange (Betfair, 2019) to determine whether the S3 Forecast Harness, operating without access to market data, can outperform aggregated crowd forecasts and translate its independent predictions into sustained capital growth under live market execution. Each model and the aggregated crowd forecast are assigned an initial bankroll of $1,000 and place



(a) LLMs vs. the Human Crowd. **(b) LLMs vs. the Market.**

Figure 5: Panel (a) compares realized betting performance of the S3 Forecast Harness and the aggregated crowd forecast, while panel (b) reports the number of LLMs that outperform the betting market throughout the tournament.

a fixed $10 wager on every match. Bets are settled using the result after 90 minutes plus stoppage time, and a 2% commission is deducted from net winnings in each market (Betfair, n.d.). Figure 5(a) reveals a clear hierarchy in realized betting performance. The S3 Forecast Harness outperforms aggregated crowd forecasts and translates its independent predictions into sustained capital growth under live market execution.

(RQ3) ***LLMs vs. the Market:*** *Can LLMs outperform betting market forecasts on real World Cup matches?* Figure 5(b) shows that a majority of the 9 evaluated LLMs outperform the betting market. At each chronological checkpoint, we count how many LLMs have accumulated lower prediction error than the betting market. The advantage holds at 94 of the 103 checkpoints, and 7 of the 9 LLMs finish the tournament ahead. This majority advantage persists across the live 103 match competition.

5.3 Learning from Sparse Feedback

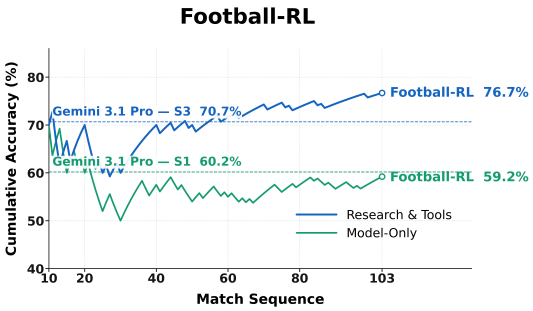


Figure 6: Cumulative forecasting accuracy of Football-RL under the Research & Tools and Model-Only settings. Football-RL progressively improves the Research & Tools model beyond the Gemini-3.1-Pro S3 reference, while the Model-Only setting gradually recovers toward the Gemini-3.1-Pro S1 reference.

(RQ4) ***Learning from Sparse Feedback:*** *Can Football-RL improve a model from sequential match level feedback?* We apply Football-RL to Qwen3.5-9B, updating the model chronologically as each match outcome becomes available. We use Gemini-3.1-Pro as the reference because it achieves the highest forecasting accuracy among the 10 frontier LLMs evaluated in our World Cup benchmark. We compare Qwen3.5-9B under 2 settings: Research & Tools, which provides the S3 Forecast Harness, and Model Only, which relies solely on the model itself.

As shown in Figure 6, Football-RL continuously improves Qwen3.5-9B under the Research & Tools setting, ultimately surpassing the strongest frontier LLM reference by 6.0 percentage points. This result shows that chronological post training can elevate a 9B model beyond the strongest frontier LLM when sparse feedback is paired with grounded research and computational tools. The Football-RL algorithm uses match outcomes as the learning signal, while the Forecast Harness supplies the evidence needed to convert that signal into reliable improvements on future matches. This resolves **RQ4**: Football-RL improves a model from sequential match level feedback.

We compare Football-RL with GRPO and DAPO on Qwen3.5-9B and Qwen3.5-27B. As shown in Table 2, Football-RL delivers the strongest performance at both scales, reaching 76.7% accuracy with Qwen3.5-9B and 79.6% with Qwen3.5-27B. Football-RL converts sparse sequential feedback into sustained gains across both backbones.

Table 2: Comparison of reinforcement learning methods across model scales.

Method	Qwen3.5-9B Acc. ↑	Qwen3.5-27B Acc. ↑
Base Policy	60.2	58.3
GRPO	59.2	56.3
DAPO	62.1	60.2
Football-RL (Ours)	**76.7**	**79.6**

Table 3: Cross-competition generalization on the 2026 Premier League (MW29–MW38).

Method	Qwen3.5-9B Acc. ↑	Qwen3.5-27B Acc. ↑
Base Policy	45.0	43.0
GRPO	41.0	44.0
DAPO	61.0	57.0
Football-RL (Ours)	**63.0**	**65.0**

Generalization to Novel Competitions We evaluate the World Cup trained Football-RL checkpoints on the 2026 English Premier League, using Matchweeks 29 to 38 (100 matches). We directly transfer the trained checkpoints without further RL updates, allowing us to test whether the policy learned from World Cup feedback generalizes to a different competition rather than merely adapting online to the new data. This interval begins after the release of the Qwen3.5 checkpoints, reducing the risk that its match outcomes were already contained in the base models' training data. Since these EPL matches had already concluded when we conducted the evaluation, however, Web search could retrieve post match information unavailable at prediction time. We therefore route every retrieved result through an independent **Supervisor Agent**, which removes final scores, outcomes, post match reports, subsequent standings, and other future information before the evidence reaches the forecasting agent, and manually inspect all retained evidence.

As shown in Table 3, Football-RL maintains its advantage when transferred from the World Cup to the Premier League, demonstrating robust cross-competition generalization.

6 Conclusion

We present CupArena, a live benchmark comparing frontier LLMs with betting markets, statistical models, and crowd forecasts across 103 World Cup matches. Results show that LLMs outperform established signals, while the S1 to S4 study shows that dynamic memory and the Forecast Harness improve forecasts by correcting systematic biases and constructing evidence grounded prediction strategies. Football-RL learns from chronological match streams and enables the open source Qwen3.5-9B model to surpass the strongest frontier LLM reference with research and computational tools. Together, these findings establish football forecasting as a real world test of sequential adaptation and show that LLM forecasts improve as outcomes arrive.

7 Limitations

Football-RL learns from the small stream of resolved matches available in a live competition, where each fixture provides one outcome and cannot be replayed. A single tournament therefore offers limited direct evidence for parameter updates. This constraint reflects the intended setting rather than an artificial benchmark restriction: real football competitions also reveal results one match at a time. We evaluate Football-RL chronologically during the World Cup, at two model scales, and after transfer to the Premier League. Observations from future competitions can extend this evidence while preserving the same live forecasting protocol.

AI USE STATEMENT

We used GPT during manuscript preparation to improve the writing and help identify LaTeX issues. We checked all AI-supported changes ourselves and are fully responsible for the submitted version.

ETHICS STATEMENT

This work evaluates forecasting systems using publicly available football data and does not involve human subjects or private personal information. Real-money wagering was used only as a controlled research evaluation with fixed stakes; our study does not advocate gambling or the deployment of these systems for financial decision-making.

REPRODUCIBILITY STATEMENT

We provide implementation and evaluation details throughout the paper and appendix. The live forecasting protocol and S1–S4 configurations are described in Section 4, with the experimental setup and chronological evaluation protocol reported in Section 5. Forecast Harness Evolution is specified in Section 4.2 and Appendix A, including its program mutation, selection, and Skill Library update rules. Appendix B details the pre-kickoff input features and time-alignment protocol, while Appendix C specifies the Dynamic Memory stores, confidence updates, and retrieval settings. Football-RL is defined in Section 4.3, with its reward design, Temporal Advantage Calibration, Adaptive Rollout Allocation, and full training configuration provided in Appendix D. We will release the benchmark data, source code, configurations, and evaluation scripts to support reproduction of our results.

REFERENCES

Anthropic. Introducing claude opus 4.8, May 2026. URL https://www.anthropic.com/news/claude-opus-4-8.

API-Football. API-Football Documentation, 2026. URL https://www.api-football.com/documentation-v3. Accessed: 2026-08-01.

bet365. Terms and conditions, 2026. URL https://help.bet365.com/s/en/terms-and-conditions. Effective: 2026-08-25. Accessed: 2026-09-20.

Betfair. What is the betfair exchange? https://betting.betfair.com/how-to-use-betfair-exchange/beginner-guides/what-is-the-betfair-exchange-010819-51.html, 2019. Accessed: 2026-08-03.

Betfair. Betfair charges. https://support.betfair.com/app/answers/detail/betfair-charges, n.d. Accessed: 2026-08-03.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078⟨0001:VOFEIT⟩2.0.CO;2.

Fan Chen, Zeyu Jia, Alexander Rakhlin, and Tengyang Xie. Outcome-based online reinforcement learning: Algorithms and fundamental limits. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/987c89953052812d9e4e295418d082b1-Abstract-Conference.html.

DeepSeek-AI. Deepseek v4 preview release, April 2026. URL https://api-docs.deepseek.com/news/news260424/.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. RLHF workflow: From reward modeling to online RLHF: A comprehensive practical alignment recipe of iterative preference learning. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a13aYUU9eU.

Ali El Lahib, Ying-Jieh Xia, Zehan Li, Yuxuan Wang, and Xinyu Pi. Temporal leakage in search-engine date-filtered web retrieval: A retrospective forecasting case study. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 787–795, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-short.65. URL https://aclanthology.org/2026.acl-short.65/.

FIFA. Fifa world cup 2026 match predictor, 2026. URL https://play.fifa.com/match-predictor/.

David Forrest, John Goddard, and Robert Simmons. Odds-setters as forecasters: The case of english football. *International Journal of Forecasting*, 21(3):551–564, 2005. doi: 10.1016/j.ijforecast.2005.03.003.

Google DeepMind. Gemini 3.1 pro: A smarter model for your most complex tasks, February 2026. URL https://deepmind.google/blog/gemini-3-1-pro-a-smarter-model-for-your-most-complex-tasks/.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models. *Advances in Neural Information Processing Systems*, 37: 50426–50468, 2024.

Lars Magnus Hvattum and Halvard Arntzen. Using ELO ratings for match result prediction in association football. *International Journal of Forecasting*, 26(3):460–470, 2010. doi: 10.1016/j.ijforecast.2009.10.002.

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.

Junfeng Liao, Qizhou Wang, Jianing Zhu, Bo Du, Rui Yan, and Xiuying Chen. Belief memory: Agent memory under partial observability. *arXiv preprint arXiv:2605.05583*, 2026. doi: 10.48550/arXiv.2605.05583. URL https://arxiv.org/abs/2605.05583.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594, 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Jens Koed Madsen. Goal-line oracles: Exploring accuracy of wisdom of the crowd for football predictions. *PLOS ONE*, 20(1):e0312487, 2025. doi: 10.1371/journal.pone.0312487.

M. J. Maher. Modelling association football scores. *Statistica Neerlandica*, 36(3):109–118, 1982. doi: 10.1111/j.1467-9574.1982.tb00782.x.

Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation, April 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/.

MiniMax. Minimax m3: Frontier coding, 1m context, native multimodality—all in one model, June 2026. URL https://www.minimax.io/blog/minimax-m3.

Mistral AI. Mistral medium 3.5, 2026. URL https://docs.mistral.ai/models/model-cards/mistral-medium-3-5-26-04.

OpenAI. Introducing gpt-5.5, April 2026. URL https://openai.com/index/introducing-gpt-5-5/.

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

Polymarket. What is a prediction market?, 2026. URL https://help.polymarket.com/en/articles/13364272-what-is-a-prediction-market. Published: 2026-07-01. Accessed: 2026-09-20.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.7: The agent frontier, May 2026b. URL https://qwen.ai/blog?id=qwen3.7.

Philipp Schoenegger, Indre Tuminauskaite, Peter S Park, Rafael Valdece Sousa Bastos, and Philip E Tetlock. Wisdom of the silicon crowd: Llm ensemble prediction capabilities rival human crowd accuracy. *Science Advances*, 10(45):eadp1528, 2024.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652, 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Martin Spann and Bernd Skiera. Sports forecasting: A comparison of the forecast accuracy of prediction markets, betting odds and tipsters. *Journal of Forecasting*, 28(1):55–72, 2009. doi: 10.1002/for.1091.

Alane Suhr and Yoav Artzi. Continual learning for instruction following from realtime feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1404. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/666cccc6376058e251315b4de7e085b9-Abstract-Conference.html.

Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U. Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, Rodrigo Perez-Vicente, Andrea Pierré, Sander Schulhoff, Jun Jet Tai, Hannah Tan, and Omar G. Younis. Gymnasium: A standard interface for reinforcement learning environments. In *NeurIPS 2025 Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=qPMLvJxtPK.

Benjamin Turtel, Danny Franklin, Kris Skotheim, Luke Hewitt, and Philipp Schoenegger. Outcome-based reinforcement learning to predict the future. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=bbhdeL8EUX.

William Hill. General sports betting rules, n.d. URL https://help.williamhill.com/hc/en-gb/articles/21754578862493-General-Sports-Betting-Rules. Accessed: 2026-09-20.

xAI. Grok 4.3, 2026. URL https://docs.x.ai/developers/models/grok-4.3.

Qi Yan, Raihan Seraj, Jiawei He, Lili Meng, and Tristan Sylvain. AutoCast++: Enhancing world event prediction with zero-shot ranking-based context retrieval. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=COYDmKkQH4.

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*, volume 35, 2022.

APPENDIX

APPENDIX OVERVIEW

SECTION	
§A	**Forecast Harness Evolution Algorithm:** We specify the chronological FHE procedure, its notation, program selection rule, and Skill Library updates.
§B	**Pre Match Input Features:** We list the pre kickoff evidence groups, time-alignment protocol, and normalized fixture record supplied to the Forecast Harness.
§C	**Dynamic Memory Details:** We define the Team Specific Memory and Global Memory stores, their confidence updates, and retrieval settings.
§D	**Football-RL Details:** We provide the reward definition, Temporal Advantage Calibration, Adaptive Rollout Allocation, and training configuration.
§E	**S3 Search Information Screening:** We describe supervisor agent filtering and human screening of market information.
§F	**Statistical Uncertainty:** We report paired comparisons and repeated-run summaries.

A FORECAST HARNESS EVOLUTION ALGORITHM

Notation. At match t, P_t is the current executable forecasting program, x_t is the pre kickoff information state, and $\mathcal{L}_t$ is the Skill Library. The official forecast $\hat{p}_t = \mathrm{Execute}(P_t, x_t)$ is submitted before kickoff and used for match scoring. FHE generates K variants $\{P_t^k\}_{k=1}^K$ and records forecasts p_t^k. After the outcome y_t arrives, r_t^k denotes the negative Brier score and $k_t^\star$ denotes the selected variant. $\mathcal{S}_t$ is the selected program's fragment set. For a fragment s, $\mathcal{E}_t(s)$ stores selected match evidence and $\bar{r}_t(s)$ is its mean reward. $n_{\min}$ is the minimum evidence count and $q_{\min}$ is the reward threshold.

FHE uses four program variants, a minimum evidence count of three matches, and a promotion threshold of -0.50.

Algorithm: Forecast Harness Evolution (FHE)
Input: P_1, $\mathcal{L}_1$, $\{x_t\}_{t=1}^T$, K, $n_{\min}$, $q_{\min}$

for $t = 1, \ldots, T$ **do**
  $\hat{p}_t \leftarrow \mathrm{Execute}(P_t, x_t)$
  Submit and record $\hat{p}_t$ before kickoff
  $\{P_t^k\}_{k=1}^K \leftarrow \mathrm{Mutate}(P_t, x_t, \mathcal{L}_t)$
  for $k = 1, \ldots, K$ **do**
    $p_t^k \leftarrow \mathrm{Execute}(P_t^k, x_t)$
  end for
  observe y_t
  Record official forecasting loss $\mathrm{Brier}(\hat{p}_t, y_t)$
  for $k = 1, \ldots, K$ **do**
    $r_t^k \leftarrow -\mathrm{Brier}(p_t^k, y_t)$
  end for
  $k_t^\star \leftarrow \arg\max_k r_t^k, \quad P_{t+1} \leftarrow P_t^{k_t^\star}$
  $\mathcal{S}_t \leftarrow \mathrm{Fragments}(P_t^{k_t^\star})$
  for $s \in \mathcal{S}_t$ **do**
    $\mathcal{E}_t(s) \leftarrow \mathcal{E}_{t-1}(s) \cup \{(t, r_t^{k_t^\star})\}$
  end for
  $\mathcal{L}_{t+1} \leftarrow \mathcal{L}_t \cup \{s : |\mathcal{E}_t(s)| \geq n_{\min}, \bar{r}_t(s) \geq q_{\min}\}$
end for

Output: P_{T+1}, $\mathcal{L}_{T+1}$

B PRE MATCH INPUT FEATURES

Every S1 to S4 forecast is conditioned on a snapshot assembled before kickoff. The snapshot excludes the target match's final score, post match reports, and subsequent standings. API-Football provides

fixture, historical match, player, lineup, and odds records (API-Football, 2026), which are normalized into a common team fixture record. Table 4 lists the input fields used by the forecasting tiers.

Table 4: Pre kickoff input fields.

Group	Fields available before kickoff
Fixture context	Fixture identifier, home and away teams, UTC kickoff, venue, city, and competition round.
Team form and strength	Elo rating, match date, competition, home or away indicator, opponent, goals for, goals against, and win, draw, or loss.
Squad and lineup	Squad number and position, player club, appearances, starts, minutes, rating, goals, assists, accumulated-card suspensions, and confirmed starting XI with formation, position, and grid location.
Performance statistics	Team expected goals, goals prevented, total and on-target shots, possession, passing, corners, fouls, yellow cards, red cards, and match events.
Head-to-head history	Meeting date, competition, teams, and score.
Market signal (S4)	Home-win, draw, and away-win decimal odds, together with bookmaker count.

For each field, the data pipeline records its retrieval timestamp and retains only records whose timestamp precedes kickoff. Missing fields are represented explicitly rather than filled with information from after the match. The Forecast Harness receives this normalized record and may call only registered retrieval and computation functions on it.

C Dynamic Memory Details

S2 maintains two persistent stores. **Team Specific Memory** indexes a record by team and stores observations that should transfer to that team's later fixtures. **Global Memory** stores tournament level observations that can apply to any fixture. Each memory record contains the match identifier and timestamp, the pre kickoff evidence, the forecast and resolved outcome, a concise lesson, and Beta confidence counts (α, β). A supporting resolved match increments α, while a contradicting one increments β. Memory confidence is $C = \alpha/(\alpha + \beta)$.

After each result, the agent first writes candidate records from the forecast outcome comparison, then deduplicates semantically equivalent records within the relevant store. Before the next forecast, retrieval ranks records by fixture relevance, team match, recency, and confidence C. The retrieved records are passed to the Forecast Harness as evidence, not as post hoc labels. Each store retains at most 12 records per team and 36 global records, applies a per matchday recency decay of 0.92, retrieves three team records and two global records, and initializes Beta confidence with $(\alpha, \beta) = (1, 1)$.

D Football-RL Details

Football-RL is a reinforcement learning algorithm that trains on the chronological match stream. A training unit contains only a pre kickoff prompt and on policy forecast rollouts. Once the corresponding match resolves, the outcome based reward is assigned to those rollouts and the algorithm updates the policy before evaluating future fixtures. The reward parser checks that a rollout supplies a valid 90 minute scoreline and a normalized win–draw–loss distribution. The reward assigns weights of 0.70 to normalized Brier reward, 0.20 to outcome accuracy, and 0.10 to output validity.

Temporal Advantage Calibration. For a resolved matchday, Football-RL computes the advantage reference from current rollout rewards together with a recency-weighted pool of earlier resolved rewards. It normalizes each current rollout reward using this reference while retaining gradients only through the current on-policy rollouts. The reference uses the latest 64 resolved rollouts, a per matchday decay of 0.92, and a numerical stabilizer of 10^{-6}.

Adaptive Rollout Allocation. Football-RL begins each matchday with a base rollout group. It measures the initial reward dispersion and, when the dispersion falls below the allocation threshold, increases the rollout budget and the rollout length for that matchday. The base and expanded rollout

Table 5: Football-RL configuration fields.

Configuration field	Qwen3.5-9B	Qwen3.5-27B
Base model and precision	Qwen3.5-9B, BF16	Qwen3.5-27B, BF16
Hardware	2*H100 (80GB)	8*H100 (80GB)
Learning rate and KL coefficient	1×10^{-6}, 0.001	3×10^{-7}, 0.001
Matchdays and rollouts per match	34, 8	34, 8
Update unit and update frequency	Matchday, after each matchday	Matchday, after each matchday
Maximum prompt and response lengths	4,096, 6,144	4,096, 6,144
Optimizer	AdamW, $(0.9, 0.95)$, 0.1	AdamW, $(0.9, 0.95)$, 0.1
Base rollout group, expanded rollout group, and allocation threshold	8, 16, 0.07	8, 16, 0.07
Base and expanded rollout lengths	4,096, 6,144	4,096, 6,144
Historical reward pool size and recency decay	64, 0.92	64, 0.92

groups are 8 and 16, the allocation threshold is 0.07, and the response lengths are 4,096 and 6,144 tokens.

E S3 Search Information Screening

S3 uses web search to retrieve football evidence while excluding market information from the forecasting context. We apply supervisor agent filtering and human screening to the information supplied to S3. A supervisor agent filters market information from search results before the Forecast Harness receives the retrieved material. Human reviewers screen all information supplied to S3 to ensure that the forecasting context contains no market information.

Screening Scope. Market information includes betting odds, bookmaker prices, market-implied probabilities, prediction market quotes, and statements that identify a favorite by reference to betting markets. Screening covers market information embedded in football news as well as explicit betting content. Football evidence such as injuries, suspensions, lineups, team form, and tactical reports remains available after screening. S3 receives the screened material rather than the raw search results.

Information Access. The S3 protocol excludes both structured market inputs and market information retrieved through web search. Human screening applies to all information supplied to S3, including the pre kickoff information package and retrieved material. S4 provides market signals as additional inputs, as described in the main text. The S3 comparisons therefore evaluate forecasts generated from screened football evidence without market data in the forecasting context.

F Statistical Uncertainty

Paired Accuracy Comparisons. Table 6 reports accuracy differences in percentage points, 95% confidence intervals, and paired-test p values. Positive differences favor S3 or Football-RL. The intervals describe differences between forecasts on the same fixtures. The comparison against Gemini-3.1-Pro reports an accuracy advantage of 6.0 percentage points.

Table 6: Paired accuracy summaries. Differences and confidence intervals are in percentage points.

Comparison	Difference	95% CI	p
S3 vs. S2	+7.3	[+2.2, +12.4]	0.008
World Cup: Football-RL vs. Base Policy, 9B	+16.5	[+9.3, +23.7]	< 0.001
World Cup: Football-RL vs. DAPO, 9B	+14.6	[+7.7, +21.5]	< 0.001
World Cup: Football-RL vs. Gemini-3.1-Pro	+6.0	[+1.4, +10.6]	0.031
Premier League: Football-RL vs. Base Policy, 9B	+18.0	[+10.4, +25.6]	< 0.001
Premier League: Football-RL vs. DAPO, 9B	+2.0	[-0.8, +4.8]	0.500

Market Comparison. The S3 comparison on 72 shared fixtures reports a mean Brier advantage of 0.018, with a 95% confidence interval of [0.006, 0.030]. The advantage is the market Brier score minus the S3 Brier score, so positive values favor S3. Comparisons across multiple LLMs require an explicit multiplicity correction before family-level significance claims.

Repeated Runs. Table 7 reports three-run summaries for Qwen3.5-9B on the World Cup. Variability across training runs and sampling uncertainty across fixtures describe distinct sources of uncertainty.

Table 7: Three-run World Cup summaries for Qwen3.5-9B. Values are accuracy percentages. SD denotes the sample standard deviation across runs.

Method	**Run 1, Run 2, Run 3**	**Mean $\pm$ SD**
Base Policy	59.2, 60.2, 61.2	60.2 ± 1.0
Football-RL	74.8, 76.7, 78.6	76.7 ± 1.9

Chronological Evaluation. Cumulative checkpoint counts describe the forecasting trajectory. Statistical comparisons use paired fixture outcomes rather than treating overlapping cumulative checkpoints as independent observations. Dependence within matchdays and variation across complete forecasting runs are distinct from fixture-level sampling uncertainty.