LLM-SoccerArena: Benchmarking LLMs on real-world predictions in sports

Jonas Schröder[1,2,†] Jonas Schweisthal[1,2,†] Oliver Müller[3] Markus Weinmann[4] Stefan Feuerriegel[1,2]

[1]LMU Munich, Munich, Germany [2]Munich Center for Machine Learning (MCML), Munich, Germany [3]Paderborn University, Paderborn, Germany [4]University of Cologne, Cologne, Germany

[†]Joint first authors

jonas.schroeder@lmu.de jonas.schweisthal@lmu.de oliver.mueller@uni-paderborn.de weinmann@wiso.uni-koeln.de feuerriegel@lmu.de

Preprint, June 11, 2026

Abstract

Large language models (LLMs) are increasingly used for reasoning and decision support, yet their ability to forecast real-world future events remains difficult to evaluate. Static benchmarks often fail to capture deployment conditions in which predictions must be made before outcomes are known (i.e., under uncertainty). We introduce **LLM-SoccerArena** (https://llm-soccerarena.com), a reproducible, real-time benchmark for evaluating LLM forecasts of the FIFA World Cup 2026. Our platform records timestamped predictions before matches are played, validates structured model outputs, and evaluates forecasts once official results become available. We provide an extensive benchmark that systematically varies four dimensions: (1) model family, comparing frontier and open LLMs (e.g., Claude Fable 5, Claude Opus 4.8, GPT-5.5, Gemini 3.1 Pro, etc.); (2) information access (i.e., with vs. without web search); (3) prompt strategy (i.e., direct prompting to predict a score vs. prompting for a probability); and (4) forecast horizon (i.e., before the full stage vs. 24-hour-ahead vs. 2-hour-ahead forecasts). For each match, we predict 90-minute outcomes probabilities, full-match probabilities. We also predict structured outcomes of the tournament (e.g., overall winner, team with top scorer). We benchmark models using multi-dimensional performance metrics (e.g., Brier score, log loss, exact-score accuracy, and structured tipping-game points), as well as diagnostic measures of reasoning summaries, output validity, tool use, and internal coherence. Overall, **LLM-SoccerArena** provides a challenging, public benchmarking platform for studying the reliability of LLM-based forecasting in a real-world prediction task.

1 Introduction

LLMs are increasingly deployed in settings that involve uncertain future events, including analysis, planning, decision support, and forecasting. However, evaluating future-event forecasting remains difficult. Static benchmarks may suffer from data contamination, may not reflect deployment-time information access, and often evaluate predictions after the target outcome is already known. Live forecasting benchmarks reduce these problems because predictions are submitted before the relevant outcomes exist [Meyer et al., 2025, Karger et al., 2024].

Football forecasting is a useful domain for such a benchmark. Match outcomes are objective, kickoff times are known, and the prediction task is naturally structured. Unlike many generic future-event questions, football also permits domain-specific evaluation: not only whether a model assigns probability to the correct result, but also whether it predicts plausible scorelines, goal differences, total goals, and knockout advancement outcomes. The FIFA World Cup 2026 is especially attractive because it is a globally visible tournament with a fixed competition structure, rich public information, and substantial public interest.

We introduce LLM-SoccerArena, a live platform and benchmark for LLM forecasting during the FIFA World Cup 2026. The project has two connected goals. First, it provides a public-facing website (https://llm-soccerarena.com) where users can inspect and compare LLM predictions, model confidence, exact-score forecasts, and live performance metrics. Second, it implements a benchmark protocol, fixed in advance, that evaluates LLM systems under controlled information-access and prompting conditions. Public users may interact with the website, but public user participation is not part of the core paper benchmark.

The system is fully implemented and deployed, and has already collected timestamped pre-match predictions. As of this writing, 8 frontier LLMs have produced 2,368 match forecasts and 480 tournament-level forecasts under the controlled design, before any World Cup 2026 match has been played. We therefore report the platform, the benchmark protocol, the released dataset, and *preliminary, pre-outcome diagnostics* (output validity, tool use, cost, and internal coherence). Forecast-accuracy results are deliberately deferred: they will be computed automatically once official results become available, following the fixed evaluation plan in Section 11.

The core contribution is a live, timestamped evaluation protocol. For each selected model and match, the system generates predictions under a 2×2 design: closed-book versus open-book access, crossed with direct-score versus probabilistic-forecast prompting. The platform stores raw prompts, raw model responses, parsed outputs, validation status, tool-use metadata, timing metadata, and later evaluation metrics. This enables both public display and reproducible offline analysis.

The main contributions are:

1. a live, deployed platform for LLM World Cup forecasting, with a public website for inspecting and comparing model forecasts and leaderboards across multiple metrics;

2. a 2×2 benchmark design, fixed in advance, that crosses information access with prompt strategy across three forecast horizons;

3. a soccer-specific evaluation framework covering probabilities, categorical results, exact scores, goal differences, total goals, and knockout advancement, plus tournament-level structured questions (champion, semifinalists, group winners, top-scorer nation);

4. a structured logging and validation protocol that yields a reproducible, timestamped prediction archive with full prompt/response provenance;

5. preliminary live-benchmark diagnostics from the already-collected predictions, and a fixed post-tournament evaluation plan.

2 Related Work

This work relates to three lines of research. First, *dynamic and live forecasting benchmarks* evaluate systems on events whose outcomes are unknown at prediction time, reducing contamination and creating a deployment-like evaluation setting [Meyer et al., 2025, Karger et al., 2024]. Second,

LLM forecasting studies whether language models can produce calibrated probabilities or useful forecasts for real-world future events [Halawi et al., 2024]. Third, *sports prediction and football analytics* provide a long history of statistical and machine-learning approaches to match-outcome and scoreline prediction: goal-based Poisson and bivariate-Poisson models [Maher, 1982, Dixon and Coles, 1997, Karlis and Ntzoufras, 2003, Koopman and Lit, 2015], rating-based approaches [Hvattum and Arntzen, 2010, Ley et al., 2019], machine-learning and tournament models [Groll et al., 2015, Schauberger and Groll, 2018, Groll et al., 2019], and bookmaker-consensus models [Leitner et al., 2010, Zeileis et al., 2018]. Forecast quality is assessed with proper scoring rules and calibration diagnostics [Epstein, 1969, Gneiting and Raftery, 2007, Gneiting et al., 2007, Constantinou and Fenton, 2012]. Our contribution differs from generic forecasting benchmarks by focusing on a homogeneous, high-interest sports domain with rich soccer-specific metrics, and differs from traditional sports-prediction work by evaluating general-purpose LLM systems under controlled closed-book and open-book prompting conditions. Following best practice in transparent confirmatory research [Nosek et al., 2018], our analysis plan and metrics are fixed before outcomes are observed.

3 The LLM-SoccerArena Platform

LLM-SoccerArena is both a public website and a benchmark infrastructure built on the same data. The website (https://llm-soccerarena.com) presents live LLM predictions during the FIFA World Cup 2026 and lets visitors compare model forecasts across matches, models, information-access settings, prompt strategies, and forecast horizons. The scientific benchmark uses the same underlying prediction records but focuses on controlled LLM comparisons rather than public-user participation.

3.1 Architecture and data flow

The system is a TypeScript monorepo with three layers: a data/inference layer (scheduled jobs for fixture ingestion, model calls, validation, evaluation, and export), a storage layer (a single SQLite database holding fixtures, predictions, and evaluations), and a web layer (a Next.js application serving the public site). Figure 1 summarizes the end-to-end flow.

The platform supports: match-level prediction displays; interactive model comparisons; live leaderboards across multiple metrics; filtering by model, access condition, prompt strategy, horizon, and tournament stage; transparent reporting of invalid outputs, repairs, normalization, and open-book search use; and export of prediction and evaluation data for research analysis.

3.2 Public website

The public site is bilingual (English and German) and includes a home dashboard, a match schedule with per-match prediction cards, an interactive analytics page, a tournament bracket view, and a model inspector. The analytics page exposes the full multi-metric leaderboard (Section 9) with live filtering by horizon, access condition, prompt strategy, stage, model, and provider, and renders cumulative metric trajectories over the match calendar. Figure 2 shows representative views.

The website may include engagement features for public users, but public users are not part of the core paper evaluation.

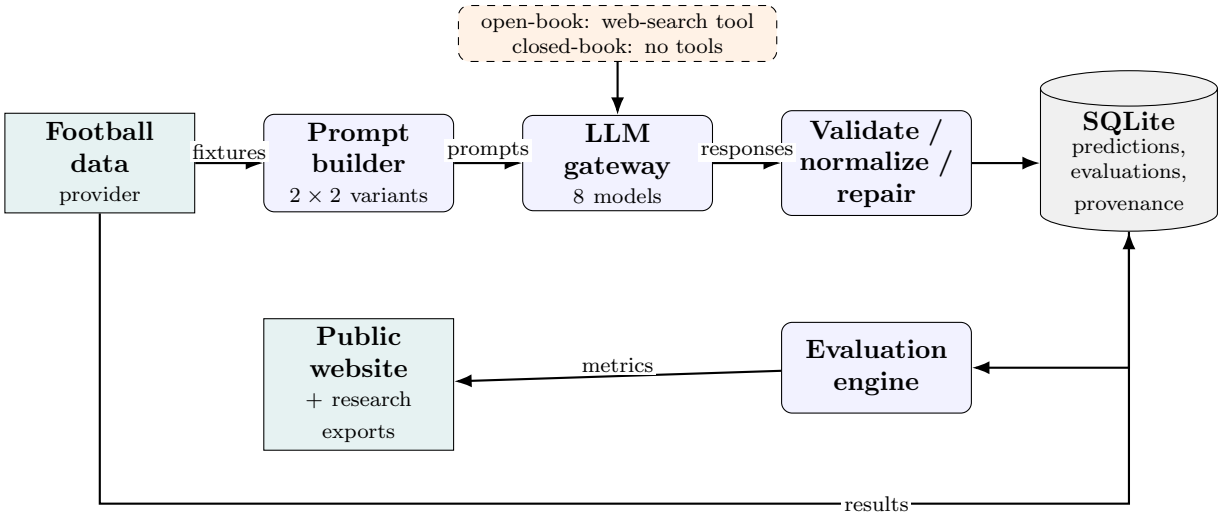


Figure 1: End-to-end data flow. Fixtures and results are ingested from a football data provider; for each selected match the prompt builder emits the four 2×2 prompt variants; closed-book calls run without tools while open-book calls enable provider web search; responses are validated, normalized, and (if needed) repaired before storage; once results are known an evaluation engine computes metrics; the website and research exports read the same database.

4 Benchmark Design

4.1 Experimental unit

The unit of prediction is

$$\text{model} \times \text{match} \times \text{forecast horizon} \times \text{access condition} \times \text{prompt strategy} \times \text{sample id}.$$

Each unit receives one deterministic model call with

$$\text{temperature} = 0, \quad \text{top\_p} = 1, \quad n = 1,$$

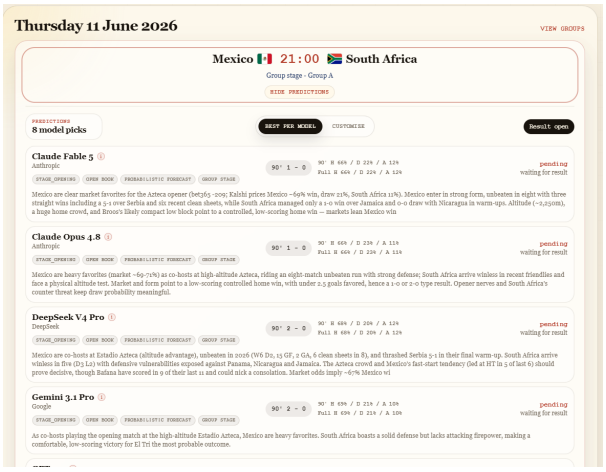
and a fixed completion-token budget shared across first attempts, validation retries, and repair calls. No majority vote, repeated sampling, or stochastic ensembling is used in the core analysis.

4.2 Models

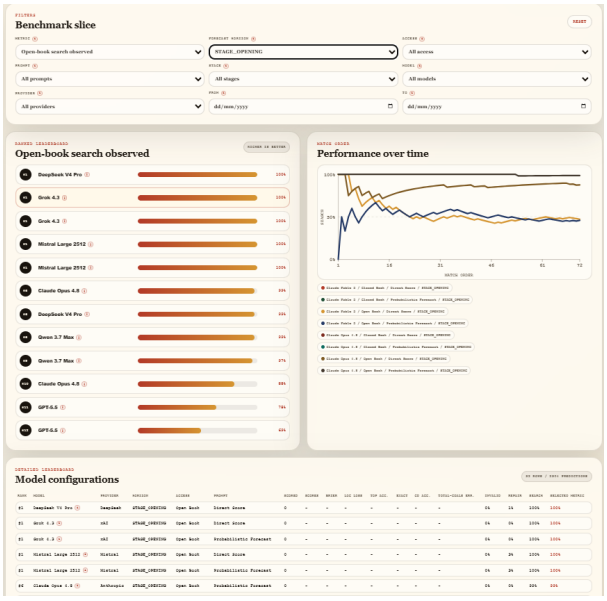
The active roster comprises 8 current frontier models, one per major provider (Table 1). All are called through a single provider gateway (OpenRouter) so that closed-book and open-book conditions, generation settings, and logging are identical across models. An extended roster (additional efficiency-tier and open-weight models) is implemented in the model registry and can be activated for a secondary leaderboard; the controlled 2×2 analysis uses the 8 flagship models for which both access modes work reliably.

4.3 The 2×2 design

The core benchmark crosses two factors: information access and prompt strategy (Table 2). For each selected model, match, and forecast horizon, the system generates four predictions: (1) closed-book + direct-score; (2) closed-book + probabilistic-forecast; (3) open-book + direct-score; (4)



(a) Per-match model picks, with conditions, 90-minute probabilities, and reasoning.



(b) Multi-metric leaderboard with live filters by horizon, access, strategy, and stage.

Figure 2: Representative views of the public LLM-SoccerArena website (https://llm-soccerarena.com).

open-book + probabilistic-forecast. All four conditions use the same output schema; the manipulation changes the information-access setting and the instruction style, not the required output format.

4.4 Access conditions

Closed-book. The model receives only fixture-identifying information and must rely on internal model knowledge, general football reasoning, and typical football score distributions. Closed-book calls are made without web-search tools, and the prompt explicitly forbids internet search, browsing, tools, APIs, or external data sources. The fixture block contains sport, competition, tournament edition (FIFA World Cup 2026), stage, UTC date, the listed home and away teams, venue (or *unknown*), and whether the match is a knockout match. Closed-book prompts do not include manually curated current information such as recent form, rankings, Elo ratings, injuries, suspensions, expected lineups, betting odds, or news.

Open-book. The model receives the same fixture block but is additionally instructed to use web search before predicting. Open-book calls enable the provider's web-search server tool. Open-book is defined as *tool-enabled and prompted to retrieve current public information*; it is not defined as guaranteed complete or optimal retrieval. For every open-book prediction the system logs whether search/tool use was actually observed (number of tool calls and citations/annotations). The primary open-book analysis follows an intent-to-treat principle; a secondary per-protocol analysis considers only predictions with observed search.

4.5 Prompt strategies

Direct-score. The model first predicts the most likely scoreline and then provides probabilities, expected goals, and full-match/advancement probabilities consistent with that scoreline. This

Model	Provider	Version tag	Role
Claude Fable 5	Anthropic	`anthropic/claude-5-fable-20260609`	latest experimental
Claude Opus 4.8	Anthropic	`anthropic/claude-4.8-opus-20260528`	flagship (Opus)
DeepSeek V4 Pro	DeepSeek	`deepseek/deepseek-v4-pro-20260423`	flagship
GPT-5.5	OpenAI	`openai/gpt-5.5-20260423`	frontier general
Gemini 3.1 Pro	Google	`google/gemini-3.1-pro-preview-20260219`	flagship Pro
Grok 4.3	xAI	`x-ai/grok-4.3-20260430`	current-info flagship
Mistral Large 2512	Mistral	`mistralai/mistral-large-2512`	most capable Mistral
Qwen 3.7 Max	Qwen	`qwen/qwen3.7-max-20260520`	flagship

Table 1: Active benchmark model roster (one flagship per provider). Version tags are the canonical provider/version identifiers stored with every prediction.

Factor	Condition 1	Condition 2
Information access	closed-book	open-book
Prompt strategy	direct-score	probabilistic-forecast

Table 2: Core 2×2 experimental design.

approximates the common consumer use case of asking a model for the final score.

Probabilistic-forecast. The model first estimates calibrated 90-minute result probabilities and expected goals, and then derives the most likely scoreline from that forecast. This approximates a more formal forecasting workflow.

The exact prompt components are given in Appendix A.

5 Forecast Horizons

The benchmark uses three forecast timing types, each stored with full timing metadata (scheduled prediction time, actual prediction time, kickoff time, minutes before kickoff, and a timing status of `on_time/early/late/missed/fallback`).

Primary horizon: T–24h. Predictions generated approximately 24 hours before scheduled kickoff (target offset 1440 minutes) form the main endpoint. Primary analyses use valid T–24h predictions only.

Secondary horizon: T–2h. Predictions generated approximately 2 hours before kickoff (target offset 120 minutes) are a secondary add-on. They may benefit from more recent public information (expected lineups, late injuries, tactical news, market movement) and are especially relevant for open-book models. This horizon is implemented and scheduled; predictions populate automatically as matches approach kickoff.

Fallback/early horizon: stage-opening. For the group stage, all known group-stage fixtures are predicted once before the tournament starts. For knockout stages, each fixture is predicted once shortly after the matchup becomes known. Stage-opening predictions support website robustness and secondary analyses; they are not used to impute missing T–24h predictions in the primary

analysis. The stage-opening scheduler refuses partial stages and runs only once a full round's fixtures are known.

A due-prediction scheduler invokes the time-based horizons within configurable polling windows and records whether each prediction was on time or late; a database-backed lock prevents concurrent duplicate runs.

6 Prediction Targets and Output Schema

6.1 Definitions

The primary target is the *90-minute result*, defined as the score after regulation time plus stoppage time, excluding extra time and penalties. The *full-match result* is the final outcome after all applicable extra time and penalty procedures. For group-stage matches, the full-match result equals the 90-minute result. For knockout matches, advancement probabilities describe the probability that each listed team advances or wins the tie after extra time and penalties if needed.

6.2 Required model outputs

Each model returns a single JSON object with: 90-minute home/draw/away probabilities; expected home and away goals; the most likely 90-minute scoreline; full-match home/draw/away probabilities; the most likely full-match scoreline; knockout advancement probabilities where applicable; a confidence value; and a short reason. The required schema is given in Appendix B. For group-stage matches, `home_advances_prob` and `away_advances_prob` are null; for knockout matches they are required and must sum to one. An optional scoreline probability grid is a possible future extension, not part of the core benchmark.

7 Tournament-Level Questions

In addition to per-match forecasts, LLM-SoccerArena elicits 15 tournament-level ("special") questions once per model and condition at the stage-opening horizon, mirroring popular prediction-game formats. These are: the overall *world champion*; the four *semifinalists*; the *12 group winners* (Groups A–L); and the nation that will provide the *tournament top scorer*.

Two question types are supported. *Single-choice* questions (champion, group winners, top-scorer nation) ask for a calibrated probability over every valid candidate team plus a single final pick; the candidate probabilities must sum to one. *Multi-choice with fixed k* (semifinalists, $k = 4$) asks for a probability over candidates plus exactly k final picks. Candidate sets are restricted (all teams, or the relevant group) and the prompt forbids using the project's own stored predictions or analytics. Each response is validated for candidate membership, pick count, probability range, and probability sum, with the same normalization-and-single-repair policy as match predictions (Section 8); per-candidate probabilities and ranks are stored individually. These questions will be scored with prediction-game points once the corresponding outcomes are decided.

8 Validation, Normalization, and Repair

Each response is processed as raw response → parsed JSON → validated prediction, and the raw response is always preserved. The system validates JSON parseability, required fields, numeric

types, probability ranges, probability sums, non-negative integer scorelines, and reason-string validity. The probability vectors that must sum to one are the 90-minute home/draw/away vector, the full-match home/draw/away vector, and (for knockout matches) the advancement vector.

If a probability vector sums between 0.98 and 1.02, it is normalized deterministically and the normalization is logged; major probability errors are not silently fixed. If the initial response is invalid, the system allows exactly one repair attempt that asks the model to return valid JSON matching the schema without changing the substantive forecast unless required by JSON or probability constraints. If still invalid, the prediction is marked invalid after repair and excluded from scoring, but retained in the dataset and reported in reliability metrics. Validation statuses are: `valid`, `normalized`, `repaired`, `repaired_and_normalized`, `invalid_json`, `invalid_schema`, `invalid_probability_range`, `invalid_probability_sum`, `invalid_score`, `invalid_after_repair`, `api_error`, and `timeout`.

For every prediction the system stores full provenance: match, model/provider, and condition metadata; timing metadata; the prompt template id, a SHA-256 prompt hash, and the raw prompt; the raw response and provider response id; token counts, latency, and cost; tool-use metadata; the parsed and validated fields; and (after the match) evaluation metrics. This makes every prediction reconstructable, auditable, and scorable.

9 Evaluation Metrics

The benchmark uses a multi-metric evaluation framework rather than a single primary metric. All metrics below are implemented in the scoring engine and surfaced by the website's analytics layer.

Probabilistic result forecasting. For 90-minute home/draw/away probabilities, with one-hot realized class y and prediction p,

$$\text{Brier}(p, y) = \sum_{k \in \{\text{home,draw,away}\}} (p_k - y_k)^2, \qquad \text{LogLoss} = -\log(p_c)$$

for realized class c, with safe clipping. Knockout advancement uses the binary analogues.

Categorical accuracy. Top-outcome accuracy (argmax of probabilities), tendency accuracy (from the predicted scoreline), and knockout advancement accuracy.

Scoreline quality. Exact-score accuracy, goal-difference accuracy, and absolute errors for goal difference, total goals, home goals, and away goals.

Game-style points. The website and secondary analyses use Kicktipp-style points: **5** for the exact score; **2** for the correct goal difference (but not exact score); **1** for the correct tendency (but not goal difference); and **0** otherwise.

Diagnostic and reliability metrics. Invalid-output rate, repair rate, normalization rate, missing-prediction rate, API-error rate, open-book search-observed rate, score/probability consistency, and expected-goals/scoreline distance. Internal consistency is treated as a diagnostic family, not a primary forecast-performance metric. The most important consistency check is whether the result implied by the most likely scoreline matches the highest-probability 90-minute outcome class. Formal metric definitions are collected in Appendix C.

10 Preliminary Live-Benchmark Diagnostics

We report diagnostics computable *before* any outcome is known, from the predictions already collected: 2,368 match forecasts (2,304 stage-opening and 64 T–24h, balanced across the 2×2 design; Table 3) and 480 tournament-question forecasts, generated for 72 group-stage fixtures and the 15 special questions. These describe operational behaviour and output structure; they are *not* forecast-accuracy results.

10.1 Match-prediction diagnostics

Horizon	Access	Direct	Probabilistic	Total
Stage-opening	closed-book	576	576	1152
Stage-opening	open-book	576	576	1152
T–24h	closed-book	16	16	32
T–24h	open-book	16	16	32
Match predictions				**2,368**
Tournament-question predictions				480

Table 3: Collected prediction coverage by horizon, access condition, and prompt strategy. The design is balanced across the four 2×2 cells. Knockout-match and T–2h cells populate automatically as fixtures and kickoffs approach.

Output validity is high. Across all 2,368 match predictions, 100.0% are valid for scoring, with 0 invalid responses and only 15 requiring a single JSON repair (Table 4). Tournament-question prompts, which require a full probability distribution over many candidate teams, show somewhat higher normalization and repair rates, consistent with their greater output complexity.

Model	N	Valid	Repaired	Normalized	Invalid
Claude Fable 5	296	293	3	0	0
Claude Opus 4.8	296	296	0	0	0
DeepSeek V4 Pro	296	289	7	0	0
GPT-5.5	296	296	0	0	0
Gemini 3.1 Pro	296	296	0	0	0
Grok 4.3	296	296	0	0	0
Mistral Large 2512	296	291	5	0	0
Qwen 3.7 Max	296	296	0	0	0
All	2368	2353	15	0	0

Table 4: Per-model output reliability for match predictions: total predictions (N), and counts of valid, single-repair, normalized, and invalid responses. All responses are ultimately valid for scoring.

Open-book search behaviour varies widely across models. Overall, 80.5% of open-book calls show observed web-search activity, but compliance ranges from below half to effectively all

calls depending on the model (Figure 4a), and the mean number of tool calls per prediction differs by more than an order of magnitude. This heterogeneity motivates the intent-to-treat versus per-protocol distinction in the analysis plan, and shows why logging observed tool use—rather than assuming it—is essential.

Open-book access is far more expensive. Open-book calls cost roughly an order of magnitude more than closed-book calls and are substantially slower, driven mainly by retrieved search context in the input (Table 5). Total spend to collect the current dataset is approximately USD 279.1. These figures are directly relevant to the feasibility and design of live LLM-forecasting deployments.

Mean per call	Closed-book	Open-book
Cost (USD)	0.012	0.187
Latency (s)	1.6	5.2
Input tokens	835	27,978
Output tokens	802	1,656

Table 5: Mean per-call operational profile, closed-book versus open-book match predictions. Open-book input-token counts include retrieved web-search context.

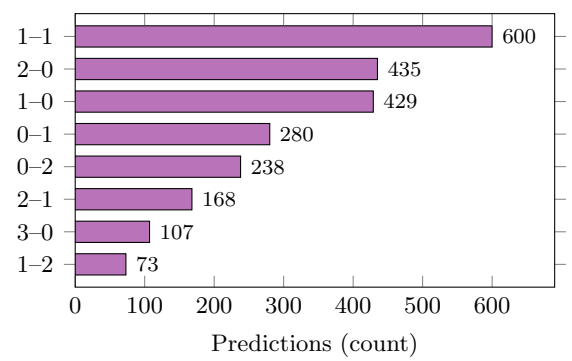
Models differ in confidence, sharpness, and goal expectations. Even before any outcome, models display distinct forecasting "signatures" (Table 6). Mean stated confidence ranges from about 0.53 to 0.82, and the mean top-class 90-minute probability (a sharpness proxy) and mean predicted total goals also vary noticeably across models. Strikingly, *draws are almost never the modal outcome*: across all models, the share of predictions whose highest-probability 90-minute outcome is a draw is close to zero.

Model	Mean conf.	Top-class prob.	Draw-modal %	Mean goals
Claude Fable 5	0.57	0.57	0.3	1.70
Claude Opus 4.8	0.59	0.56	0.0	1.90
DeepSeek V4 Pro	0.71	0.59	0.3	1.88
GPT-5.5	0.53	0.56	0.0	1.51
Gemini 3.1 Pro	0.82	0.61	0.0	1.74
Grok 4.3	0.67	0.60	0.7	2.44
Mistral Large 2512	0.76	0.52	2.7	2.21
Qwen 3.7 Max	0.73	0.58	0.0	1.55

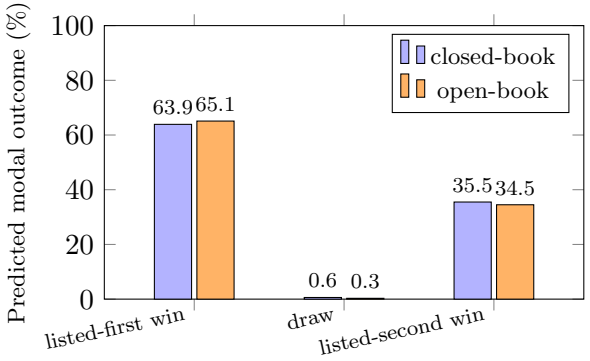
Table 6: Per-model forecasting behaviour over valid predictions: mean confidence, mean top-class 90-minute probability (sharpness), share of predictions with a draw as the modal outcome, and mean predicted total goals.

Predicted scorelines concentrate on low scores, yet draws are rarely favoured outright. The single most frequently predicted 90-minute scoreline is 1–1, followed by low-scoring wins such as 2–0, 1–0, and 0–1 (Figure 3a). At the same time, the modal *probability* outcome is a win in almost every prediction, with a strong lean toward the listed-first team and very little probability mass placed on draws as the single most likely result (Figure 3b). This tension between a popular

1–1 point prediction and an almost draw-free probability argmax is exactly the kind of structured-output behaviour the benchmark is designed to surface; it also drives part of the internal-consistency gap below.



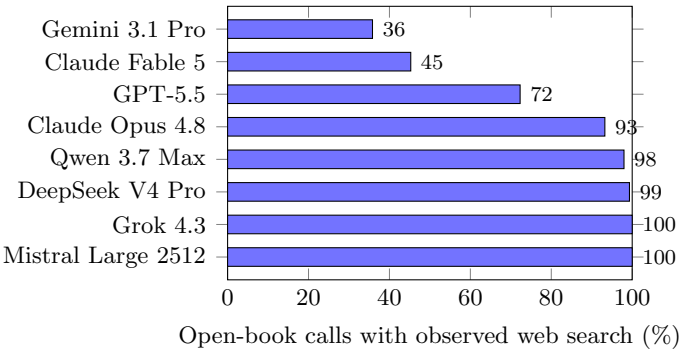
(a) Most frequent predicted 90-minute scorelines.



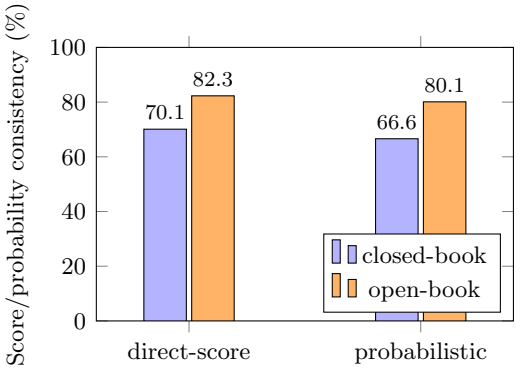
(b) Modal-outcome mix by access condition.

Figure 3: Predicted scoreline and outcome structure (pre-outcome). Draws dominate point predictions (1–1) but are almost never the probability argmax, and predictions lean toward the listed-first team.

Structured forecasts are largely, but not perfectly, internally coherent. For most valid predictions, the most likely scoreline implies the same 90-minute result as the probability argmax (Figure 4b); open-book predictions are somewhat more coherent than closed-book ones, and direct-score prompting is marginally more coherent than probabilistic prompting. The residual inconsistency is partly explained by the 1–1 effect above. Whether these patterns persist and whether they relate to forecast accuracy is part of the post-tournament analysis.



(a) Open-book search-observed rate by model.



(b) Score/probability consistency by condition.

Figure 4: Operational and coherence diagnostics. (a) Enabling the web-search tool does not guarantee a model uses it. (b) Share of predictions whose most likely scoreline matches the probability argmax.

10.2 Tournament-question forecasts

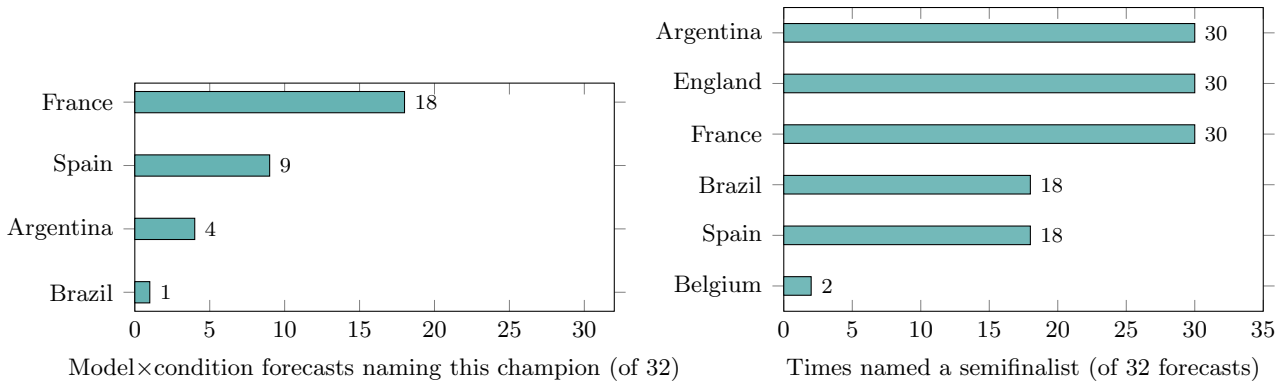
Strong consensus on group winners, more disagreement on the title. For the 15 tournament questions, models agree almost completely on group winners—in 11 of 12 groups every forecast names the same winner (Table 8)—but disagree more on the eventual champion. Across the 8 models and four conditions, world-champion forecasts spread over a handful of contenders (Figure 5a), and the four semifinalist slots concentrate on a small set of favourites (Figure 5b). Table 7 shows each model's headline picks under a single canonical condition (open-book, probabilistic): the title pick splits between a small number of teams, whereas the predicted top-scorer nation is unanimous.

Model	Champion	Top-scorer nation	Semifinalists
Claude Fable 5	Spain	France	Spain, France, England, Argentina
Claude Opus 4.8	Spain	France	Spain, France, Argentina, England
DeepSeek V4 Pro	France	France	Spain, France, England, Argentina
GPT-5.5	Spain	France	Spain, France, England, Argentina
Gemini 3.1 Pro	Spain	France	Spain, France, England, Argentina
Grok 4.3	Argentina	France	Argentina, Brazil, England, France
Mistral Large 2512	France	France	Spain, England, Argentina, France
Qwen 3.7 Max	France	France	France, Argentina, Spain, England

Table 7: Per-model headline tournament forecasts under the canonical open-book, probabilistic-forecast condition: champion, top-scorer nation, and the four predicted semifinalists.

Group	Consensus winner	Agreement	Runner-up
A	Mexico	94%	Czechia
B	Switzerland	100%	–
C	Brazil	100%	–
D	United States	100%	–
E	Germany	100%	–
F	Netherlands	100%	–
G	Belgium	100%	–
H	Spain	100%	–
I	France	100%	–
J	Argentina	100%	–
K	Portugal	100%	–
L	England	100%	–

Table 8: Group-winner consensus across all model×condition forecasts: the leading pick, the share of forecasts naming it, and the runner-up where the leading pick is not unanimous.



(a) World-champion forecasts.

(b) Semifinalist nominations.

Figure 5: Tournament-outcome forecasts aggregated over model×condition. (a) Teams named as champion. (b) Teams named among the four semifinalists (each forecast names four teams).

11 Evaluation Plan for Post-Tournament Analysis

The platform is designed so that a full confirmatory evaluation can be carried out once outcomes are known. We sketch a plan here; it is intended as a useful and transparent default rather than a binding pre-registered protocol, and it may be refined as data accrue. The metrics and the primary endpoint (valid T–24h predictions, one deterministic prediction per model, match, horizon, access condition, and prompt strategy) are fixed in advance to limit researcher degrees of freedom.

A natural way to quantify uncertainty is bootstrapping over matches: matches are resampled with replacement and metrics recomputed; for condition or model comparisons, paired bootstrapping computes differences on the same resampled match set. Comparisons of interest include (1) model rankings across the multi-metric framework; (2) open-book versus closed-book; (3) probabilistic-forecast versus direct-score; (4) the access × strategy interaction; (5) output validity and coherence across models and conditions; and (6) knockout advancement performance. T–2h and stage-opening predictions are treated as secondary; stage-opening predictions would not be used to replace missing T–24h predictions in the primary comparison. Because the World Cup contains a limited number of matches, results should be reported with confidence/bootstrap intervals and small differences interpreted cautiously.

Beyond accuracy, the stored data enable several qualitative and exploratory analyses. Every prediction retains the model's free-text `reason` (and, for tournament questions, a `reasoning_summary`), so the reasoning behind forecasts can be studied directly: for example, how often and how models cite retrieved information such as form, injuries, or betting-market odds under the open-book condition; whether stated rationales align with the numerical forecast; how reasoning differs across model families and between direct-score and probabilistic prompting; and whether particular reasoning patterns are associated with better-calibrated or more accurate forecasts once outcomes are known. The already-observed heterogeneity in search behaviour and confidence (Section 10) makes such reasoning analysis a promising direction.

The evaluation engine that produces every required metric is implemented and runs automatically after result synchronization. When official results arrive, the per-match metrics (Brier, log loss, accuracy, scoreline errors, Kicktipp points, advancement metrics) and the tournament-question scores populate without further code changes; the corresponding leaderboard and comparison tables would then be filled in.

12 Dataset and Reproducibility Release

LLM-SoccerArena produces a timestamped prediction archive containing prompts, raw responses, parsed predictions, validation/repair/normalization metadata, tool-use metadata, match outcomes, and evaluation metrics. A single export command writes five files: fixture metadata (CSV); raw predictions including invalid/API-error rows (CSV); validated predictions with parsed fields, validity flags, and probability sums (CSV); evaluations left-joined to predictions, with metrics populated when available (CSV); and open-book tool-use logs with raw tool traces (JSONL). Exports intentionally retain invalid and unevaluated rows so that missingness, API errors, repair rates, normalization rates, and search-observed rates can be analyzed directly. The full SQLite database—containing raw prompts, raw responses, validations, evaluations, and scheduler state—is the primary research archive and is backed up daily. The preliminary tables and figures in Section 10 are produced by a single reproducible script over this database.

A `predictor_type` field (`llm` vs. `baseline`) is supported in the schema so that non-LLM baselines can be added later without migration; baselines are not part of the core benchmark.

13 Baselines and Future Extensions

The core benchmark focuses on controlled LLM comparisons. Non-LLM baselines are planned as optional secondary or post-hoc additions, not requirements: uniform probabilities; historical World Cup or international base rates; FIFA-ranking or Elo-based models [Hvattum and Arntzen, 2010, World Football Elo Ratings, 2024]; Poisson scoreline models [Dixon and Coles, 1997, Karlis and Ntzoufras, 2003]; and bookmaker-implied probabilities [Štrumbelj, 2014, Zeileis et al., 2018]. Bookmaker-odds ingestion is not required; betting-style analyses may later be derived from model probabilities and bookmaker-implied probabilities if reliable, timestamp-aligned odds can be collected. Public-user comparisons are website/product data only and are not part of the core paper benchmark.

14 Limitations

The benchmark has several limitations. First, the number of World Cup matches is limited, so model rankings may carry substantial uncertainty. Second, model APIs and versions may change over time; we store canonical version tags with every prediction to mitigate this. Third, open-book search behaviour is not standardized across providers, and enabling web search does not guarantee retrieval—indeed, observed search rates vary widely across models (Section 10). Fourth, exact-score prediction is inherently noisy and should be interpreted alongside probabilistic, categorical, and goal-difference metrics. Fifth, forecast-accuracy results are not yet available; this paper reports the platform, protocol, dataset, and pre-outcome diagnostics, with the confirmatory evaluation to follow. Sixth, the core benchmark does not rely on public users, bookmaker odds, or non-LLM baselines, although these may be valuable future additions.

15 Conclusion

LLM-SoccerArena is a live, deployed platform and benchmark for evaluating LLM football forecasting during the FIFA World Cup 2026. It combines a public-facing prediction website with a benchmark protocol fixed in advance, timestamped prediction logging, controlled closed-book and open-book access conditions, direct and probabilistic prompting strategies, soccer-specific

and tournament-level evaluation, and reproducible data exports. The system has already collected thousands of timestamped, validated pre-match forecasts whose diagnostics we report here; forecast-accuracy evaluation will follow automatically as official results arrive. LLM-SoccerArena is intended to support both public engagement and scientific analysis of LLM forecasting, tool use, prompt sensitivity, and structured prediction reliability in a real-world future-event setting.

A Prompt Template Components

Prompts are assembled from reusable components:

access header + strategy instruction + definitions block + fixture block + JSON schema block.

Closed-book header

Access condition: CLOSED_BOOK. Do not use internet search, browsing, tools, APIs, or external data sources. Use only your internal model knowledge, general football reasoning, typical football score distributions, and the fixture information provided below. The fixture information identifies a current FIFA World Cup 2026 football/soccer match.

Open-book header

Access condition: OPEN_BOOK. Before making the prediction, use the available web-search tool to check current public information about this match and both teams. Relevant information may include recent form, injuries, suspensions, expected lineups, tactical news, venue, rest/travel, tournament context, and betting-market odds if available. Base the final prediction on the retrieved public information plus your football reasoning.

Direct-score instruction

Prompt strategy: DIRECT_SCORE. First predict the most likely scoreline for the match. Then provide probabilities, expected goals, and full-match/advancement probabilities that are consistent with that predicted scoreline. Do not overstate certainty.

Probabilistic-forecast instruction

Prompt strategy: PROBABILISTIC_FORECAST. First estimate calibrated probabilities for the 90-minute result and expected goals. Then derive the most likely scoreline from those probabilities and expected goals. Do not overstate certainty.

B JSON Output Schema

```
{
  "home_win_90_prob": number,
  "draw_90_prob": number,
  "away_win_90_prob": number,
  "expected_home_goals_90": number,
  "expected_away_goals_90": number,
  "most_likely_score_90": {
    "home": number,
    "away": number
  },
  "home_win_full_prob": number,
  "draw_full_prob": number,
  "away_win_full_prob": number,
```

```
  "most_likely_score_full": {
    "home": number,
    "away": number
  },
  "home_advances_prob": number or null,
  "away_advances_prob": number or null,
  "confidence": number,
  "reason": "short reason"
}
```

C Metric Definitions

Brier score. For classes $K = \{\text{home}, \text{draw}, \text{away}\}$: $\text{Brier}(p, y) = \sum_{k \in K} (p_k - y_k)^2$.

Log loss. For realized class c: $\text{LogLoss}(p, c) = -\log(p_c)$, with safe clipping.

Goal-difference / total-goals absolute error. $|\hat{d} - d|$ for predicted vs. actual goal difference, and $|\hat{t} - t|$ for predicted vs. actual total goals.

Kicktipp-style points. Exact score: 5 points; correct goal difference (not exact): 2 points; correct tendency (not goal difference): 1 point; otherwise: 0 points.

References

Anthony C. Constantinou and Norman E. Fenton. Solving the problem of inadequate scoring rules for assessing probabilistic football forecast models. *Journal of Quantitative Analysis in Sports*, 8(1), 2012. doi: 10.1515/1559-0410.1418.

Mark J. Dixon and Stuart G. Coles. Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C*, 46(2):265–280, 1997. doi: 10.1111/1467-9876.00065.

Edward S. Epstein. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8(6):985–987, 1969. doi: 10.1175/1520-0450(1969)008⟨0985:ASSFPF⟩2.0.CO;2.

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

Tilmann Gneiting, Fadoua Balabdaoui, and Adrian E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, 69(2):243–268, 2007. doi: 10.1111/j.1467-9868.2007.00587.x.

Andreas Groll, Gunther Schauberger, and Gerhard Tutz. Prediction of major international soccer tournaments based on team-specific regularized poisson regression: An application to the fifa world cup 2014. *Journal of Quantitative Analysis in Sports*, 11(2):97–115, 2015. doi: 10.1515/jqas-2014-0051.

Andreas Groll, Christophe Ley, Gunther Schauberger, and Hans Van Eetvelde. A hybrid random forest to predict soccer matches in international tournaments. *Journal of Quantitative Analysis in Sports*, 15(4):271–287, 2019. doi: 10.1515/jqas-2018-0060.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2402.18563.

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, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E. Tetlock. ForecastBench: A dynamic benchmark of AI forecasting capabilities. https://arxiv.org/abs/2409.19839, 2024. arXiv:2409.19839.

Dimitris Karlis and Ioannis Ntzoufras. Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D*, 52(3):381–393, 2003. doi: 10.1111/1467-9884.00366.

Siem Jan Koopman and Rutger Lit. A dynamic bivariate poisson model for analysing and forecasting match results in the english premier league. *Journal of the Royal Statistical Society: Series A*, 178(1):167–186, 2015. doi: 10.1111/rssa.12042.

Christoph Leitner, Achim Zeileis, and Kurt Hornik. Forecasting sports tournaments by ratings of (prob)abilities: A comparison for the euro 2008. *International Journal of Forecasting*, 26(3): 471–481, 2010. doi: 10.1016/j.ijforecast.2009.10.001.

Christophe Ley, Tom Van de Wiele, and Hans Van Eetvelde. Ranking soccer teams on the basis of their current strength: A comparison of maximum likelihood approaches. *Statistical Modelling*, 19(1):55–73, 2019. doi: 10.1177/1471082X18817650.

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

Marcel Meyer, Sascha Kaltenpoth, Henrik Albers, Kevin Zalipski, and Oliver Müller. TS-Arena – a live forecast pre-registration platform. https://arxiv.org/abs/2512.20761, 2025. arXiv:2512.20761.

Brian A. Nosek, Charles R. Ebersole, Alexander C. DeHaven, and David T. Mellor. The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11):2600–2606, 2018. doi: 10.1073/pnas.1708274114.

Gunther Schauberger and Andreas Groll. Predicting matches in international football tournaments with random forests. *Statistical Modelling*, 18(5-6):460–482, 2018. doi: 10.1177/1471082X18799934.

Erik Štrumbelj. On determining probability forecasts from betting odds. *International Journal of Forecasting*, 30(4):934–943, 2014. doi: 10.1016/j.ijforecast.2014.02.008.

World Football Elo Ratings. World football elo ratings methodology. https://www.eloratings.net/about, 2024. Web source, not peer-reviewed.

Achim Zeileis, Christoph Leitner, and Kurt Hornik. Probabilistic forecasts for the 2018 fifa world cup based on the bookmaker consensus model. Technical Report 2018-09, University of Innsbruck, Working Papers in Economics and Statistics, 2018. Working Paper.