Calibration over Accuracy: A Pre-Registered Live Benchmark of Three AI Reasoning Architectures on the 2026 FIFA World Cup

Mo Kahlain
inocta.io, Toronto, Canada
mo@inocta.io

Compiled July 21, 2026
Pre-registration: OSF ke35t, DOI 10.17605/OSF.IO/R5SBJ (frozen before the first scored match)
Live dashboard: https://match.inocta.io

Abstract

When an AI system reports 70 percent confidence, is the number worth anything? We present inocta-bench, a pre-registered live benchmark that compares three AI reasoning architectures on probabilistic *calibration*, not accuracy: a single-call generalist (Solo), a sequential specialist pipeline (Pipeline), and a multi-model council (Council). All three forecast the 104 matches of the 2026 FIFA World Cup in public and in real time. Instead of comparing model classes on accuracy, we compare reasoning architectures on calibration, with a pre-registered protocol and the strongest published baseline (Elo with a home dummy, per Robberechts and Davis 2019) as our comparison line, alongside the betting market's closing odds and a human forecaster. Two ablation controls (a context-free twin of the generalist and a learning-loop-disabled twin of the pipeline) were registered to separate the contribution of context from the contribution of architecture, and genuine learning from retrieval. The learning-loop layer was never populated in production, so the learning question is reported as not testable as designed rather than as a finding (Section 7). Every prediction, cost, and miss is published before each match begins. The protocol, prompts, and scoring rules were hash-locked and registered (OSF 10.17605/OSF.IO/R5SBJ) before the first scored match. The tournament is complete, and the answer an operator can act on is on the calibration axis. Across the full tournament the council, an ensemble of three independent model families with a step that weighs their disagreement, posts the lowest expected calibration error of every machine tested, the human reference aside (0.100), below the single-call and pipeline architectures, the deterministic baseline, and the market's own closing line. It is the architecture whose stated confidence tracks reality most closely, and for an organization choosing an AI shape to trust with a forecast it is the one to reach for. We report the margin honestly: on a chance-dominated event the bootstrap intervals overlap, so the lead is directional rather than resolved, and a second tournament would sharpen it. Accuracy was never the question, and it behaved as a chance-dominated task predicts: no architecture reliably called more matches than another, and the mild inversion of our stated prior, the AI lanes edging below the market's closing line on the held-out set, is a directional reading we do not press. That the AI lanes land near the market at all is a property of football, which has a deep public betting market already pricing every match. A business forecasting its own pipeline has no such reference, and the calibrated AI probability is the closest thing to one it can build, which is what makes the calibration axis, and the council on it, the operational result of the study. We publish every prediction, cost, and the full dataset with the paper, so the direction the result gives can be checked against the evidence it rests on.

Keywords: calibration, probabilistic forecasting, large language models, reasoning architectures, pre-registration, sports forecasting

1 Introduction

An AI system that reports 70 percent confidence is making a promise. If picks issued at 70 percent confidence win about 70 percent of the time, the number carries information a decision-maker can act on: it tells you what to delegate, what to hedge, and when to escalate to a human. If picks at 70 percent win 50 percent of the time, the system is not merely imprecise. It is miscalibrated, and every downstream decision built on its confidence inherits the defect. Inside organizations adopting AI, this is the operative question. Whether a system is right slightly more or less often than an alternative matters far less than whether its own account of its uncertainty can be trusted, because trust in the reported confidence is what determines which decisions can be handed over at all.

This paper asks that question experimentally, at the level where most applied AI decisions are actually made: not the choice of base model, but the choice of *reasoning architecture*: how calls, roles, and context are arranged around models. We compare three architectures that mirror the patterns organizations deploy today: a single-call generalist with rich context (SOLO), a sequential pipeline of specialist roles (PIPELINE), and a council of three independent model families whose outputs are merged by a decider (COUNCIL). The research question, fixed in the pre-registration before any match was scored, is:

> *On a chance-dominated forecasting task (World Cup match outcomes), how do three AI reasoning architectures differ in probabilistic calibration, not accuracy, relative to each other and to non-AI baselines?*

Football match forecasting is the testbed, not the subject. It offers what calibration measurement needs and what retrospective benchmarks cannot supply: a stream of 104 genuinely uncertain events on a fixed schedule, an unambiguous ground truth within two hours of each prediction, a hard information horizon (predictions lock 60 to 75 minutes before kickoff), and, decisively, a public, liquid reference forecast in the betting market's closing odds, which decades of forecasting literature treat as the strongest available aggregation of public information. A task dominated by irreducible chance is precisely where calibration separates from accuracy: when no forecaster can be right much more often than the odds allow, the honest description of one's own uncertainty becomes the measurable skill.

Three design commitments distinguish this benchmark from retrospective evaluations. **First, it is pre-registered and live.** The protocol, prompts, model pins, scoring rules, and claim conditions were hash-locked and registered on OSF (DOI 10.17605/OSF.IO/R5SBJ) before the first scored match. Every prediction is published on a public dashboard (`https://match.inocta.io`) before kickoff, with per-prediction cost and latency. There is no possibility of selective reporting, post-hoc metric choice, or quiet test-set reuse. **Second, it carries its own controls.** A context-free twin of the generalist (SOLO-ZERO) isolates what context contributes versus what the underlying model already knows. A learning-loop-disabled twin of the pipeline (PIPELINE-STATIC) was registered to turn "does the system actually learn during the tournament?" into a measured ablation rather than a marketing claim, sharpened further by two synthetic patterns seeded into the context stream and absent from any public data. (The loop layer was never populated in production. Section 7 reports what the control pair can and cannot say as a result.) **Third, it states a falsifiable prior.** We expect no AI lane to beat the closing market line on raw accuracy. The contribution is calibration measurement, cost transparency, and the learning-loop test, and we commit to publishing whatever the data shows, including inversions of our thesis.

The AI lanes are benchmarked against three non-AI reference points: an internally computed Elo rating with a host-nation home adjustment (the strongest published baseline class for this task, per Robberechts and Davis [12]), the de-vigged closing odds (MARKET), and a single named human forecaster serving as a non-blind reference line (HUMAN, Section 3). Section 3 specifies the architectures, controls, and dispatch discipline. Section 4 covers the data sources and the construction of the context packs and the Elo baseline. Section 5 gives the scoring contract and the pre-committed claim conditions. Results follow the tournament (Section 6), and the learning-loop analysis (Section 7) addresses the second, subordinate research question fixed in the registration:

> *Does an architecture with an active learning loop demonstrate genuine learning (versus retrieval), measured by (a) the gap between the looped architecture and its loop-disabled control, and (b) pickup of two seeded synthetic patterns absent from public data?*

Contributions. This paper makes four. The first is a pre-registered, fully public, live benchmark of AI reasoning architectures with calibration as the headline metric. The second is a set of matched ablation controls registered to separate context from architecture and learning from retrieval, including a seeded-pattern probe (the learning half was not exercised in production and is reported as such, Section 7). The third is complete cost and latency transparency per prediction, treating operational economics as a first-class experimental output. The fourth is an open dataset, evaluation harness, and prompt set, hash-verified against the pre-registration and published with this paper.

2 Related work

This benchmark sits at the intersection of three literatures. Statistical football forecasting supplies the baselines and the feature families. Proper scoring and calibration supply the measurement frame. LLM-based forecasting and multi-agent architectures are the class of system under test.

Statistical World Cup forecasting. The feature families our context packs operationalize descend from the regularized-regression line of Groll et al. [7] and Groll et al. [8], whose LASSO-selected covariates, home and host advantage decomposed into host, continent, and confederation effects, squad structure, and market-value proxies, are the inputs a strong statistical model draws on for this task. We borrow these as the content of context layer L0, and we deliberately skip the macro covariates their variable-importance analyses find weak, GDP and population, which carry little marginal signal once Elo and market odds are present. The point of interest for us is not to improve on this line's accuracy but to ask whether an AI system handed the same class of information reports its confidence honestly.

The Elo-plus-home-dummy precedent. Robberechts and Davis [12] show that an ordered logit on Elo difference plus a home dummy beats every more complex model they test on the 2002–2014 World Cups, including bookmaker-augmented models and random forests (RPS 0.1860 versus 0.1877 for bookmakers). This is why ELO is a first-class lane rather than a footnote: it is the strongest simple baseline class for this task, so an architecture that cannot match a home-adjusted Elo has not earned its complexity. Our internal Elo (Section 4) follows this precedent in spirit, an Elo difference with a host-nation home term, computed transparently from match history so the baseline is reproducible rather than borrowed from an opaque third-party rating.

Proper scoring rules and calibration. The measurement frame descends from Brier [2] and Murphy [11] through the proper-scoring-rules synthesis of Gneiting and Raftery [6]. The Brier score is the proper score we report as the headline, its Murphy decomposition separates calibration (reliability) from discrimination (resolution), and ECE isolates the calibration piece on its own, the single number for the question this study asks: when a system says seventy percent, does it happen seventy percent of the time. Sharpness completes the picture, distinguishing a well-calibrated forecaster from a well-calibrated coward, so we report calibration, sharpness, and the decomposition together rather than any one in isolation.

Modern ML upper bounds. Al-Bustami and Ghazal [1] represents the contemporary accuracy-oriented approach, player-level features with bookmaker odds and Elo differentials as the dominant input class. We treat this line as framing rather than a target: it marks how far accuracy can be pushed on this task, which is precisely the axis a chance-dominated benchmark expects to saturate, and it is the reason we measure calibration instead.

Bayesian uncertainty in football models. Egidi et al. [5] and Robberechts et al. [13] carry an uncertainty-quantification thread that our calibration focus parallels from the AI side. Where these models express uncertainty through a posterior over outcomes, we ask the same underlying question of a language-model architecture, whether the uncertainty it reports is honest, and answer it with the calibration machinery above rather than with a generative model of the score line.

LLM forecasting and multi-agent architectures. On the AI side, a growing line asks whether language models can forecast uncertain future events at a useful level of skill. Halawi et al. [9] build a retrieval-augmented system that approaches, and in places exceeds, the aggregate of competitive human forecasters, and Karger et al. [10] introduce ForecastBench, a dynamic, contamination-free benchmark of LLM forecasting against human comparison groups. Both share our commitment to questions whose answers are unknown at prediction time, which removes the test-set-reuse confound that troubles retrospective evaluation. Where they measure accuracy against human crowds, we hold accuracy fixed by choosing a chance-dominated task and measure calibration instead, and we compare architectures rather than retrieval pipelines.

The council architecture we test descends from work on multi-agent reasoning. Du et al. [4] show that having several model instances debate over multiple rounds improves factual accuracy and reduces hallucination relative to a single call, and Chen et al. [3] propose ReConcile, a round-table of diverse model families that reaches a confidence-weighted consensus, which is the closest published relative of our Council decider. The evidence that more agents help is not one-directional, and we cite the tension rather than hide it. Zhu et al. [16] find that vanilla multi-agent debate often underperforms simple majority voting, and recovers an advantage only when the agents are diverse and their confidence is calibrated, precisely the two properties our council is built around. Tran and Kiela [15] go further and show that a single agent can match or beat multi-agent systems on reasoning once the thinking-token budget is held equal, attributing reported multi-agent gains to extra test-time compute rather than to the interaction itself. Our study speaks to this debate from a different angle: we do not claim the council reasons better, we ask whether its *confidence* is more trustworthy, which is a property none of these works measures directly. Calibration of LLM confidence is itself an active problem, with methods such as Thermometer [14] developed to recalibrate model probabilities after the fact, which is one more reason to measure calibration in situ rather than assume it.

3 Benchmark design

The benchmark runs eight forecasting lanes over every match of the 2026 FIFA World Cup (104 fixtures: 72 group-stage matches across three matchdays, then knockout rounds as brackets resolve). The scored target for every match is the 90-minute result, three-way (home / draw / away), at every tournament stage. All design elements in this section were fixed in the pre-registration (OSF 10.17605/OSF.IO/R5SBJ) before the first scored match. The registered documents and the production prompt and gateway code are bound by SHA-256 checksums (Appendix B).

3.1 Lanes

Lane	Type	Definition
Solo (A1)	architecture	Single pinned-model call, full 7-layer context pack
Solo-Zero (A1-zero)	control	Same code path as A1, fixture metadata only, zero context layers
Pipeline (A2)	architecture	3-step sequential pipeline, per-step schema validation, one retry
Pipeline-Static (A2-static)	control	A2 with the learning-loop layer removed, excluded from loop writes
Council (A3)	architecture	Three model families in parallel + a decider
Elo	baseline	Internal Elo from match history (Section 4)
Market	baseline	Closing odds, single T−1 h snapshot, proportionally de-vigged
Human	reference	One named human (the registrant), non-blind reference line, pick locked at kickoff

Table 1: The eight lanes: three architectures, two ablation controls, three non-AI baselines. Lane definitions are verbatim commitments from the pre-registration §2.

3.2 The three architectures

Solo (single-call generalist). One call to a pinned model with a shared analyst system prompt and the full context pack. The structured output (winner, scoreline, confidence 0–100, at most three risk factors, a reasoning note of at most 100 words) is enforced by a JSON schema, with exactly one retry on schema failure, and a second failure is recorded as an abstention, never silently repaired.

Pipeline (sequential specialists). Three calls in sequence on the same pinned model family. A *statistician* reads only the quantitative layers of the context pack. An *operations reader* then combines the statistician's output with the narrative layers (news, rivalries, prior insights, the learning loop when present). A *voice editor* synthesizes both into the final structured prediction. Each step carries its own output schema and the same one-retry rule. An escalation gate is part of the registered architecture: if the pipeline's final confidence falls below 60, it adopts the Council's prediction for that match, and the adoption is logged and cost-attributed: the pipeline pays for its own three steps regardless.

Council (multi-model council). Three members from distinct model families (Claude, GPT, Gemini, with exact version pins in Appendix B) receive the identical prompt and context in parallel. A decider from the first family receives the three submissions *with model identities removed* (anonymous "member 1/2/3"), is instructed that its confidence should reflect member disagreement, and produces the lane's single prediction. Model identities are logged in telemetry and published, but are not visible to the decider

at decision time. Architecture and model mix are therefore confounded by design. All claims in this study concern the pinned systems as specified, not reasoning architectures in the abstract.

3.3 Controls

SOLO-ZERO runs the literal SOLO code path with the context pack withheld: fixture metadata only. The gap between the twins prices the context pack: any calibration or accuracy difference is attributable to context, not architecture, because nothing else differs. PIPELINE-STATIC runs the full pipeline with the learning-loop layer nulled out of its context and is excluded from learning-loop writes, and it runs on every match PIPELINE runs. This control carries the Level-2 research question (Section 7).

3.4 Dispatch discipline and information horizon

AI lanes predict exactly once per match, dispatched when the match enters the T−75 min pre-kickoff window on a 15-minute scheduler (in practice 60 to 75 minutes before kickoff). Predictions are never revised after dispatch. Dispatch is claim-guarded: redundant clocks race on a transactional per-match claim, so duplicate scheduling cannot produce duplicate predictions. The MARKET baseline is a single T−1 h snapshot, so the AI lanes and the market sit at approximately the same information horizon. The market line never enters any AI lane's context: the odds pipeline is a separate code path with no contact with the context packs. Official lineups publish roughly 35 minutes before kickoff, after the dispatch window closes, so no lane ever sees them. The HUMAN lane may be entered any time before kickoff and locks at kickoff.

The human lane is a non-blind reference line, not a controlled baseline. The named human forecaster was the project's registrant and operator. In that role, the human had access, before locking each pick, to the AI lanes' published predictions and reasoning for the same match, and, across the tournament, to the running scoreboard and the lane-performance feedback that a blinded forecaster would not see. The human pick is therefore informed by the system it is compared against, on both a per-match channel (visibility of the AI outputs) and a tournament-long channel (feedback and adaptation). We report the HUMAN lane for transparency and completeness, but its results carry no claim of an independent human-versus-machine comparison, and any calibration advantage it shows is confounded by information access. A clean human baseline would require a forecaster blind to all lane outputs and to the running results. We did not run one, and we flag this as a limitation (Section 10) rather than obscure it.

3.5 Methodological commitments (M1–M8)

The registration binds eight commitments.

- **M1.** Pre-registration on OSF before the first scored match.
- **M2.** Multi-judge evaluation of narrative outputs (three model families as judges, ground truth from the data provider, a 10 % stratified human spot-check).
- **M3.** Version pinning: every call logs both the requested pin and the provider-reported effective model identifier, so silent model substitution is detectable in the published telemetry.
- **M4.** The loop control runs on every match the looped lane runs.

- **M5.** Two synthetic patterns, sealed in an annex before the tournament, injected into the context stream at deterministic locations, with pickup measured by a detector, the probe for learning versus retrieval.
- **M6.** The final 20 matches of the tournament are frozen as the held-out set, and no prompt, rubric, or configuration changes are permitted after the 84th tournament match.
- **M7.** Human baselines: the registrant from launch, named public writers as a possible mid-tournament addition under the registered amendment process.
- **M8.** The calibration reporting contract of Section 5.

3.6 Development window and amendments

Matches before the registration lock are labeled development data wherever they appear and enter no headline claim, and headline claims are carried by held-out performance. Any protocol amendment must be versioned and published before any match it affects, and the amendment log is part of the public registry (Appendix B).

4 Data

4.1 Sources

Fixtures, results, team statistics, historical match data, and closing odds come from a single commercial provider (API-Football, PRO tier), verified by a pre-registration probe on response shapes and coverage. News and rivalry narratives are harvested by a search-grounded model (Perplexity Sonar) with URL citations stored per item. Weather comes from venue coordinates at forecast time.

Coverage is imperfect and the design accounts for it rather than hiding it. Several provider fields, lineup availability and some historical statistics in particular, read as unavailable before the tournament and remained sparse for lower-profile fixtures, so every context layer is nullable and a missing layer is logged per match rather than silently substituted. The governing principle is that a gap is a gap for every lane equally: when a layer fails to fetch, no lane receives it, so no lane is advantaged or disadvantaged by a coverage hole. The one lane that receives no context layers at all by design, the context-free twin Solo-Zero, is therefore not a special case of missing data but the deliberate zero-context control against which the others are read.

4.2 The 7-layer context pack

Every context-consuming lane draws from a single harvested pack of up to seven layers (Table 2). The pack is assembled once per match at harvest time, serialized with a cap of 120,000 characters, and the set of layers successfully fetched or failed is logged per match so that any prediction can be traced to exactly the context that was available when it was made. The context-free twin Solo-Zero runs the identical code path with the pack withheld, receiving only fixture metadata.

Layer	Content	Source	Nullable
L0	Team and match statistics	API-Football	yes
L1	Venue weather at forecast time	Venue coordinates	yes
L2	News narrative, with citations	Perplexity Sonar	yes
L3	Rivalry narrative, with citations	Perplexity Sonar	yes
L5	Generated pre-match insights	Model, from L0–L3	yes
L6	Pre and post-match analysis	Model	yes
L7	Learning-loop entries	Loop layer (never populated)	yes

Table 2: The context pack. Layer numbering follows the implementation and skips L4. Every layer is nullable and a failed fetch is logged, not substituted. The context lanes (Solo, Pipeline, Pipeline-Static, Council) consume the assembled pack. Solo-Zero receives none of it, and the deterministic baselines (Elo, Market) and the human reference use no pack. L7 was never populated in production (Section 7).

4.3 The Elo baseline construction

Elo is computed internally from provider match history rather than scraped from third-party rating sites, since no clean API exists and scraping is fragile and methodologically opaque. The registered parameters are as follows.

- Standard Elo expected score $W_e = (1 + 10^{-(R_h+H-R_a)/400})^{-1}$.
- $K = 30$ for World Cup matches (finals and qualifiers), $K = 20$ for all other internationals.
- A goal-difference multiplier on K (1 for a draw or one-goal margin, 1.5 for two, 1.75 for three, $+1/8$ per additional goal).
- All teams start at 1500, over a backfill window from June 1, 2022 to tournament start (2,001 fixtures across the 48 qualified teams at computation).

The home adjustment $H = 100$ applies to the listed home team in historical updates and, at tournament prediction time, only when the home team is a host nation playing at home: every other World Cup venue is treated as neutral. Three-way probabilities derive from the Elo expected-score identity $W_e = P(\text{win}) + \frac{1}{2}P(\text{draw})$ with the draw mass set to the empirical draw share of the processed backfill fixtures. The computation is deterministic and rebuilt from scratch daily, so any later replay reproduces the lane's predictions exactly, the property under which the registration permits retroactive computation. Full parameter disclosure is in the public registry (Appendix B). Two known limitations, off-graph opponents entering at 1500 and the listed-home simplification for neutral-venue tournaments in history, are stated there and in Section 10.

4.4 Information flow and isolation

Isolation between lanes is structural, not procedural. The market line reaches the scored path only at scoring time, through a code path that never touches the context packs, so no AI lane can read the odds it is measured against. Official lineups publish roughly 35 minutes before kickoff, after the dispatch window has closed, so no lane ever forecasts with a team sheet in hand. The display-layer artifacts a reader sees on the public dashboard, translations, rendered lineups, and post-match verdicts, are produced off the scored path and are written after the prediction is locked, so nothing shown to the public can flow back into a

scored forecast. The effect is that contamination is prevented by the shape of the system rather than by a reviewer remembering to avoid it: a lane cannot see what the architecture gives it no path to.

5 Metrics and statistical methods

All metrics, their exact operationalization, and the conditions under which any comparative claim may be made were locked with the pre-registration (`statistical-tests.md`, hash-bound, Appendix B). The evaluation code implements exactly these definitions. Unit tests assert known values on synthetic data, and the first scored matchday carried an additional integrity gate described below.

5.1 Scoring a single prediction

The primary metric is the multiclass Brier score over the three-way result (home / draw / away) [2]. Each lane emits a predicted winner and a confidence $c \in [0, 100]$. The lane's probability vector is constructed by the registered rule: the predicted class receives $c/100$, and the remaining mass is split between the other two classes *proportionally to the market's de-vigged priors at T−1 h*. This split rule is declared in the registration precisely to avoid the uniform-split strawman. The market prior never enters any AI lane's own prediction: it is used only at scoring time, identically for every lane. With $\mathbf{p}$ the constructed vector and $\mathbf{o}$ the one-hot outcome,

$$\text{Brier} = \sum_{k \in \{\text{H,D,A}\}} (p_k - o_k)^2,$$

so lower is better, a confident hit beats a timid one, and a confident miss costs the most.

5.2 The two probability spaces

Two constructions coexist, both stored and both labeled wherever they appear. The *multiclass space* of Section 5.1 carries the season Brier score, the locked headline metric. The *binary predicted-winner space* carries everything calibration-shaped: for each scored, non-abstained prediction, $p = c/100$ is the mass the lane put on its predicted winner and $y = 1$ if that winner was correct, else 0. The reliability diagrams, ECE, the Murphy decomposition, and sharpness all live in this binary space, which is exactly the space of the question a reader asks of a confidence number ("when it says 70, is it right 70 percent of the time?"), so the public curve and the reported calibration metrics are internally consistent by construction.

5.3 Reported metrics

Per lane, the benchmark reports the following. Bins, wherever they appear, are ten equal-width intervals on $[0, 1]$, half-open with the top bin closed. n_b is the count in bin b, N the total, conf_b the mean p in the bin, and acc_b the mean y.

Season Brier score (headline). The mean multiclass Brier score of Section 5.1. Perfect is 0, and an always-uniform three-way split scores 0.667.

Murphy decomposition. The binary Brier score partitions over the bins [11] as Brier = REL − RES + UNC, where

$$\mathrm{UNC} = \bar{y}(1-\bar{y}), \qquad \mathrm{RES} = \sum_b \tfrac{n_b}{N}\,(\mathrm{acc}_b - \bar{y})^2, \qquad \mathrm{REL} = \sum_b \tfrac{n_b}{N}\,(\mathrm{conf}_b - \mathrm{acc}_b)^2.$$

Uncertainty is the base-rate variance of the task itself (no forecaster can change it), resolution rewards forecasts that separate wins from losses (higher is better), and reliability is miscalibration proper (lower is better). The identity is exact for the binned forecast. The evaluation code asserts it on every computation and fails loudly on violation, and the raw (per-point) Brier score is reported as the headline alongside the binned decomposition.

ECE. Count-weighted expected calibration error over the same bins:

$$\mathrm{ECE} \;=\; \sum_b \frac{n_b}{N}\,\big|\mathrm{acc}_b - \mathrm{conf}_b\big|.$$

The plain reading: an ECE of 0.10 means that, averaged over its predictions, the lane's stated confidence and its realized hit rate disagree by ten percentage points. Zero is perfect calibration.

Sharpness. Implemented as the variance of the forecast probabilities in the binary space, $\mathrm{mean}\big((p - \bar{p})^2\big)$: a lane that always says 55 has near-zero sharpness regardless of its calibration, and sharpness is what separates a well-calibrated coward from a well-calibrated forecaster. Disclosed deviation: the registered text phrased sharpness as mean distance from the uniform prior, but the implementation measures spread around the lane's own mean forecast. Both capture "willingness to commit". The implemented form is the one reported, and the registered phrasing is recoverable from the published per-prediction dataset.

Bootstrap intervals. 95 % confidence intervals from 1,000 resamples with the match as the resampling block. The same resampled match-set is applied to every lane within an iteration, preserving cross-lane correlation so per-lane intervals are comparable, and the generator is seeded with a fixed, published constant so every interval in this paper reproduces bit-for-bit on re-run.

Hit rate. Fraction of correct predicted winners. Context only, never a headline: a bold guesser can post a high hit rate while being badly calibrated, which is the confusion this benchmark exists to separate.

Stage-stratified Brier score. The registered strata are group, round of 32, round of 16, quarter-finals, semi-finals with the third-place match, and the final. Disclosed implementation note: the published pipeline collapses the knockout strata into a single group-versus-knockout split because per-stage counts are too thin for stable intervals. The finer registered strata remain computable from the published per-prediction dataset.

A reliability diagram accompanies the ECE results (Section 6). Abstentions are reported per lane (picks made / refused) so that Brier score comparisons are apples-to-apples. A reader-facing reference defining every lane and every metric in plain terms is provided in Appendix E.

5.4 Pre-committed claim conditions

Three conditions, fixed before any data existed, govern what this paper is allowed to claim. **(i)** A lane is "better calibrated" than another only if its held-out Brier score confidence interval and the comparison lane's do not overlap. Overlapping intervals are reported as no resolvable difference. **(ii)** The Level-2 learning claim requires *all* of: the looped lane's held-out Brier score better than its control's with non-overlapping intervals, *or* a seeded-pattern pickup demonstrably routed through the learning-loop layer, *and* at least one concrete behavior-change instance (the same pattern, pre-loop versus post-loop decisions differing) documented in this paper. **(iii)** An abstention asymmetry exceeding 10 percentage points between compared lanes disqualifies the direct Brier score comparison. The asymmetry is reported instead. Additionally, no comparison is published below $N = 30$ scored matches per lane, and development-window results are labeled as development data wherever they appear.

5.5 Evaluation integrity

Two integrity mechanisms are part of the registered design. First, on the first scored matchday, the Brier score was hand-computed independently for the required five outputs and cross-checked against the production pipeline. The check was executed on matchday one (June 11, 2026) across all eight lanes of the opening match, with agreement at three decimals on every output (Appendix B). Second, for the narrative-quality judging of M2, inter-judge agreement is reported as mean pairwise Krippendorff's α. Reported, not repaired: disagreement between judges is data, not noise to be averaged away silently.

All reported values for these metrics are generated by the published evaluation pipeline and inserted from `data/`, never hand-typed. The held-out and full-tournament leaderboards, the Brier decomposition, the stage-stratified results, and the reliability diagram appear with the results (Section 6).

6 Results

Reading rule. The registration fixes a deliberately strict test for a single pairwise claim on the held-out window: two lanes are called "resolvably different" there only when their held-out Brier score bootstrap intervals do not overlap. On the twenty-match held-out set every pair overlaps, which is what a sample that small returns on a chance-heavy task, so we make no single-pairwise significance claim on the held-out window and read its ordering as directional. The full 104-match tournament, where the sample is an order of magnitude larger, is where the calibration ordering is read: the point ordering is clear there, the council lowest on ECE of every machine, though its bootstrap interval still overlaps the others', so we present that lead as directional rather than as a resolved pairwise separation. We report both windows, are explicit throughout about which window a given statement rests on, and hold every ordering at the strength its intervals license.

Simplifications established during the window (structural, not outcome-dependent). No lane abstained on any match, so all Brier score comparisons are apples-to-apples by construction and the registered abstention-asymmetry disqualifier never triggers. The dev-window clause (§5 of the registration) was never needed: the OSF lock preceded the first kickoff, so every scored match is a scored-window match and no dev-window stratum exists. The held-out set (the final 20 matches by kickoff) is entirely knockout-stage, a consequence of the M6 freeze point falling at the group/ knockout boundary. We note this where it bears on interpretation.

6.1 Headline calibration (held-out set)

Table 3 reports every lane on the held-out set: multiclass Brier score with its block-bootstrap 95 % interval, ECE, sharpness, hit rate, and cost.

Lane	n	Brier score	95% CI	ECE	ECE 95% CI	Sharp.	Hit	Cost
Solo	20	0.422	[0.299, 0.536]	0.192	[0.110, 0.361]	0.008	75%	$8.22
Solo-Zero	20	0.375	[0.239, 0.519]	0.181	[0.071, 0.341]	0.012	80%	$1.06
Pipeline	20	0.430	[0.314, 0.549]	0.200	[0.125, 0.377]	0.010	75%	$21.93
Pipeline-Static	20	0.431	[0.316, 0.548]	0.202	[0.124, 0.377]	0.009	75%	$21.93
Council	20	0.425	[0.309, 0.546]	0.196	[0.120, 0.369]	0.011	75%	$15.03
Elo	20	0.503	[0.416, 0.581]	0.309	[0.190, 0.447]	0.006	65%	n/a
Market	20	0.457	[0.336, 0.578]	0.166	[0.119, 0.341]	0.019	65%	n/a
Human	20	0.483	[0.317, 0.655]	0.127	[0.056, 0.325]	0.006	65%	n/a

Table 3: Held-out results, the final 20 matches (Section 3, M6). Intervals are 1000-resample block bootstraps (block = match, all lanes resampled together). The ECE interval uses the same seeded resample recomputing binned ECE. All intervals overlap, so the ordering is suggestive, not resolved. Cost is the lane's cumulative API spend over the full tournament, not per match. Per-match economics and the marginal cost figure are in Appendix C.

Fact. On the held-out set the context-free twin Solo-Zero records the lowest multiclass Brier score (0.375), ahead of every full-context AI lane, the deterministic Elo baseline, and the market's de-vigged closing line. All five AI lanes record a lower held-out Brier score than Market. Every lane's 95 % interval overlaps every other lane's.

Reading the two twin pairs (they are not the same kind of pair). Two pairs of lanes are twins by design, and they must be read differently. Solo and Solo-Zero are a genuine ablation: the same single-call architecture, with and without the context stack, so their gap measures the context stack. Pipeline and Pipeline-Static are *not* an ablation in this run. The learning loop that was meant to separate them was never populated in production (Section 7), so the two lanes received *identical* inputs on every match, and any gap between them is run-to-run noise, not a learning effect. We report their two numbers because they were scored, but the near-identity of Pipeline and Pipeline-Static on the held-out set is exactly what a same-system pair should show, and it bounds the noise floor against which every other gap in this table must be read. It carries no information about learning, and we make no learning claim from it.

Interpretation. The registered prior stated that no AI lane would beat the closing line on raw accuracy. On the held-out set, measured by Brier score, the ordering runs the other way: the AI lanes sit below the market line, and the leanest of them, a single call with no context stack, sits lowest of all. We report this as the pre-stated question resolving against the prior in direction. We do *not* claim any lane is significantly better than the market or than another lane: the intervals overlap, so the honest statement is that no elaboration of architecture, and no addition of the context stack, produced a Brier score improvement that survives the sample. The economic contrast is the sharper reading of the accuracy result: the lowest-Brier score lane also cost roughly a twentieth of the pipeline lanes.

Accuracy was the wrong question, by design. The paragraphs above report Brier score because it is the locked headline metric, but the benchmark was built to compare calibration, not accuracy, and on a

chance-dominated task that distinction is the whole point. No architecture owns a coin-flip outcome, so the held-out accuracy ranking is dominated by match noise and, as the intervals show, resolves nothing. The question that a business actually asks of a forecast is different and is answered by calibration: when the system states a confidence, can that number be trusted? A forecast that says thirty percent and is wrong has reported its uncertainty honestly. A forecast that says seventy percent and is wrong has not. The next subsection reads the lanes on that axis.

6.2 Calibration: which architecture's confidence can be trusted

Calibration is the axis the study was built to measure, and it is where the architectures give an operator something to act on. Over all 104 matches (Table 7), the COUNCIL architecture records the lowest ECE of every machine tested (0.100), below the single-call and pipeline architectures, the deterministic ELO baseline, and the market's own closing line. Its stated confidence tracks realized frequency more closely than any other machine's, which is the property an operator consumes when an AI-supplied probability enters a real decision. On the operational question the study exists to answer, which architecture reports the most trustworthy confidence, the evidence points to the council.

We report the strength of that lead precisely. Once each ECE carries its bootstrap interval, the council's, $[0.076, 0.216]$, overlaps the market's $[0.066, 0.209]$ and the other AI lanes', so the ordering is directional rather than resolved: a real point lead inside overlapping intervals that a second tournament would sharpen. This is the honest limit on the margin, and it does not unseat the recommendation. A directional best is still a best, and an organization choosing an architecture to trust with a forecast is choosing under exactly this kind of evidence, a clear point ordering with intervals that a single event cannot yet separate. The overlap tells it how firmly to hold the choice, not to abandon it. What the operator learns is both the quantity to demand, calibration, and the shape that reported it best here, the council.

The reliability diagram (Figure 1) shows the shape behind the numbers. A perfectly calibrated forecaster's points lie on the diagonal, where stated confidence equals observed frequency. A point below the diagonal is overconfidence, the failure mode that costs a business the most, and a point above it is caution. The curves are visibly noisy, which is the honest picture of what a hundred-odd binned matches support and the reason the ECE intervals overlap and the council's lead reads as directional rather than resolved. The value of the diagram is not to declare a winner by eye but to show that no lane exhibits a gross, systematic overconfidence bias: the points scatter around the diagonal rather than sitting consistently below it. The ECE lead the council holds above lives inside this scatter, which is exactly why the paper treats it as directional and reports every interval alongside every point.

6.3 The two pre-stated hypotheses

Two patterns were logged as hypotheses on 2026-07-03, while 20 of the held-out matches were still unplayed, so their held-out verdict is a stated-in-advance question rather than a post-hoc story:

(a) The context-free twin matches or beats the full-context lanes on Brier score. *Verdict: supported in direction, not resolvable in magnitude.* On the held-out set SOLO-ZERO records a lower Brier score than its full-context twin SOLO and than every other AI lane. The direction matches the hypothesis, and the marginal calibration value of the context stack for a single-call architecture is, at best, not positive on this evidence. The intervals overlap, so we stop short of claiming the context stack *hurt*. We claim only that it did not measurably help, at a cost of roughly eight times the context-free call.

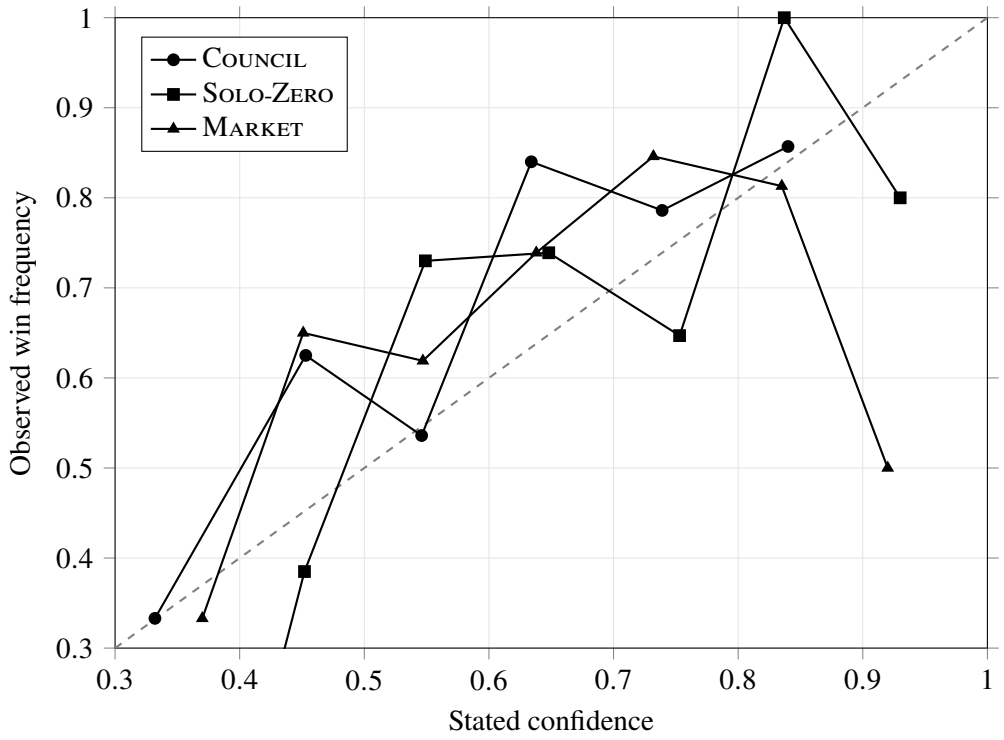

Figure 1: Reliability diagram, full tournament. Each point is a confidence bin: the horizontal axis is the mean stated confidence, the vertical axis the observed win frequency in that bin. The dashed diagonal is perfect calibration. Points below the diagonal are overconfident. Full-tournament (development plus held-out) data, shown because the per-bin counts on the held-out set alone are too thin to plot stably.

(b) Council leads on ECE while not leading on Brier score. *Verdict: not confirmed on the held-out window.* This hypothesis was pre-stated about the twenty-match held-out set specifically, and that window is too small to confirm it: among the lanes blind to the other lanes' outputs the lowest held-out ECE belongs to Solo-Zero rather than Council, the lowest overall belongs to Human (a non-blind reference, Section 3, carrying no comparative claim), and none of these held-out ECE gaps survives the interval test in any case. So on the held-out window alone the hypothesis is not borne out. Nor is it resolved on the full tournament, where the sample is an order of magnitude larger, the council holds the lowest *point* ECE of every machine, though its interval overlaps the others' (Section 6.2), so the lead there is directional rather than resolved. The honest verdict is therefore split by window: the strict twenty-match test does not confirm the hypothesis, while the full tournament supports it in direction, with the council posting the best calibration of any machine but by a margin a single event cannot yet separate from noise. We report both, and we read the full-tournament ordering as the directional answer to the operational question, the council is the best-calibrated architecture here, while holding it at the strength the intervals license and no further.

6.4 Controls: the price of context and the cost of calibration

The registered control pair isolates the context stack: Solo and Solo-Zero are the same single-call architecture, one with the full context pack and one with none. On the held-out set the context-free member records the lower Brier score (Table 3) at a small fraction of the cost. Read with the claim condition,

this is a null result on the value of context for a single call, not a demonstration that context is harmful. It is, however, a demonstration that the context stack's cost bought no measurable calibration gain here. The Brier decomposition (Table 4) locates where the lanes' scores come from: reliability (miscalibration, lower is better) and resolution (discrimination, higher is better) are reported per lane on the held-out set.

Lane	Reliability	Resolution	Uncertainty
Solo	0.047	0.033	0.188
Solo-Zero	0.035	0.010	0.160
Pipeline	0.054	0.034	0.188
Pipeline-Static	0.054	0.034	0.188
Council	0.051	0.034	0.188
Elo	0.130	0.123	0.227
Market	0.038	0.073	0.227
Human	0.021	0.034	0.227

Table 4: Murphy decomposition of the binary Brier score on the held-out set (Brier = reliability − resolution + uncertainty). In the predicted-winner space each lane's uncertainty is $\bar{y}(1-\bar{y})$ on its own hit rate $\bar{y}$, so a lane that is right more often carries a lower uncertainty term. It is the part of the score no calibration can change, not a lane's skill.

6.5 Escalation: the Pipeline–Council entanglement

The Pipeline gate escalates a low-confidence match to Council, and the published Pipeline output on an escalated match *is* Council's. The two lanes are therefore not independent. The gate fired on 15 of the 20 held-out matches (75 %) and 65 of the full tournament (63 %), so the entanglement is not a footnote. We report Pipeline three ways (Table 5) to make it measurable: as published (the mix that scored), pre-escalation (the pipeline's own output, from the logged pre-escalation prediction), and the adopted-Council subset alone.

Pipeline view	Held-out Brier score	Full Brier score
As published (Council where escalated)	0.430	0.481
Pre-escalation (Pipeline's own output)	0.429	0.494
Adopted-Council subset only	0.518 (n = 15)	0.546 (n = 65)

Table 5: The escalation triple-report, multiclass Brier score. "As published" is what scored. "Pre-escalation" is what Pipeline would have scored on its own output. The "adopted-Council subset" is the escalated matches alone, which the gate selected precisely because they were low-confidence, so a higher Brier score on that subset reflects match difficulty, not a lane defect.

Interpretation. The gate routed the matches it judged hardest to Council, so the adopted subset carries the harder matches by construction and scores worse than the tournament average on both sets. The as-published and pre-escalation columns differ modestly, which bounds how much the escalation machinery changed Pipeline's scored calibration: on the held-out set the two are within rounding, so escalation neither helped nor hurt the pipeline's held-out Brier score to any resolvable degree.

6.6 Stage-stratified results and the judging layer

Because the held-out set is entirely knockout, the group-versus-knockout contrast is visible only across the full tournament, reported here as development context (Table 6). The narrative-quality judging (M2) is a separate layer from the calibration results above: it scores reasoning text, not predictions, and its inter-judge agreement fell below the registered threshold. That result, and its consequences for what the M2 scores can support, is reported with the other measurement limitations in Section 10 rather than here, so that the calibration findings, which are scored against real outcomes, are not read through the judging layer's disagreement.

	Group		Knockout	
Lane	Brier	*n*	Brier	*n*
Solo	0.212	72	0.204	32
Solo-Zero	0.211	72	0.187	32
Pipeline	0.224	72	0.207	32
Pipeline-Static	0.221	72	0.207	32
Council	0.222	72	0.204	32
Elo	0.243	72	0.230	32
Market	0.220	72	0.194	32
Human	0.221	72	0.210	32

Table 6: Stage-stratified binary Brier score, full tournament (development context). The held-out set is entirely knockout, so this split is not a held-out result and carries no headline claim.

6.7 Full-tournament results, the calibration evidence base

Table 7 reports every lane over the full tournament, all 104 matches. This is the evidence base for the calibration finding of Section 6.2: the registration fixes the held-out set as a pre-committed stress window, and the full event is the larger sample against which the calibration point ordering is read. On accuracy the two windows tell the same story in their extremes, the context-free lane lowest on Brier score and the deterministic baseline highest, though the market sits mid-pack over the full tournament rather than behind the AI lanes as it does on the held-out knockout stretch, a difference we attribute to the knockout-only composition of the held-out set rather than to a change in what the systems do. On calibration the full tournament carries the point ordering of Section 6.2, the council lowest on ECE of every machine. Its bootstrap interval overlaps the others', so the lead is directional rather than resolved on this larger window, which is the strength we claim for it and no more.

7 The learning-loop test

7.1 What was registered

The registered Level-2 question asked whether an architecture with an active learning loop demonstrates genuine learning rather than retrieval, measured by (a) the held-out gap between Pipeline and its loop-disabled twin Pipeline-Static under the pre-committed claim conditions (Section 5.4), and (b) pickup of two sealed synthetic patterns injected into the context stack from match 1.

Lane	*n*	Brier score	95% CI	ECE	ECE 95% CI	Sharp.	Hit	Cost
Solo	104	0.472	[0.409, 0.545]	0.114	[0.074, 0.214]	0.015	66%	$8.22
Solo-Zero	104	0.453	[0.379, 0.532]	0.135	[0.085, 0.219]	0.018	68%	$1.06
Pipeline	104	0.481	[0.416, 0.555]	0.112	[0.084, 0.222]	0.014	67%	$21.93
Pipeline-Static	104	0.477	[0.414, 0.550]	0.136	[0.097, 0.237]	0.014	67%	$21.93
Council	104	0.477	[0.411, 0.553]	0.100	[0.076, 0.216]	0.017	67%	$15.03
Elo	104	0.534	[0.488, 0.584]	0.114	[0.067, 0.212]	0.007	62%	n/a
Market	104	0.468	[0.394, 0.553]	0.104	[0.066, 0.209]	0.023	68%	n/a
Human	104	0.505	[0.424, 0.597]	0.074	[0.042, 0.170]	0.010	62%	n/a

Table 7: Full-tournament results, all 104 matches. This is the evidence base for the calibration reading (Section 6.2). The council posts the lowest ECE of every machine. Every ECE also carries its bootstrap interval, and these overlap, so the lead is directional rather than resolved. The held-out set (Table 3) is the pre-registered stress window, reported alongside. Cost is cumulative tournament API spend per lane, not per match (Appendix C).

7.2 What actually ran

The seeded-pattern injection (M5) executed as registered from match 1 with deterministic location rotation. The loop-write side did not: the evaluation component that populates the learning-loop layer was not deployed during the tournament, so the L7 context layer was empty for every match, for both lanes of the control pair. Two consequences follow by construction, and we state them plainly:

- Pipeline and Pipeline-Static received *identical* inputs on every match. The pair therefore functions as a same-architecture *replication control*, an unplanned but usable estimate of run-to-run variance under identical conditions, and any observed gap between the twins bounds the noise floor against which architecture-level claims elsewhere in this paper must be read. It carries no information about learning.
- Seeded-pattern pickup *through the loop layer* is impossible by construction. Pattern-pickup analysis is accordingly limited to detection in output text (did any lane's reasoning surface the seeded patterns from the injected context?), which tests context attention, not learning.

7.3 The Level-2 verdict this design change allows

The pre-committed claim conditions required a held-out twin gap with non-overlapping intervals *or* pickup demonstrably routed through the loop layer, *and* a concrete behavior-change instance. With the loop never active, neither routing evidence nor behavior change is observable: the registered Level-2 claim is reported as **not testable as designed**, not as a null result. The distinction matters: a null result would say "the loop did not help." The honest statement is "the loop was never exercised, so this benchmark cannot say."

7.4 The replication gap, and what it bounds

Because the twins received identical inputs, the gap between them is a direct read on run-to-run variance, the noise floor of the benchmark. On the held-out set Pipeline and Pipeline-Static record multiclass Brier score values that differ by roughly one thousandth (Table 3), and over the full tournament by roughly four thousandths (Table 7). Both gaps are an order of magnitude smaller than any architecture-level

difference elsewhere in the study, and both sit far inside the overlapping bootstrap intervals. This is the quantitative reason the paper claims no resolved architecture difference: the identical-twin gap shows that a difference of a few thousandths in Brier score is what two runs of the *same* system produce, so a gap of that size between two *different* systems carries no signal. The replication control was not the one we registered, but it earns its place by calibrating how much of any observed gap is noise.

7.5 Why the loop did not ship, and what a corrected re-run needs

The absence is an engineering-scope fact, not a silent protocol change. The component that writes learning-loop entries into the L7 context layer was not deployed during the tournament, so L7 was null for every match and for both lanes of the control pair. No prompt, rubric weight, model pin, context-pack construction rule, or scoring rule was altered: the registered design was run as written, with one registered layer left unpopulated because the writer behind it did not ship. We report this in the open rather than narrow the claim quietly around it. The seeded-pattern injection (M5) did run from match one as registered, but with no loop entries the only observable form of pickup is in output text, which tests context attention rather than learning through the loop. We report the injection as executed and leave the output-text detection sweep to the open dataset rather than assert a learning result the design cannot support. A corrected successor benchmark would need the loop writer active from match one, so that L7 carries genuine prior-match entries and the twin pair becomes the ablation it was meant to be. That is the single change that would move the Level-2 question from *not testable* to answered.

8 Discussion

The introduction argued that the operative question for an organization adopting AI is not whether a system is right slightly more often than an alternative, but whether the confidence it reports can be trusted. The results let us restate that claim with evidence, and the evidence points in one direction on the axis that matters and in no direction at all on the axis that does not.

8.1 What the calibration result means for delegating to AI

On accuracy, the study resolves nothing, and that is the correct outcome for a chance-dominated task. No architecture reliably called more matches than another, the held-out intervals overlap without exception, and even the inversion of our stated prior, every AI lane below the market's closing line on held-out Brier score, is a directional reading we decline to inflate into a significance claim. An organization that chose an AI forecasting architecture on its win rate would be choosing on noise.

Calibration behaves differently, and it is the axis the study was built to isolate. Across the full tournament the council reported the most trustworthy confidence of any machine: its stated probabilities tracked realized frequencies more closely than the single-call and pipeline architectures, the deterministic baseline, and the market line itself, and it posts the lowest ECE of the eight lanes, the human reference aside. This is the result an adopter should act on, because calibration is the property an adopter actually consumes. A weighted pipeline forecast, a reserve set against a risk-adjusted number, a decision to auto-approve above a confidence threshold and escalate below it, each of these inherits the quality of the confidence estimate, not the raw hit rate. A system whose seventy percent means seventy percent can be delegated a class of decisions that a better-guessing but overconfident system cannot, and across this tournament the council is the system that earned that trust.

We state the strength of the reading exactly. It rests on the full tournament, 104 live matches scored by a pre-registered pipeline, which is a substantial basis for the direction it gives. On a chance-dominated event the bootstrap intervals on ECE overlap, so the council's lead is directional rather than a resolved separation, and we present it as such: a strong directional result for anyone choosing an architecture to trust, which a second event would sharpen. The overlap is the honest limit on the margin, not a retraction of the recommendation. We name the tightest comparison plainly: the council's calibration error, 0.100, sits four thousandths below the market's 0.104, and our own control (Section 7) puts the run-to-run noise floor at about the same magnitude. Against the market the council's lead is therefore a tie at the resolution this single tournament can measure, and we do not claim it beats the market. That comparison, however, is the one that does not transfer. The reason the machines cluster near the market is that football has a deep public betting market already pricing every fixture, a hard reference that leaves little room to separate. A business forecasting its own pipeline has no such market and no closing line. There is no incumbent probability to tie against, only the company's own slow actuals, so the calibrated AI forecast is not squeezed against a reference, it is the reference the company would otherwise lack. The council's lead over the other machines, the ones a company would actually be choosing between, is where the finding lands, and it is a lead the market comparison neither creates nor takes away.

8.2 Architecture economics: calibration is not bought with spend

The cost record sharpens the finding rather than complicating it. The context-free single call, the cheapest system in the study by more than an order of magnitude, was no worse calibrated than the elaborate pipelines and, on the held-out set, posted the lowest Brier score of all. Whatever the context stack and the sequential specialist roles added, calibration was not it, and their cost was real. The council, more expensive than a single call but far cheaper than the pipelines, is the architecture that turned its extra spend into the study's best calibration, while the pipelines spent more and bought no calibration edge over the cheapest call, which is the cost reading in its sharpest form. The operating implication is that spend and trustworthiness are not the same axis. An organization can pay for orchestration it does not need, and the way to tell is to measure calibration per dollar rather than to assume that a more elaborate pipeline is a more trustworthy one. This benchmark measured it, and elaboration did not pay on this task. Costs throughout are cumulative tournament API spend per lane, with the per-match economics detailed in Appendix C.

8.3 What the controls taught

The two twin pairs isolate two questions. The genuine ablation, the single call with and without the context pack, returned a null on the value of context for this architecture: the context-free twin held its own on calibration and led on accuracy, at a fraction of the cost. We resist the stronger reading that context is worthless, because the intervals do not license it, but the evidence gives an organization no reason to assume a rich context stack improves a single-call system's calibration, and a clear reason to price it before adopting it. The second pair was not an ablation in this run: the learning loop that was meant to separate the pipeline from its static twin was never populated (Section 7), so the two received identical inputs and their near-identity is a measurement of run-to-run noise, not a finding about learning. It bounds the noise floor against which every other gap in the study must be read, and it is the reason none of those gaps is claimed as resolved.

The reading holds here: on a task built to test whether calibration rather than accuracy is the axis to measure, calibration is where the architectures separate, the council reporting the most trustworthy

confidence of any machine while the cheapest single call is calibrated as well as the most expensive pipeline. The margin is directional on a chance-dominated event, and we say so, but the operational answer is clear enough to act on: measure calibration, and on this evidence the council is the shape to trust. The applied translation of that finding, how an operator maps this benchmark to a business forecast such as a sales pipeline or a demand plan and what procurement questions follow, is the subject of the next section.

9 From benchmark to business: choosing a forecasting architecture

Football matches are the testbed. The class of problems this benchmark actually addresses is any recurring business forecast with four properties: the decision repeats on a schedule, the outcome is genuinely uncertain, a hard deadline forces the forecast to lock before the truth arrives, and a verifiable result lands soon enough to score it. Sales pipelines, demand plans, cash-flow projections, churn predictions, and delivery-date commitments all have this shape. What changes between them is the nouns, not the method.

9.1 Why calibration is a business number

The clearest instance is the weighted sales pipeline. Multiplying deal size by close probability and summing is standard revenue arithmetic, and it is *only* arithmetic if the probabilities are calibrated: a book of deals marked "70 percent" must close about 70 percent of the time or the weighted total is fiction with decimal places. When an AI system starts supplying those probabilities, the question this benchmark asks of a World Cup lane becomes a finance question verbatim: when the system says 70, is it right 70 percent of the time? ECE is that question as a single number (Appendix E.2), and a reliability curve is the audit trail behind it. An operator who adopts AI-generated confidence without measuring calibration has not automated forecasting. They have automated the production of unaudited numbers.

A finance reader will object that a company already has a validator its forecast can be checked against, last period's actuals, and that objection is correct as far as it goes. A quarterly backtest against realized revenue is a real calibration check, and where a company has enough closed history it should run one. What calibration adds is what the backtest cannot give. The World Cup had a public liquid reference forecast in the betting market, and a company does not, so a company's only ground truth is its own actuals, which arrive slowly and in aggregate. Calibration is the signal available before those actuals accrue, when a new segment, a new product, or a young company has too few closed deals to backtest (the N floor this study respects), and it is a *per-forecast* signal a quarterly backtest cannot produce: actuals tell an operator that last quarter's number missed, while a reliability curve tells them *which* confidence bands are the ones that lie. The claim is not that a company cannot validate its forecasts. It is that calibration is the honesty check it can run before it has enough actuals, and at a resolution its actuals never reach.

9.2 The mapping

Two rows of Table 8 deserve emphasis. First, the incumbent process plays the role of the closing line: it is the benchmark to beat, it embeds information the models do not have, and the registered prior of this study (expect the incumbent aggregate to win on raw accuracy) is the honest starting posture for any AI forecasting initiative. The realistic near-term goal is not out-predicting the incumbent. It is matching it with better-calibrated, cheaper, auditable confidence. Second, the Solo-Zero row is the cheapest experiment most organizations never run: the same model with the context withheld. If the context-free

Benchmark element	Business equivalent
A match	One forecast event (a deal, a SKU-month, an invoice)
Kickoff lock (T−75 min)	The forecast-commit deadline
The 90-minute target	The locked outcome definition (closed-won by date X)
MARKET (closing odds)	The incumbent forecast (reps, finance, the current process)
ELO	The cheap deterministic baseline (historical base rates)
HUMAN	The experienced operator's instinct call
The context pack	CRM fields, call notes, email signals, external data
SOLO-ZERO (the twin)	"Does all that context actually help?", measured
The escalation gate	Low-confidence forecasts routed to human review
Abstention	The registered right to say "no forecast" without penalty
$N = 30$ floor, held-out set	How much history before the scorecard is trusted

Table 8: The translation table. Every methodological element of the benchmark has a direct operational equivalent. Nothing in the method is specific to football.

twin performs comparably, the data integration was cost, not signal, and that is worth knowing before the pipeline is built, not after.

9.3 A worked use case: the quarterly sales forecast

Consider a company forecasting quarterly revenue from a pipeline of deals. The benchmark's method transfers as follows. The *event* is each open deal above a materiality threshold, scored on the locked question "closed-won by quarter end, yes or no" (the three-way match target becomes a two-way outcome, and every metric in Section 5 has a binary form, the calibration machinery being binary already). The *lock* is the monthly forecast-commit date: the AI's probability per deal is recorded before the commit meeting and never revised, exactly as predictions lock before kickoff. The *lanes* are the build options: a single-call generalist reading the deal's CRM record and communications trail (the SOLO pattern), a staged pipeline in which a quantitative step reads deal metrics, a contextual step reads the communications, and a synthesis step commits (the PIPELINE pattern, with its registered escalation gate routing any deal below 60 confidence to the sales manager), and a multi-model committee for the calibration-critical book (the COUNCIL pattern). The *controls* are the same twins: a context-free run pricing the CRM integration, and, if a learning loop is claimed, a loop-disabled twin, because Section 7 of this study is the cautionary tale of claiming a learning loop without instrumenting proof that it runs. The *baselines* are the rep's own call, the historical close rate by stage (the deterministic floor), and the incumbent weighted pipeline. After one quarter, a company of ordinary deal volume clears the $N = 30$ floor and owns its first reliability curve: which is to say, it knows whether its AI's 70 means 70.

The same skeleton re-nouns to the neighboring problems: demand planning (event = SKU-month, outcome = demand within tolerance), cash-flow (event = invoice, outcome = paid by date), churn (event = account-quarter), and delivery commitments (event = order, outcome = on-time). The method is indifferent to the noun.

9.4 Choosing the architecture

The tournament turns the five architectures into a priced menu an operator can choose from. Read against cost, the calibration results sort into three tiers.

The cost-calibration frontier. Plotting each lane's calibration error against its tournament cost places two lanes on the efficient frontier and the rest above it. The context-free single call (SOLO-ZERO) sits at one dollar and change ($1.06) with a calibration error of 0.135, the cheapest system that forecasts at all. The council sits at $15.03 with the lowest calibration error of any machine (0.100). Everything between is dominated: the full-context single call (SOLO, $8.22) is no better calibrated than the council and costs half as much only by being worse, and the two pipelines ($21.93 each) are the most expensive systems in the study and buy nothing for the premium, one of them posting the worst machine calibration of all (PIPELINE-STATIC, 0.136). An operator prices architectures off this frontier. The question is never which system is most elaborate, it is which point on the cost-calibration curve the decision justifies.

Context spend must be validated per task. The single clean ablation in the study, SOLO against SOLO-ZERO, is the same architecture with and without the seven-layer context pack. The context-free twin was no worse calibrated at one eighth the cost. The operating lesson is not that context is worthless, it is that a context stack is a cost that must be shown to pay before the integration budget is committed, because on this task it did not. A vendor proposing a rich retrieval-and-enrichment layer should be asked for the calibration difference it buys, measured, not assumed.

When the council's premium is worth paying. The council costs roughly fourteen times a context-free call and returns the study's best calibration. That premium is worth paying when the decision is calibration-critical, when a confidence number drives an irreversible or expensive commitment such as a credit line, a capacity reservation, or revenue guidance to a board. For those the council is the architecture to reach for, holding the directional caveat this paper states throughout. For a high-volume, low-stakes forecast where a rough probability suffices, the context-free call is the rational floor, and the money saved is real. The council is a recommendation scoped to the decisions that can afford it, not a default for every forecast.

An escalation gate is an architectural decision, not a safety afterthought. The PIPELINE lane routed its self-judged-hardest matches to a heavier review step, and that gate is entangled with its results. It is a design choice with measurable cost and calibration consequences, not a bolt-on. An operator composing a multi-stage system should treat every routing and escalation rule as a lane in its own right, to be scored, because it changes the number the business consumes.

The emergent architecture. The study's own components suggest a system it did not run: a context-free base call for cost, a council step reserved for the calibration-critical subset, and an escalation gate scored rather than assumed. What that assembly would cost per calibration point against a company's incumbent process is the measurement a first internal tournament should be built to return. This paper gives the method to run it, not the finished blueprint.

9.5 The operator's checklist

Six questions, verbatim, for any vendor or internal team proposing an AI forecasting system. Each is answerable from this paper's method, and a team that cannot answer them has not measured what it is selling.

1. "Show me the reliability curve for your system's predictions at $N \geq 30$ scored events. Not accuracy. The curve."
2. "When your system reports 70 percent confidence, what does it measure to? What is its ECE, over how many events, and on what held-out window?"
3. "What is the confidence threshold below which a forecast routes to a human, and where is that logged?"
4. "What did the context integration measurably add? Show me the context-free ablation."
5. "Can you replay any past forecast bit-for-bit, with the inputs it saw at lock time? What prevents post-hoc revision?"
6. "What does one forecast cost, and what is the cost per calibration point gained over our current process?"

9.6 Running it on your own pipeline

The benchmark is a repeatable method, and its minimum viable form fits inside one business quarter. Lock the outcome definition and the commit deadline in writing before the first forecast. Record every prediction with its confidence before the deadline, with no revision. Carry the incumbent process and one deterministic base-rate as baselines. Run the context-free twin. Score nothing publicly below $N = 30$ events. Hold out the final stretch for the claims that matter. That is the whole protocol, transplanted. The instruments are the ones defined in this paper (Section 5, Appendix E), the cost of running them is small against the cost of one quarter of miscalibrated commitments, and the result is the thing this study argues no operator should delegate without: a measured answer to whether the number the machine reports deserves the trust the arithmetic quietly places in it.

10 Limitations

- **Architecture–model confound.** SOLO and PIPELINE run on one model family while COUNCIL mixes three, by design. Claims concern the pinned systems as specified, never architectures in the abstract.
- **Single tournament, chance-dominated domain.** Generalization beyond live, chance-heavy forecasting tasks is argued, not demonstrated.
- **Non-blind human reference.** HUMAN is not a controlled baseline. The forecaster was the project operator and was not blind to the AI lanes' predictions or the running results, so any calibration advantage the human shows is confounded by information access and carries no human-versus-machine claim (the full disclosure is in Section 3). It is one person, not a panel: the registered option to add a named-writer panel mid-tournament (M7) was not exercised, so no writer picks were collected.
- **Elo construction simplifications.** Off-graph opponents enter at 1500. The listed-home adjustment is applied to neutral-venue tournaments in the historical window (disclosed in the registry).
- **What we did not replicate.** The strong published baselines we cite (e.g., the ordered-logit of 12) are comparison context, not reimplementations.

- **[2026-07-03] Learning loop never populated.** The evaluation component that writes the learning-loop layer did not ship during the tournament, and L7 was empty for both twins on every match. The registered Level-2 claim is reported as not testable as designed (Section 7), and the twin pair is repurposed as a replication control. This is the largest divergence between registration and execution in the study.
- **[2026-07-03] Pipeline–Council entanglement at majority rate.** The registered escalation gate (final confidence below 60 adopts the Council's prediction) fired on a majority of matches, so PIPELINE and COUNCIL are strongly non-independent as published lanes. Per-match pre-escalation outputs were logged throughout and the entanglement is analyzable (Section 6). Lane comparisons involving PIPELINE must be read with this coupling in mind.
- **[2026-07-03] Narrative-quality evaluation (M2) executed post-hoc.** No judge scores were produced live. The three-judge evaluation ran on 2026-07-03 from the frozen outputs (provenance-safe on three counts. Judging is evaluation, not prediction. Judges were blinded to lane, model, and result. The judge prompt is dated post-hoc tooling, disclosed, outside the M1 lock). Inter-judge agreement did not merely fall below the registered 0.6 threshold, it fell below zero: the mean pairwise Krippendorff's α is negative (Table 9), meaning the three judges agreed on absolute quality less than random assignment would. The mechanism is a scale effect visible in the per-lane matrix (Table 10): the judges rank the lanes similarly at the extremes but score on very different absolute scales, one clustering near six and another near nine, so an interval-metric agreement statistic reads the disagreement as severe. We report this as a finding, not a defect to repair. For the narrative-quality task, a panel of language-model judges did not produce a reliable absolute score, and the M2 qualitative numbers are read accordingly, as indicative of gross ordering at most and not as a calibrated quality measure. The result does not touch the calibration findings, which are scored against real match outcomes rather than judged.

Dimension	Claude/GPT	Claude/Gemini	GPT/Gemini	Mean
reasoningTraceability	0.017	-0.512	-0.214	-0.236
brandDiscipline	-0.320	-0.503	0.027	-0.265
operationalClarity	-0.175	-0.568	-0.123	-0.289
Mean pairwise α (all dimensions)				-0.263

Table 9: Inter-judge agreement (M2), mean pairwise Krippendorff's α per rubric dimension. Negative values indicate agreement worse than chance on absolute severity.

Judge	SOLO	SOLO-ZERO	PIPELINE	PIPELINE-STATIC	COUNCIL
claude	6.81	5.44	6.70	6.62	6.26
gpt	7.80	7.69	7.92	7.96	7.74
gemini	8.86	8.48	9.12	9.18	9.04
Hallucination flags per verdict	0.24	0.05	0.20	0.19	0.13

Table 10: Judge-by-lane mean score matrix (M2). The scale effect is visible: the judges disagree on absolute level far more than on ordering. Hallucination flags per verdict appear in the final row.

The registered family-level self-preference matrix degenerates because every lane's published narrative is authored by the same model family (the writer or decider step), so the matrix is reported

per lane with the authorship confound named. The registered 10 % stratified human spot-check remains open (operator task).

- **[2026-07-03] Registered engagement-layer artifacts.** The §8 points ladder was not displayed live during the tournament. Because it is deterministic and the registration permits retroactive display, it is computed from the published predictions and included (Appendix F). The §8 full-bracket commitment (each lane publishing a complete knockout bracket before the first Round-of-32 kickoff) was not met and cannot be reconstructed with clean provenance, so it is noted as an unmet commitment. Both are engagement surfaces, not headline metrics, and neither bears on the calibration result.
- **[dated] Operational incidents.** Two over the tournament, both recovered with zero fixtures lost: the 2026-06-23 provider rate limit (2 fixtures) and the 2026-07-03 dispatch hang (1 fixture, recovered pre-kickoff, and the timeout fix is a post-freeze transport-level change disclosed under M6). One of the rate-limited fixtures, id 1489404, was re-harvested and its predictions regenerated from a clean pre-kickoff pack. Because its provenance differs from the untouched fixtures, we report headline results with it included and confirm they are unchanged when it is held out, so a reader can see the single regenerated fixture moves nothing. Details: `NOTES-incidents.md` in the source repository, published with the dataset.

11 Conclusion

We asked whether the confidence an AI system reports is worth anything, and made the question falsifiable: a pre-registered live benchmark, every prediction and its cost published before kickoff, the scoring rules locked before the first scored match. The answer the tournament gave is not the one a leaderboard would celebrate. On accuracy no architecture owns a chance-dominated outcome, and on the held-out set no lane is statistically distinguishable from another. The prior we stated, that no AI system would beat the market on raw accuracy, was inverted in direction and left unresolved in magnitude, both of which we publish.

The finding that survives is about trust, not accuracy. A forecast that says thirty percent and fails has told the truth about its own uncertainty. A forecast that says seventy percent and fails has not. The number that separates them is calibration, and it is the number a business actually consumes when it puts an AI-supplied probability into a real decision. Over the full tournament the council, three independent model families and a step that weighs their disagreement, posts the lowest calibration error of any machine we tested, the human reference aside. If the operational question is which architecture to trust with a forecast, the evidence points to the one whose stated confidence is most honest, not the one that calls the most games, and across this tournament that is the council. We state the strength of that reading plainly: on a chance-dominated event the bootstrap intervals overlap, so it is a directional lead a second tournament would sharpen, not a resolved separation. A directional lead is still the answer an operator asked for, which architecture to reach for, and this study gives it.

That the machines land near the market at all is the football result, not the business one. Football has a deep public betting market that already prices every match, so a good forecast has a hard reference to be measured against and little room to beat. A company forecasting its own pipeline, its own demand, its own risk has no such market and no closing line, so the calibrated AI probability is not competing with a reference, it is the closest thing to a reference the company can build. That is where this benchmark pays

off outside the stadium: it tells an operator which shape of AI reports the most trustworthy confidence when nothing else is scoring it, and on this evidence that shape is the council.

The limits are stated where they belong, in Section 10, and none of them unseats the result: this is one tournament, and it is a full one, scored in public against a locked protocol. The central claim was never that any system predicts the future. It is that a system's account of its own uncertainty is the first thing to measure about it, and the first thing to trust or distrust, and that on this evidence the council is the architecture that reports it most honestly. You cannot automate what you do not understand, and a confidence number you have not audited is not something you understand yet.

References

[1] Ali Al-Bustami and Zaid Ghazal. From players to champions: A generalizable machine learning approach for match outcome prediction with insights from the FIFA World Cup. In *2025 IEEE International Conference on Electro Information Technology (eIT)*, pages 574–578, 2025. doi: 10.1109/eit64391.2025.11103598.

[2] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

[3] Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs, 2023. Presented at NAACL 2024.

[4] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2023. Presented at ICML 2024.

[5] Leonardo Egidi, Roberto Macrì Demartino, and Vasilis Palaskas. footBayes: Fitting Bayesian and MLE football models. CRAN: Contributed Packages, 2022.

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

[7] 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), 2015. doi: 10.1515/jqas-2014-0051.

[8] 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.

[9] Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models, 2024.

[10] 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. In *International Conference on Learning Representations (ICLR)*, 2025.

[11] Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973.

[12] Pieter Robberechts and Jesse Davis. Forecasting the FIFA World Cup: Combining result- and goal-based team ability parameters. In *Machine Learning and Data Mining for Sports Analytics (MLSA), Lecture Notes in Computer Science*, pages 16–30, 2019. doi: 10.1007/978-3-030-17274-9_2.

[13] Pieter Robberechts, Jan Van Haaren, and Jesse Davis. A Bayesian approach to in-game win probability in soccer. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 3512–3521, 2021. doi: 10.1145/3447548.3467194.

[14] Maohao Shen, Subhro Das, Kristjan Greenewald, Prasanna Sattigeri, Gregory Wornell, and Soumya Ghosh. Thermometer: Towards universal calibration for large language models, 2024. Presented at ICML 2024.

[15] Dat Tran and Douwe Kiela. Single-agent LLMs outperform multi-agent systems on multi-hop reasoning under equal thinking token budgets, 2026.

[16] Xiaochen Zhu, Caiqi Zhang, Yizhou Chi, Tom Stafford, Nigel Collier, and Andreas Vlachos. Demystifying multi-agent debate: The role of confidence and diversity, 2026.

A Prompts, verbatim

The prompts below are reproduced from the hash-locked source (`web/convex/lib/prompts.ts`, `PROMPTS_VERSION` 1.0, frozen at the M1 pre-registration lock), SHA-256 `1ce8d0…a9d0966`. The full file, its checksum, and the prediction output schema are published with the dataset so this listing can be verified against what ran. Runtime templates interpolate the match header and the assembled context pack at dispatch time.

Shared analyst system prompt. Used by SOLO, SOLO-ZERO, the PIPELINE voice editor, and every COUNCIL member and decider.

```
You are a football analyst forecasting a 2026 World Cup match. Confidence must
be a numeric value 0-100 representing your actual probability estimate that your
predicted winner is correct. Do not hedge, calibration is scored. riskFactors:
max 3 short strings. reasoning: max 100 words, plain language, direct, no buzzwords,
no emojis, no semicolons, no em dashes.
```

Single-call user templates. SOLO receives the header plus the seven-layer context pack. SOLO-ZERO receives the header alone.

```
[A1] {matchHeader} + Full context pack (seven layers, some may be null): {contextPack}
+ Predict the match outcome.
```

```
[A1-zero] {matchHeader} + No additional context is provided. Predict the match
outcome from the fixture information above only.
```

PIPELINE step prompts (three sequential specialists).

```
[Step 1, statistician] You are a football statistician. From the quantitative
layers of the context pack (stats, weather), produce a strictly statistical read
of the match. No narrative, no news.

[Step 2, operations reader] You are an operations reader. Combine the statistician's
read with the narrative layers (news, rivalries, prior insights, learning loop
if present). Surface what the numbers miss: availability, momentum, stakes, conditions.

[Step 3, voice editor] You are the final voice editor of a sequential analyst
pipeline. Synthesize the statistician's and operations reads into one final prediction.
(+ shared analyst system prompt above.)
```

Council prompts. Each member runs the shared analyst prompt. The decider runs:

```
You are the decider of a three-member analyst council. You receive three independent
predictions from different model families. Weigh their agreement and disagreement,
then produce ONE final prediction. Where they disagree, your confidence should
reflect that disagreement. (+ shared analyst system prompt above.)
```

Context-harvest queries (L2 news, L3 rivalry).

```
[L2 news] Latest team news for {home} and {away} national football teams in the
last 7 days: injuries, suspensions, lineup signals, form. World Cup 2026 context.

[L3 rivalry] Historical rivalry and head-to-head narrative between {home} and
{away} national football teams. Memorable encounters, styles of play, current
tactical identities.
```

Prediction output schema. All lanes emit a JSON object validated against one schema: `winner` (home / away / draw), `score` (integer home and away), `confidence` (number 0–100), `riskFactors` (array of strings), `reasoning` (string). Structured output is enforced with one retry on schema failure, and a second failure is recorded as an abstention, never silently repaired. Each Pipeline step carries its own intermediate schema (statistical read, operational read), published in the source.

B Protocol artifacts and integrity record

[TODO: Final table generated at publication. Contents fixed now:]

- Pre-registration: OSF ke35t, DOI 10.17605/OSF.IO/R5SBJ, with the SHA-256 checksum table binding the registered documents and the production code they govern (prompts, model gateway, seeded-pattern injection). Documents bound: `pre-registration.md`, `statistical-tests.md`, `seeded-patterns-sealed.md`.
- The schema-drift disclosure: post-lock additive operations tables (registry of record, dispatch-claim ledger, display-layer tables), each disclosed in its commit, with methodology tables untouched.
- The M8 first-matchday hand-check record (June 11, 2026): independent recomputation of the Brier score for all eight lanes of the opening match, agreement at three decimals (gate required five outputs).

- The Elo parameter disclosure (registry key `elo:params`): conventions, empirical draw share, fixtures processed, as-of date.
- The amendment log, dated, versioned, each published before any match it affects.

C Cost and latency accounting

Every scored prediction carries its billed API cost in the published telemetry. The figures below are the cumulative pre-prediction spend per lane over all 104 matches, read from the frozen tournament export, so each row reconciles exactly with the Cost column of Tables 3 and 7. The deterministic baselines (Elo, Market) and the human reference incur no API cost and are omitted.

Lane	Tournament total	Per match	Relative to Solo-Zero
Solo-Zero	$1.06	$0.010	1.0×
Solo	$8.22	$0.079	7.8×
Council	$15.03	$0.145	14.2×
Pipeline	$21.93	$0.211	20.7×
Pipeline-Static	$21.93	$0.211	20.7×
Full AI slate	$68.18	$0.656	

Table 11: Cost accounting, all 104 matches. Tournament total is the cumulative billed API cost per lane, per match divides by the 104 scored fixtures, and the relative column prices each architecture against the context-free call. The full AI slate is the five-lane per-match cost an operator running every architecture in parallel would pay.

The operational reading is in Section 9.4. The whole five-lane tournament cost $68.18 in API spend, so the binding constraint on a first internal benchmark is the methodology, not the budget. The cost-per-calibration-point comparison an adopter cares about, dollars per lane against ECE per lane, is the frontier that section reads off Table 7: the council buys the study's lowest calibration error at roughly fourteen times the context-free call and well under the pipelines, which spend the most and return no calibration edge. Per-call token counts and wall-clock latency for every prediction are in the published per-prediction dataset.

D Adversarial review

Before release the final draft was put through an internal adversarial read: four independent reviewers, each given a distinct lens and instructed to find real defects rather than to approve. The lenses were statistical rigor (does the calibration math hold, are the intervals labeled honestly), operator and finance credibility (does the recommendation give a decision-maker something to act on, does the market-to-business bridge follow), business-purpose fit (does the paper stay a decision-grade artifact or lapse into method-for-its-own-sake), and brand and internal coherence (voice discipline, section-to-section consistency). Each reviewer read the compiled paper independently and reported findings ranked by severity.

The review left the central finding standing: all four reviewers confirmed the directional council recommendation holds on the full-tournament evidence, with the overlapping calibration interval disclosed as a limit rather than a retraction. Their substantive corrections were adopted before release. The sharpest was the market comparison: the council's calibration error sits about four thousandths below the market's,

inside the study's own replication noise floor, so the paper now names that comparison as a tie at this resolution and resolves it through the point that a business has no market line to tie against (Section 8.1). Others sharpened the cost-versus- calibration reading of the architecture choice (Section 9.4) and the placement of the operator result in the abstract.

Independent external reads, by a domain journalist, a practicing operator, and an outside AI specialist, are planned post-publication and will be reproduced here with responses as they arrive. This appendix records the pre-release adversarial process, not a claim of external peer review.

E Lane and metric reference

A reader-facing reference: each of the eight analysts and each reported metric, defined in one place. Everything here transcribes the registered protocol and the disclosed implementation notes. Nothing in this appendix is a new commitment, and no tournament result appears here.

E.1 The eight analysts

Solo (A1), architecture. One call to a pinned model with the shared analyst system prompt and the full seven-layer context pack (statistics, weather, news, rivalries, generated insights, prior analysis, learning loop). Structured output enforced by JSON schema with one retry. A second failure records an abstention. The simplest thing an operator could deploy, and therefore the reference architecture.

Solo-Zero (A1-zero), control. The literal Solo code path with the context pack withheld: fixture metadata only (teams, stage, venue, kickoff). Everything else is identical, so the gap between the twins prices the context pack itself. It answers: what did all that context buy?

Pipeline (A2), architecture. Three sequential calls on one pinned model family: a statistician reads only the quantitative layers, an operations reader adds the narrative layers, a voice editor produces the final structured prediction. Each step has its own schema and the one-retry rule. A registered escalation gate is part of the architecture: when the pipeline's final confidence falls below 60 it adopts the Council prediction for that match, logged and cost-attributed (the pipeline still pays for its own three steps).

Pipeline-Static (A2-static), control. The full pipeline with the learning-loop layer nulled from its context, excluded from loop writes, run on every match Pipeline runs. Registered to isolate the learning loop. (Section 7 reports how this pair must actually be read, given that the loop layer was never populated.)

Council (A3), architecture. Three members from distinct model families receive the identical prompt and context in parallel. A decider from the first family receives the three submissions with model identities removed (anonymous member 1/2/3), is instructed that its confidence should reflect member disagreement, and produces the lane's single prediction. Architecture and model mix are confounded by design and disclosed as such.

Elo, deterministic baseline. An Elo rating system computed from international match history (window 2022-06-01 to tournament start, all fixtures of the 48 qualified teams, deduplicated). Expected score

$W_e = 1/(1 + 10^{-(R_h+H-R_a)/400})$. History-side home advantage $H = +100$ for the listed home team. For tournament predictions $H = +100$ only when the team is a host nation (Mexico, USA, Canada), 0 otherwise, per the host dummy of Robberechts and Davis [12]. $K = 30$ for World Cup fixtures, $K = 20$ otherwise, scaled by a goal-differential multiplier (1 for a draw or one-goal margin, 1.5 for two, 1.75 for three, plus 1/8 per additional goal). Three-way probabilities derive from the identity $W_e = P(\text{win}) + P(\text{draw})/2$ with the draw share taken empirically from the backfill window and probabilities clamped to $[0.01, 0.97]$. Off-graph opponents enter at 1500. Deterministic: the same history reproduces the lane's every prediction bit-for-bit. No AI is involved at any step.

Market, deterministic baseline. The betting market's closing line, one snapshot at T−1 h per match, Pinnacle preferred with a registered fallback chain (Bet365, then the first available 1X2 book). Bookmaker odds embed a margin, so the implied probabilities are de-vigged proportionally: with decimal odds d_k for the three outcomes, $p_k = (1/d_k)/\sum_j (1/d_j)$. The lane's pick is the argmax outcome and its confidence is that outcome's de-vigged probability. The market line never enters any AI lane's context. It is the hardest widely available benchmark in sports forecasting, and the registered prior expects no AI lane to beat it on raw accuracy.

Human, non-blind human reference. One named human (the registrant). Picks were entered by the registrant, who was not blind to the AI lanes' outputs or the running results (see the non-blind-reference disclosure, Section 3), any time before kickoff, locking at kickoff. Re-entry before kickoff replaces the prior pick. A missing pick is shown as missing and scores nothing. The reasoning text is published verbatim. This lane is reported for transparency, not as a controlled human-versus-machine test, and any calibration advantage it shows is confounded by information access. It is one person, not a panel, and is read accordingly.

E.2 The metrics, in plain terms

Formal definitions live in Section 5. This is the reading guide. Throughout, a lane's confidence c on its predicted winner maps to $p = c/100$, and y records whether that winner won.

Season Brier score (multiclass). The penalty score for a full three-way probability vector against what happened, summed over home/draw/away. Lower is better. Perfect foresight scores 0, an always-uniform split scores 0.667, and confident misses cost the most. This is the locked headline metric. The probability vector is constructed by the registered rule of Section 5.1.

Binary Brier score. The same idea collapsed to the predicted winner: $\text{mean}\big((p - y)^2\big)$. A lane that always said 50 would score 0.25 regardless of outcomes, and skill and honesty both push below that. This is the space the calibration analysis lives in.

ECE (expected calibration error). Groups predictions into ten confidence bins and asks, per bin, how far stated confidence sits from the realized hit rate, weighted by how often the lane spoke at that confidence. The single number to read first: an ECE of 0.05 means stated confidence tracks reality within about five points on average, an ECE of 0.15 means the number the lane reports is off by fifteen points on average. Zero is a lane whose 70 means 70.

Reliability, resolution, uncertainty. The three-way anatomy of the binary Brier score. Reliability is miscalibration itself (lower is better): the squared version of what ECE measures in absolute terms. Resolution rewards a lane whose confident and unconfident predictions actually sort wins from losses (higher is better). Uncertainty is the task's own base-rate variance and belongs to the World Cup, not to any forecaster. It is the floor that chance sets for everyone.

Sharpness. The spread of a lane's stated confidences (variance around its own mean forecast). Calibration alone can be gamed by never committing. Sharpness is the metric that catches the lane which is well-calibrated only because it always says 55. The pair to read together is ECE low *and* sharpness high: honest *and* willing to commit.

Hit rate. The plain fraction of correct winners. Deliberately demoted: it is the number everyone asks for first and the one this benchmark trusts least, because it rewards bold guessing and hides miscalibration entirely.

Bootstrap 95 % intervals. The stability band around every reported metric, from 1,000 match-block resamples with a fixed published seed. The registered claim rule: two lanes differ only if their held-out intervals do not overlap. Overlap reads as "no resolvable difference," and that phrase appears wherever it is true.

Stage-stratified Brier score. The same score split by tournament stage, because knockout football is harder to forecast than group play and a lane's aggregate can hide that. Published as group versus knockout (the finer registered strata are computable from the released dataset).

F The points ladder

The points ladder is a public engagement surface, registered in the protocol to make the benchmark legible to non-specialist visitors. It is not a calibration claim and enters no headline result: it rewards confident correct calls and penalizes confident wrong ones, which is a different question from whether a lane's confidence is well-calibrated. We include it here, computed retroactively from the published predictions, because the registration defines it as a deterministic function of the record and permits retroactive display.

The rule, fixed in the pre-registration, awards each lane (confidence if the pick is correct, else $-$ confidence) $\times$ round multiplier per scored match, with the multiplier set to 1.0 for the group stage, 1.5 for the Round of 32 through the quarter-finals, and 2.0 for the semi-finals, third-place match, and final. Abstentions and missing picks score zero. Table 12 reports the final standings over all 104 matches.

The council finishes third on this ladder, behind the context-free single call and the market. That ordering is expected and it does not qualify the study's finding: the ladder rewards confident accuracy, the axis this paper shows resolves nothing on a chance-dominated task, while the calibration axis that carries the finding puts the council first. A lane can rank high here by being boldly right on the matches it happened to call and still be less honest about its own uncertainty across the full set. The ladder is the engagement surface; the reliability curve is the result.

Rank	Lane	Points
1	Solo-Zero	3436
2	Market	3094
3	Council	2944
4	Solo	2940
5	Pipeline-Static	2891
6	Pipeline	2860
7	Human	2205
8	Elo	1673

Table 12: Final points ladder, computed retroactively over all 104 matches under the registered rule. The ladder rewards confident accuracy, so it tracks the accuracy ordering rather than the calibration ordering that carries the study's finding, and the two are deliberately reported side by side.