SCALAR RECALIBRATION CAN OVERSTATE CONTEXTUAL GAINS IN PAIRWISE PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Does a contextual predictor add value if its advantage survives recalibration of a simpler model? In pairwise prediction, this conclusion can depend on whether the calibrator observes which competitor occupies a known side. An ordinary Platt intercept in blue-first or white-first coordinates is exactly a side-dependent intercept in randomly ordered coordinates. We use this identity to compare ordinary and side-aware Platt scaling while holding each model's forecast stream fixed. Across seven annual League of Legends splits and nine weekly Lichess splits, covering 70,670 and 1,340,775 evaluation matches, respectively, an online side term improves a calendar rating model's log loss after ordinary Platt scaling by 0.001144 and 0.000119. After both readouts receive side information, the differences reverse to +0.000139 and +0.000013, with 95% paired calendar-block intervals of $[-0.000087, 0.000325]$ and $[0.000004, 0.000024]$. A four-arm League of Legends ablation also finds no proper-score benefit from the existing frozen rank-five interaction term; shared and independent validation selection yield identical forecasts. These results do not establish equivalence or rule out richer contextual models. They show that survival under scalar recalibration is insufficient evidence of contextual value beyond a simpler predictor with the same side information.

1 INTRODUCTION

Pairwise predictors routinely use context: blue or red side in a competitive game, white or black pieces in chess, and home or away status in a sporting event. A natural question is whether an apparent improvement reflects a useful change to the predictive model or a correction that a simpler probability readout could supply. Comparing models after recalibration is one way to ask that question. Its answer, however, depends on the information available to the readout, as well as its functional form.

Consider the probability that the first recorded competitor wins. When the recording order is randomized, a single Platt intercept cannot directly express an advantage for the blue team: the correction must change sign when blue is recorded second. When the same game is written blue-first, an ordinary intercept can express that advantage. The physical game has not changed, but the readout has gained access to side through the definition of its target. Consequently, survival under ordinary scalar calibration need not identify a contextual contribution that survives an information-matched comparison.

We examine this ambiguity through a rating-model archive, then test whether the finding persists beyond its implementation and dataset. The historical Contextual Dynamic Relational Skill (CDRS) extension combines a calendar rating backbone with an online side offset and a frozen pairwise interaction. Its accuracy advantage over the backbone survives ordinary Platt scaling. Applying a side-aware readout to the same stored forecasts largely removes that advantage. Because the archive also contains implementation defects and uses overlapping tuning and calibration data, this reconstruction alone cannot establish a general evaluation lesson.

Our main experiments therefore use a separately implemented, swap-coherent calendar filter with daily feedback and disjoint tuning and calibration periods. On League of Legends (LoL), we evaluate the backbone, side only, the existing frozen rank-five interaction only, and both additions. On public Lichess games, we repeat the side comparison using the same numerical kernel and temporal ordering, with white/black replacing blue/red. The ordinary-Platt side advantage is absent at the

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

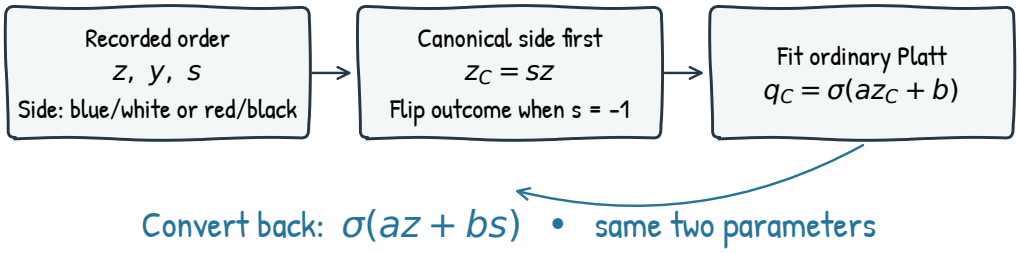


Figure 1: **Canonical ordering supplies side to the calibrator.** The canonical competitor is blue in League of Legends and white in chess. Fitting ordinary Platt scaling in canonical coordinates is exactly equivalent to fitting a side-dependent intercept in the original coordinates. Both readouts have two parameters. This is a change to the readout, with the underlying prediction stream held fixed.

point-estimate level under side-aware Platt in both domains. The frozen interaction does not restore a proper-score advantage on LoL.

The contribution is a documented empirical failure of an attribution criterion, rather than a new calibration algorithm. We provide (i) an exact accounting of the orientation-dependent information available to a two-parameter calibrator; (ii) fixed-stream controls in two pairwise prediction domains; and (iii) a four-arm ablation that separates the side effect from a frozen interaction under common grids and temporal controls. All comparisons retain unfavorable outcomes. Our conclusion concerns the evaluated binary side variables and rating recipes; it does not imply that context or intransitivity is generally unhelpful.

2 Related Work

Pairwise ratings and context. Bradley–Terry represents a comparison by a skill difference, while Elo updates ratings incrementally (Bradley & Terry, 1952; Elo, 1978). Glicko models uncertainty, and Glicko-2 adds volatility (Glickman, 1999; 2022). Calendar dynamics and uncertainty propagation are also central to TrueSkill Through Time (Herbrich et al., 2006; Dangauthier et al., 2007). Context-dependent preferences and intransitive matchup models are established extensions (Chen & Joachims, 2016a;b). We use a frozen antisymmetric interaction as an ablation of an existing implementation; neither that parameterization nor the calendar filter is presented as a new rating method.

Calibration and the available information. Platt scaling fits an affine logistic readout (Platt, 1999); temperature and beta calibration make different choices of readout class (Guo et al., 2017; Kull et al., 2017). Calibration assessment depends on conditioning information and the evaluation procedure (Vaicenavicius et al., 2019). Multicalibration explicitly considers subpopulations, while omniprediction asks for predictions that support a family of downstream losses (Hebert-Johnson et al., 2018; Gopalan et al., 2022). The control studied here is much narrower: it supplies one known binary covariate to both readouts and claims neither guarantee. Its algebra is elementary. The empirical question is how much an otherwise surviving contextual gain depends on omitting this control.

Probability quality and decisions. Log loss and Brier score are strictly proper scores; they evaluate the predictive distribution, not calibration alone (Gneiting & Raftery, 2007). Decision calibration links probabilities to specified decision classes (Zhao et al., 2021). We report proper scores as primary evidence and fixed asymmetric decision losses as a descriptive supplement. Improvement at the accuracy threshold is not treated as evidence of improvement for every decision cost.

3 A FIXED-STREAM CONTROL FOR SIDE INFORMATION

Let $Y \in \{0, 1\}$ indicate whether the first recorded competitor wins, let $P \in (0, 1)$ be its predicted probability, and write $Z = \text{logit}(P)$. Let $S = +1$ when that competitor occupies the canonical side and $S = -1$ otherwise. For a coherent predictor, swapping the competitors maps (Y, Z, S) to $(1 - Y, -Z, -S)$. Define canonical coordinates by

$$Y_C = \frac{1 + S(2Y - 1)}{2}, \qquad Z_C = SZ. \tag{1}$$

The two readouts we compare in the original coordinates are

$$q_{\text{scalar}}(Z) = \sigma(aZ + b), \qquad q_{\text{side}}(Z, S) = \sigma(aZ + bS), \qquad a > 0. \tag{2}$$

Both fit a slope and one intercept coefficient. The difference is the covariate carrying the intercept, not parameter count.

Proposition 1 (Canonical and side-aware Platt equivalence)**.** *For any fixed array (z_i, y_i, s_i), fitting $\sigma(az_{C,i} + b)$ by binary log loss in canonical coordinates is equivalent to fitting $\sigma(az_i + bs_i)$ in the original coordinates. Their objectives agree for every (a, b), including any identical penalty on these coefficients. Mapping predictions back gives the same probabilities and any loss invariant to simultaneous complementation of the probability and label. The side-aware map preserves swap complementarity; the ordinary affine map preserves it for all z only when $b = 0$.*

Proof sketch. For $s = -1$, mapping the canonical probability back gives $1-\sigma(-az+b) = \sigma(az-b)$. Binary log loss satisfies $\ell(1 - q, 1 - y) = \ell(q, y)$. Thus every sample has the same loss in the two representations. Complementarity follows from $\sigma(-x) = 1 - \sigma(x)$; setting $z = 0$ gives necessity of $b = 0$ for the ordinary map. Appendix A gives the optimization statement explicitly.

The equivalence is an identity for stored predictions. It does not require retraining a rating model under a different ordering, and it does not assert that an incoherent model would produce complementary forecasts if rerun after a swap. This distinction is useful for auditing the legacy implementation.

For a model with side term and its backbone, our estimand is the difference in held-out proper score after each receives the same specified readout class. Comparing that difference under the two maps in Eq. (2) measures its dependence on readout information and constraints. We fit each map on an earlier calibration period and freeze it during evaluation. The raw stream for each model is identical in both conditions. Equal information is a necessary comparison control, not a guarantee that two arbitrary readout classes are equally expressive or equally well estimated.

4 EXPERIMENTAL DESIGN

4.1 TWO ARCHIVES WITH OBSERVED SIDE

League of Legends. The existing Oracle's Elixir-derived table contains 93,213 matches from 2014–2025 (Oracle's Elixir, 2025). We use data through 2024, leaving 83,172 rows. For the main experiment, removing 12 self-comparisons and 39 repeated canonical time/team keys leaves 83,121 matches; 70,670 are evaluated in seven annual folds, 2018–2024. For evaluation year t, interaction fitting uses all observations before $t - 1$, hyperparameters are selected on January–June of $t - 1$, and readouts are fitted on July–December. The rating states advance through both validation halves and the test year. Team identifiers, timestamps, side, and past outcomes are the only model inputs.

Lichess. We use an existing table built from the public March–May 2026 standard rated game dumps (Lichess, 2026). The upstream cleaner retains 180+0 time controls, valid integer rating headers, timestamps, and results, then systematically subsamples to approximately two million rows. It stores the actual `white_is_p1` flag; we do not reconstruct side from a random seed. Of 2,028,767 rows, 86,739 are draws. Removing these leaves 1,942,028 decisive games and 379,559 player identifiers, with no additional missing-identifier, self-comparison, or canonical-key exclusions. The target is the winner *conditional on a decisive outcome*; these are not unconditional win or draw probabilities. Neither site ratings nor moves are predictive inputs.

We evaluate nine consecutive weeks beginning March 30 through May 25, totaling 1,340,775 games. Each test week is preceded by one calibration week, one tuning week, and all earlier available history. This uses the same train–tune–calibrate–test ordering as LoL, with a shorter calendar unit rather than an identical window length. Every test game appears once. Training histories overlap across folds, so the folds are not independent replications. The historical archives and earlier findings were already available when this study was designed; temporal splitting does not make this a prospective study.

4.2 One implementation and a four-arm LoL ablation

All new experiments use the same numerical kernel. The calendar model maintains a mean μ_i, deviation ϕ_i, and volatility σ_i for each competitor. Uncertainty increases with elapsed calendar days. Its public logit is

$$z_{ij,t} = g\Big(\sqrt{\phi_i^2 + \phi_j^2}\Big)\,(\mu_i - \mu_j) + \beta_t s_{ij,t} + h_{ij}, \qquad g(x) = \big(1 + 3x^2/\pi^2\big)^{-1/2}. \tag{3}$$

The first term is a symmetric uncertainty-attenuated rating difference. The side coefficient updates by $\beta \leftarrow \beta + \eta s(y - p)$, using the current model probability at feedback. Competitor updates use a common pre-update snapshot. All predictions for a UTC day are made before any outcome from that day is applied; feedback then follows chronological order. This prevents uncertain within-day outcome availability from entering the predictions. The filter is a calendar-time Glicko-style adaptation, not an exact Bayesian posterior or the standard batch Glicko-2 algorithm. Appendix B specifies all updates.

For LoL, the four arms activate neither addition (*base*), β only (*side*), h only (*interaction*), or both. We reuse the archive's rank-five term

$$h_{ij} = u_i^\top v_j - u_j^\top v_i, \qquad u_i, v_i \in \mathbb{R}^5. \tag{4}$$

Its unchanged fitting routine minimizes training log loss for $\sigma(h_{ij})$ plus an L_2 penalty, without a joint skill or side component. The embeddings are then frozen and used as offsets in Eq. (3) and the conditional updates. An unseen competitor gives $h_{ij} = 0$. This deliberately tests the existing frozen interaction recipe, rather than a redesigned residual or joint interaction estimator.

Common grids and selection control. The calendar period scale is in $\{3, 7, 14, 30, 60\}$ days, the active side learning rate in $\{0.01, 0.05, 0.1\}$, and the interaction penalty in $\{1, 5, 20, 100\}$. We first select a common triple by the mean tuning log loss of the four arms over all 60 triples. Each arm uses that triple's active parameters. Inactive parameters do not create extra fits: the distinct candidate counts are 5/15/20/60. We also select each arm independently by its own tuning log loss over the same grids. Both rules select identical forecasts in all seven folds. Thus the result does not depend on forcing an unfavorable common configuration.

Embedding fits use the original seed zero and a 300-iteration L-BFGS-B cap. Every selected fit reports convergence. Because 13 unselected weak-penalty fits hit the cap, we additionally repeat all grids with a 1,200-iteration cap. Selected configurations and predictions are unchanged. Five penalty-1 fits still reach this larger cap; no claim of global optimization is made.

On chess, we compare calendar with and without side and repeat the comparison for Elo. The latter uses $K \in \{16, 24, 32, 48, 64\}$ and the same side learning rates and daily feedback. Each model is selected by tuning log loss. No chess interaction is introduced. Numerical checks against a separately written Python reference and under mirrored histories are documented in the artifact.

4.3 Readouts, scores, and conditional uncertainty

Both readouts minimize summed calibration-period binary log loss plus $10^{-3}[(a-1)^2 + b^2]$, with $a \geq 10^{-8}$. The slope constraint, penalty, optimizer, and fitting data are identical. Raw and transformed probabilities are clipped to $[10^{-12}, 1 - 10^{-12}]$ for finite scoring. No readout is selected by test performance. Log loss is primary; Brier score and accuracy are secondary, with half credit for a probability exactly equal to $1/2$.

We report macro means over seven years or nine weeks, alongside pooled means in the artifact. Paired 95% percentile intervals use 5,000 circular calendar-block resamples within each test fold,

Table 1: **The same readout comparison in two domains.** Macro means over seven LoL years and nine chess weeks. Accuracy is in percent; lower log loss is better. LoL uses the shared configuration; independent selection gives identical predictions. Absolute scores should be compared within a domain. Brier and raw scores appear in Appendix D.

		Log loss ↓		Accuracy (%) ↑	
Data	Model	Ordinary	Side-aware	Ordinary	Side-aware
LoL	Base	0.643493	0.641981	62.432	62.626
	Side	0.642349	0.642121	62.699	62.610
	Interaction	0.643635	0.642126	62.442	62.606
	Both	0.642488	0.642262	62.693	62.573
Chess	Calendar	0.691867	0.691265	51.649	52.344
	Calendar + side	0.691749	0.691278	51.786	52.330
	Elo	0.691806	0.691204	51.474	52.185
	Elo + side	0.691583	0.691338	51.746	52.146

then average fold differences. Main intervals use 30-day blocks for LoL and three-day blocks for chess. We also report 7/90-day and 1/2-day choices, respectively, and resample whole chess weeks as an additional sensitivity (K"unsch, 1989). These are conditional intervals for the fitted forecast streams and maps: they omit training, selection, and calibrator-fitting uncertainty. Weekly blocks cannot establish independence at longer horizons. They are descriptive uncertainty assessments, not equivalence tests or multiplicity-adjusted discovery claims.

5 Results

5.1 The fixed-stream historical comparison motivates the control

We first recover the original selected Elo, backbone (P2), and contextual (P2X) forecast streams. All 42 archived model–year–raw/recalibrated accuracy entries reproduce exactly; the maximum log-loss discrepancy is 4.7×10^{-12}. This reconstruction retains the original event feedback and full-year validation, including overlap between selection and calibration; it covers 70,720 evaluation matches. It is kept separate from the main experiment.

With ordinary Platt scaling, P2X minus P2 has log-loss difference -0.001393 and accuracy difference $+0.327$ percentage points (pp). Supplying side to both readouts, with exactly the same raw streams, changes these to $+0.000091$ and $+0.035$ pp. The corresponding intervals are $[-0.000258, 0.000413]$ and $[-0.136, 0.203]$ pp. Thus the historical "survives recalibration" observation depends on which information the calibrator sees. The old implementation's swap and state-update defects make it unsuitable as the sole evidence for this claim; Appendix C records them without treating a bug fix as a new model contribution.

5.2 Side gains do not survive the matched control at the macro level

Table 1 reports the new daily-feedback experiments, and Table 2 reports paired contrasts. On LoL, side improves the calendar backbone's ordinary-Platt log loss by 0.001144, with interval $[0.000724, 0.001562]$ for the improvement. Its accuracy gain is 0.266 pp, with interval $[0.112, 0.437]$. Under side-aware Platt, side minus base instead has log-loss difference $+0.000139$ and accuracy difference -0.016 pp; both intervals include zero. This is consistent with the apparent advantage being accessible to a simpler model's readout. It is not proof of population equivalence, and no smallest meaningful effect was specified.

On chess, calendar plus side improves ordinary-Platt log loss by 0.000119, with interval $[0.000086, 0.000151]$, and accuracy by 0.137 pp, with interval $[0.097, 0.176]$. After both readouts observe white/black side, the log-loss difference is $+0.000013$ with interval $[0.000004, 0.000024]$, slightly favoring the backbone, while accuracy differs by -0.014 pp. Whole-week resampling gives a log-loss interval $[0.000002, 0.000027]$. Elo exhibits the same qualitative change: its side extension improves ordinary-Platt log loss by 0.000223 but worsens side-aware log loss by 0.000134. The

Table 2: **Paired model differences and conditional 95% intervals.** Log-loss differences are multiplied by 10^3; accuracy differences are in percentage points. Negative log loss favors the first model. Intervals use 30-day blocks for LoL and three-day blocks for chess. Ordinary-Platt intervals and block-length sensitivities are in Appendix D.

	$10^3 \times \Delta$ log loss ↓		Δ accuracy (pp) ↑
Contrast	Ordinary	Side-aware [95% interval]	Side-aware [95% interval]
Legacy: P2X − P2	−1.393	+0.091 [−0.258, +0.413]	+0.035 [−0.136, +0.203]
LoL: side − base	−1.144	+0.139 [−0.087, +0.325]	−0.016 [−0.156, +0.136]
LoL: interaction − base	+0.142	+0.145 [+0.023, +0.267]	−0.020 [−0.069, +0.022]
LoL: both − side	+0.139	+0.142 [+0.023, +0.262]	−0.038 [−0.093, +0.009]
Chess: side − calendar	−0.119	+0.013 [+0.004, +0.024]	−0.014 [−0.037, +0.010]
Chess: side − Elo	−0.223	+0.134 [+0.085, +0.186]	−0.038 [−0.090, +0.010]

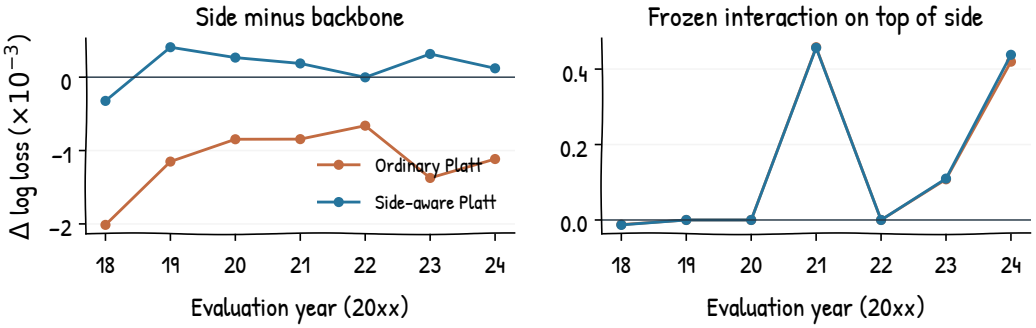


Figure 2: **Annual heterogeneity in the LoL ablation.** All seven annual log-loss contrasts are shown, without selecting favorable years. Negative values favor adding the indicated term. Lines connect annual estimates and do not represent a fitted temporal trend. Empirical coordinates are exact; the hand-drawn style affects lettering and decoration only.

reversal is not specific to the calendar filter. An exploratory three-parameter readout retaining both ordinary and side intercepts gives the same qualitative results (Appendix D.1).

These are small effects in a sparse-history chess sample. For example, side-aware calendar accuracy is approximately 52.3%. We make no claim to strong chess prediction or superiority over a system supplied with site ratings. Also, the raw chess side model does not improve calendar log loss before readout fitting (Appendix D); the claim concerns the ordinary-Platt comparison actually being tested.

5.3 The frozen interaction does not restore a proper-score benefit

The four-arm ablation tests whether the conclusion merely omits the archive's second contextual component. Under side-aware Platt, adding the frozen interaction to the backbone changes log loss by $+0.000145$, with interval $[0.000023, 0.000267]$. Adding it on top of the side model changes log loss by $+0.000142$, with interval $[0.000023, 0.000262]$. Both comparisons favor omitting the interaction. Brier score also becomes worse at the macro point-estimate level. Ordinary Platt does not yield an interaction advantage either. Shared selection and independent per-arm selection coincide, and raising the fitting cap from 300 to 1,200 iterations leaves all reported selected forecasts exactly unchanged.

The selected penalties are 20 in four years and 100 in three. At penalty 100, the selected offsets are nearly zero; at penalty 20, their mean absolute test logit ranges from 0.0023 to 0.0528. Thus this is partly a shrinkage result, not evidence that every fit expresses a strong matchup effect. The interaction was trained in isolation and frozen, which may limit its usefulness after a skill filter is introduced. The supported negative result is about this recipe and budget; it cannot exclude better jointly fitted or time-varying interactions.

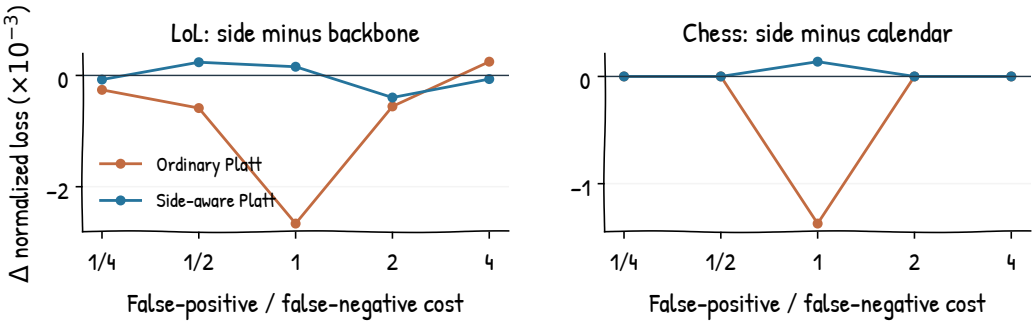


Figure 3: **One fixed set of decision costs.** Side model minus backbone in canonical blue/white coordinates; lower is better. These are descriptive point estimates, with the five cost ratios fixed before the new comparisons. No threshold is chosen by test performance.

Figure 2 shows why the macro result should not be read as a uniform per-period law. Side-aware differences vary across years, and the interaction has negligible effect in some selected folds. All annual and weekly contrasts are retained in Appendix D, rather than replacing the macro comparison with whichever period supports the preferred interpretation.

5.4 A DECISION CHECK AT FIXED COSTS

We evaluate the same forecasts in canonical coordinates at false-positive to false-negative cost ratios $c \in \{1/4, 1/2, 1, 2, 4\}$. Predicting a canonical-side win incurs cost c when wrong; predicting a loss incurs cost 1 when wrong. The decision is $\mathbf{1}[p_C > c/(1+c)]$, with realized cost divided by $\max(1, c)$. Figure 3 reports macro differences for the side extension. The matched-information control does not imply identical performance at every threshold. These five descriptive estimates do not establish an operational benefit or universal decision dominance; they prevent the accuracy result at $c = 1$ from standing in for an unspecified deployment objective.

6 SCOPE AND IMPLICATIONS

The practical lesson is to specify both the event and the information available to recalibration before attributing a residual gain to a contextual model. For these archives, a blue/white side covariate is available before the outcome and costs one coefficient to expose at the readout. The resulting control changes the conclusion of the ordinary scalar comparison without retraining or reordering any fixed forecast stream. A useful default for pairwise evaluation is therefore to compare predictors under a declared common context set, check swap coherence, separate selection from calibration, and report proper scores alongside any claimed decision objective.

The limitations matter to the interpretation. Both domains contain one binary side variable; this study does not test rich lineup, patch, or player-context interactions. The chess archive spans only three months, is systematically subsampled, and excludes draws, so its estimand differs from unconditional game outcomes. Its short rolling windows are not directly comparable to annual LoL windows. The seven years and nine weeks share histories, and the conditional bootstrap intervals omit parameter-selection uncertainty. Historical results informed research choices, so the protocol is a retrospective extension rather than a preregistration. The frozen interaction is one nonconvex recipe with one seed; increasing its iteration budget is not a global-optimality certificate. Finally, a failure to demonstrate a positive advantage is not an equivalence claim, particularly for the LoL side effect whose interval crosses zero.

AI USE STATEMENT

AI assistants were used for research ideation and execution, including code inspection, implementation, experiment orchestration, figure generation, adversarial critique, and analysis. They were also used to draft and polish sections of the manuscript and to develop and check the mathematical argument, including proof development. These uses correspond to assistance with writing, research ideation or execution, drafting, and proving mathematical claims. AI-generated proposals were checked against executable results and source materials; this does not constitute an independent human review. Responsibility for the submission remains with the authors.

ETHICS STATEMENT

The study uses existing public competitive-game records. Lichess releases its database under CC0. Player identifiers serve only to connect prior observations; we do not infer personal attributes or publish individual skill assessments. The result concerns evaluation design and does not validate a betting, matchmaking, or other deployed decision system. The artifact reports aggregate findings and avoids adding private information to the public records.

REPRODUCIBILITY STATEMENT

The anonymous artifact contains the numerical kernel and Python reference, unchanged legacy source snapshots, selected per-match prediction streams, readout parameters, calibration inputs for the new experiments, full selection grids, optimizer reports, data hashes, generated tables, and figure code. All headline scores and intervals can be recomputed from the saved streams without downloading the original game dumps. Replaying model selection requires the specified input tables; acquisition and preprocessing instructions distinguish this from replaying the saved predictions. Appendix B gives the updates and Appendix D supplies the full temporal partitions and sensitivity results. The review notes document unresolved limitations rather than certifying acceptance or an independent external review.

REFERENCES

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3–4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324.

Shuo Chen and Thorsten Joachims. Modeling intransitivity in matchup and comparison data. In *Proceedings of WSDM*, pp. 227–236, 2016a. doi: 10.1145/2835776.2835787.

Shuo Chen and Thorsten Joachims. Predicting matchups and preferences in context. In *Proceedings of KDD*, pp. 775–784, 2016b. doi: 10.1145/2939672.2939764.

Pierre Dangauthier, Ralf Herbrich, Tom Minka, and Thore Graepel. TrueSkill through time: Revisiting the history of chess. In *Advances in Neural Information Processing Systems*, volume 20, 2007.

Arpad E. Elo. *The Rating of Chessplayers, Past and Present*. Arco Publishing, 1978.

Mark E. Glickman. Parameter estimation in large dynamic paired comparison experiments. *Journal of the Royal Statistical Society: Series C*, 48(3):377–394, 1999.

Mark E. Glickman. Example of the Glicko-2 system. https://www.glicko.net/glicko/glicko2.pdf, 2022. March 22 revision.

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

Parikshit Gopalan, Adam Tauman Kalai, Omer Reingold, Vatsal Sharan, and Udi Wieder. Omnipredictors. In *Proceedings of ITCS*, 2022. doi: 10.4230/LIPIcs.ITCS.2022.79. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2022.79.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of ICML*, volume 70, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of ICML*, volume 80, pp. 1939–1948, 2018. URL https://proceedings.mlr.press/v80/hebert-johnson18a.html.

Ralf Herbrich, Tom Minka, and Thore Graepel. TrueSkill: A Bayesian skill rating system. In *Advances in Neural Information Processing Systems*, volume 19, 2006.

Meelis Kull, Telmo Silva Filho, and Peter Flach. Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Proceedings of AISTATS*, volume 54, pp. 623–631, 2017. URL https://proceedings.mlr.press/v54/kull17a.html.

Hans R. K"unsch. The jackknife and the bootstrap for general stationary observations. *The Annals of Statistics*, 17(3):1217–1241, 1989.

Lichess. Lichess open database. https://database.lichess.org/, 2026. Standard rated game exports, March–May 2026; CC0.

Oracle's Elixir. League of legends esports match data. https://oracleselixir.com/, 2025. Historical public match-data snapshot.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas B. Sch"on. Evaluating model calibration in classification. In *Proceedings of AISTATS*, volume 89, pp. 3459–3467, 2019. URL https://proceedings.mlr.press/v89/vaicenavicius19a.html.

Shengjia Zhao, Michael P. Kim, Roshni Sahoo, Tengyu Ma, and Stefano Ermon. Calibrating predictions to decisions: A novel approach to multi-class calibration. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://arxiv.org/abs/2107.05719.

A Orientation equivalence and fitting details

For a sample with $s = +1$, canonical and recorded coordinates are identical. For $s = -1$, they are $(z_C, y_C) = (-z, 1 - y)$. If $q_C = \sigma(az_C + b)$, mapping back gives

$$q = \begin{cases} q_C & s = +1, \\ 1 - q_C & s = -1, \end{cases} = \sigma(az + bs).$$

For binary log loss $\ell(q, y) = -y \log q - (1 - y) \log(1 - q)$, complementation preserves the loss exactly. Consequently the summed objectives agree pointwise in (a, b), as does any identical penalty. Their feasible sets are also the same under $a \geq 10^{-8}$, so their minimizer sets correspond even if the unpenalized problem has multiple optima. With our strictly convex quadratic penalty the coefficient optimum is unique. Symmetric probability clipping preserves the complementation identity.

For the side-aware readout, swapping gives $\sigma(-az - bs) = 1 - \sigma(az + bs)$. For an ordinary intercept, equality for every z requires $\sigma(b) = 1 - \sigma(b)$ at $z = 0$, hence $b = 0$; this is also sufficient. These statements describe the readout's algebra. For an incoherent base model, defining a canonical score as sz does not establish coherence of the underlying model under a separately rerun swapped history.

Each readout is fitted by L-BFGS-B with initial coefficients $(1, 0)$, `ftol`$= 10^{-13}$, `gtol`$= 10^{-8}$, and at most 1,000 iterations. Fitting uses only the calibration slice. We do not fit side-aware and ordinary readouts to different sets of examples or choose between them on test labels. The two-coefficient comparison changes the conditioning variable and imposes coherence on the side-aware map; it is not a nested-model likelihood-ratio test.

B Numerical model and temporal protocol

B.1 Calendar update

Let $q_0 = 400/\log 10$. Every competitor starts with $\mu = 0$, $\phi = 350/q_0$, and $\sigma = 0.06$; the side coefficient starts at zero. For elapsed UTC days Δd and selected period scale r,

$$\phi \leftarrow \min\left\{350/q_0, \sqrt{\phi^2 + \sigma^2 \Delta d / r}\right\}.$$

The elapsed interval is applied once, when the competitor is next encountered. The public prediction is Eq. (3). During feedback, take a snapshot of both competitors and of β. For competitor i, with outcome y,

$$e_i = \sigma(g(\phi_j)(\mu_i - \mu_j) + \beta s + h_{ij}), \qquad v_i = \{g(\phi_j)^2 e_i(1 - e_i)\}^{-1}, \qquad \Delta_i = v_i g(\phi_j)(y - e_i).$$

The inverse-variance term is floored at 10^{-12} for numerical stability. Following the Glicko-2 volatility root construction (Glickman, 2022), solve with $\tau = 0.5$ and $A_0 = \log \sigma_i^2$:

$$f(x) = \frac{e^x(\Delta_i^2 - \phi_i^2 - v_i - e^x)}{2(\phi_i^2 + v_i + e^x)^2} - \frac{x - A_0}{\tau^2} = 0, \qquad \sigma_i' = e^{x/2}.$$

The bracket tolerance is 10^{-6}. The updated deviation and mean are

$$\phi_i' = (\phi_i^{-2} + v_i^{-1})^{-1/2}, \qquad \mu_i' = \mu_i + \phi_i'^2 g(\phi_j)(y - e_i).$$

The second competitor uses outcome $1 - y$, context $-\beta s - h_{ij}$, and the same pre-update snapshot. The new volatility affects subsequent calendar inflation; an additional event-based inflation is not applied after the root solve. This is the specific adaptation evaluated, rather than standard rating-period Glicko-2. Updating β uses the public probability recomputed at feedback, before either skill update. Within-day feedback can affect later state updates but cannot change the already-issued predictions for that day.

The Elo sensitivity uses public logit $\mu_i - \mu_j + \beta s$ in these same units, then updates $\mu_i \leftarrow \mu_i + (K/q_0)(y - p)$ and $\mu_j \leftarrow \mu_j - (K/q_0)(y - p)$. It shares the side update and daily prediction-before-feedback ordering.

B.2 FROZEN INTERACTION, GRIDS, AND CONVERGENCE

The unchanged interaction routine solves

$$\min_{u,v} \sum_{(i,j,y)\in\mathcal{T}} \left[\log(1+e^{h_{ij}}) - y h_{ij}\right] + \frac{\lambda}{2}\sum_{i}(\|u_i\|_2^2 + \|v_i\|_2^2), \qquad h_{ij} = u_i^\top v_j - u_j^\top v_i.$$

It starts from independent $N(0, 0.1^2)$ coordinates with seed zero, with identifiers sorted before initialization. Its fitting data end before the tuning period. The routine has no skill, side, or calendar term in its training objective. Once fitted, the embeddings do not change on tuning, calibration, or test data; the online skill and side states do. An offset is zero if either identifier is absent from the interaction training set.

The four arms share period scale r in the primary configuration. Side and both share η; interaction and both share λ. The common configuration minimizes the mean of their four tuning losses over the 60 triples. Because the base ignores two coordinates and the single-addition arms ignore one, counting each duplicate triple as an independent model search would be misleading. The independent-selection sensitivity uses all 5, 15, 20, and 60 distinct configurations. In every fold, its minimizers coincide with the respective active coordinates of the common minimizer.

All seven folds select $r = 3$ and $\eta = 0.01$. The selected penalties for 2018–2024 are $(20, 100, 100, 20, 100, 20, 20)$. Under the 300-iteration cap, 15 of 28 embedding fits report convergence, including every selected fit. The 1,200-iteration repeat increases this to 23 of 28; the five remaining capped fits have $\lambda = 1$ in 2020–2024 and are never selected. All selected configurations, test arrays, and fitted readout coefficients are identical between caps. This budget check was added after inspecting the original convergence reports and aggregates, and is an exploratory sensitivity rather than a prespecified result. Per-fit objectives, gradient norms, and termination messages are retained in the artifact.

B.3 DATA PROVENANCE AND PARTITION SIZES

The LoL cleaner aggregates public team records, uses deterministic random recording order, and carries the observed blue-side flag. The chess cleaner parses the three specified monthly Lichess dumps, filters to `180+0`, requires parseable timestamps and numeric rating headers, and retains decisive or drawn results before sampling. It sorts by time and takes every $\lfloor N/2{,}000{,}000 \rfloor$-th row. Its recorded ordering uses a seed-42 pool of 2^{20} swap flags, repeated as needed. This is deterministic pseudorandom recording order, not an assumption that every orientation draw is independent. The stored white-side flag, rather than that generator, is used by our analysis. Dropping draws occurs after upstream sampling, and all subsequent training, tuning, calibration, and evaluation condition on decisive outcomes.

The existing tables omit original match/game identifiers. Consequently, time/canonical-competitor key checks are conservative collision checks, not proof that every repeated key is a duplicate physical game. We remove those collisions consistently in the new LoL experiment. The chess table has none. We do not infer outcome availability within the day from recorded start times; all feedback is delayed until that day's prediction batch is complete.

C HISTORICAL RECONSTRUCTION AND IMPLEMENTATION BOUNDARIES

The legacy reconstruction uses the original validation-selected settings and event-by-event feedback, then fits readouts on the complete validation year. Hyperparameter selection and calibration therefore share outcomes, unlike the new split protocol. The historical P2X extension contains side and a frozen rank-five interaction; it is not a side-only arm. The fixed-stream intervention tests its aggregate contextual advantage without attributing it entirely to either component. Table 4 preserves these results as a reconstruction, not as directly comparable scores for the new implementation.

The archive's public prediction attenuates only by the recorded opponent's deviation. When the two deviations differ, this violates swap complementarity; maximum observed same-state errors are 0.08435 for P2 and 0.10024 for P2X. It also updates the second competitor after modifying the first, and the volatility root uses squared rating mean where the derivation requires squared rating

Table 3: Temporal partition sizes. The first column is a LoL evaluation year or the first date of a seven-day chess evaluation window. Training is the cumulative prefix; the four slices do not overlap within a fold.

Test fold	Training	Tuning	Calibration	Evaluation
2018	6,947	3,641	1,863	6,732
2019	12,451	3,831	2,901	8,125
2020	19,183	4,821	3,304	9,740
2021	27,308	5,340	4,400	12,263
2022	37,048	7,669	4,594	12,519
2023	49,311	7,591	4,928	11,097
2024	61,830	7,102	3,995	10,194
2026-03-30	309,769	145,058	146,426	147,852
2026-04-06	454,827	146,426	147,852	151,106
2026-04-13	601,253	147,852	151,106	149,523
2026-04-20	749,105	151,106	149,523	151,158
2026-04-27	900,211	149,523	151,158	148,183
2026-05-04	1,049,734	151,158	148,183	148,146
2026-05-11	1,200,892	148,183	148,146	149,283
2026-05-18	1,349,075	148,146	149,283	145,891
2026-05-25	1,497,221	149,283	145,891	149,633

Table 4: Historical fixed-stream control, macro means over 70,720 matches. Accuracy is in percent. These absolute values use a different feedback and selection protocol from Table 1.

	Log loss		Accuracy (%)	
Model	Ordinary	Side-aware	Ordinary	Side-aware
Elo	0.641443	0.639538	62.781	63.024
P2	0.640879	0.639181	62.834	63.105
P2X	0.639485	0.639272	63.162	63.140

deviation. The new filter uses symmetric public attenuation, snapshot updates, and the corrected volatility expression. Other archived evaluation issues include a smoothed routine labeled isotonic regression and an uncentered bootstrap tail described as a p-value. Neither procedure is used for the new results. We preserve the old source for traceability rather than silently changing the procedure whose output is being reconstructed.

For the ordinary-Platt historical contrast, the 30-day interval for P2X minus P2 log loss is $[-0.001872, -0.000888]$ and for accuracy is $[0.134, 0.515]$ pp. Matched side-aware intervals appear in Table 2. The mathematical coordinate identity can be applied to these stored arrays despite the old model's lack of swap coherence; it does not repair its internal predictions.

D Additional scores, heterogeneity, and uncertainty

D.1 Nested-readout sensitivity

The primary two-parameter maps are not nested: the side-aware map replaces the ordinary intercept and imposes swap coherence. During manuscript review, after the primary results were known, we added the exploratory readout

$$q_{\mathrm{nested}}(z, s) = \sigma(az + b_0 + b_s s), \qquad a \geq 10^{-8}.$$

It contains both primary classes and tests whether retaining the ordinary intercept changes the conclusion. We use the same calibration slices, optimizer, and summed-log-loss penalty, now centered at $(1, 0, 0)$ with coefficient 10^{-3}. The rating streams are unchanged. Table 5 reports every tested contrast under this readout. Side does not gain a macro proper-score advantage in either domain, and neither LoL interaction contrast becomes favorable. The result therefore does not depend on excluding an ordinary intercept from the side-aware map. This is a post-analysis sensitivity, not ad-

ditional prespecified confirmation; the main two-parameter control retains equal coefficient counts and an exact canonical-coordinate interpretation.

Table 5: Exploratory three-parameter readout, paired differences and conditional 95% intervals. Log loss is multiplied by 10^3; accuracy is in percentage points. Blocks are 30 days for LoL and three days for chess.

Contrast	$10^3\Delta$ log loss	Δ accuracy (pp)
LoL: side − base	$+0.139$ $[-0.088, +0.324]$	-0.163 $[-0.291, -0.018]$
LoL: interaction − base	$+0.139$ $[+0.016, +0.262]$	-0.026 $[-0.075, +0.019]$
LoL: both − side	$+0.136$ $[+0.017, +0.257]$	-0.037 $[-0.093, +0.013]$
LoL: both − base	$+0.275$ $[+0.008, +0.509]$	-0.200 $[-0.337, -0.056]$
Chess: side − calendar	$+0.013$ $[+0.004, +0.024]$	-0.001 $[-0.028, +0.028]$
Chess: side − Elo	$+0.134$ $[+0.085, +0.186]$	-0.033 $[-0.082, +0.013]$

D.2 Raw scores and complete temporal contrasts

Table 6 reports raw scores and recalibrated Brier scores. The calendar model is markedly overconfident in this subsampled chess setting before readout fitting. Its large raw log loss is a limitation of the specified rating recipe on sparse histories, not an omitted successful baseline. The paired comparisons use a common recipe and the same calibration windows. Site-provided ratings are deliberately outside the input contract, so these experiments should not be read as chess prediction benchmarks.

Table 6: Raw and Brier scores; macro means. Lower is better in every column.

		Raw log loss	Brier score		
Data	Model		Raw	Ordinary Platt	Side-aware Platt
LoL	Base	0.650716	0.228687	0.226368	0.225654
	Side	0.649872	0.228263	0.225850	0.225740
	Interaction	0.650913	0.228776	0.226433	0.225719
	Both	0.650064	0.228348	0.225912	0.225803
Chess	Calendar	0.748866	0.272343	0.249363	0.249062
	Calendar + side	0.749097	0.272338	0.249304	0.249069
	Elo	0.694239	0.250546	0.249336	0.249036
	Elo + side	0.694315	0.250574	0.249225	0.249102

For each fold, the bootstrap constructs a sequence of calendar days spanning the entire evaluation window, including days without games. We sample circular block starts uniformly and concatenate blocks to the original number of days, truncating the last block if needed. Loss-difference sums and match counts are resampled together, yielding a per-match mean within each fold; these fold means are then averaged. The random seed is fixed and recorded in code. Resampling complete chess weeks instead treats the nine weekly differences as the sampling units. This sensitivity also conditions on fitted predictions and cannot account for overlapping training histories or common shocks beyond the observed window. No p-values are inferred from uncentered bootstrap tails.

Table 7: Every annual and weekly side-minus-backbone contrast. Log loss is multiplied by 10^3 and accuracy is in percentage points. The LoL backbone is base and the chess backbone is calendar.

	Ordinary Platt		Side-aware Platt	
Test fold	$10^3\Delta$ log loss	Δ accuracy	$10^3\Delta$ log loss	Δ accuracy
2018	-2.013	+0.520	-0.323	+0.282
2019	-1.151	+0.234	+0.408	-0.677
2020	-0.846	+0.154	+0.267	-0.175
2021	-0.845	-0.024	+0.187	-0.049
2022	-0.661	+0.511	-0.002	+0.152
2023	-1.373	+0.460	+0.316	-0.045
2024	-1.116	+0.010	+0.121	+0.402
2026-03-30	-0.149	+0.294	-0.000	+0.022
2026-04-06	-0.109	+0.164	+0.002	+0.032
2026-04-13	-0.101	+0.124	-0.007	+0.064
2026-04-20	+0.025	+0.005	+0.028	-0.107
2026-04-27	-0.147	+0.159	+0.026	-0.025
2026-05-04	-0.085	+0.111	+0.019	-0.043
2026-05-11	-0.136	+0.108	+0.052	-0.033
2026-05-18	-0.152	+0.124	+0.007	-0.024
2026-05-25	-0.213	+0.146	-0.006	-0.011

Table 8: Block-length sensitivity of paired 95% log-loss intervals. All values are multiplied by 10^3. Whole-week resampling samples nine complete weekly differences with replacement. None of these intervals includes model-fitting or hyperparameter-selection uncertainty.

Contrast	Block	Ordinary Platt	Side-aware Platt
LoL: side − base	7	-1.144 $[-1.557, -0.730]$	$+0.139$ $[-0.105, +0.372]$
LoL: side − base	30	-1.144 $[-1.562, -0.724]$	$+0.139$ $[-0.087, +0.325]$
LoL: side − base	90	-1.144 $[-1.559, -0.728]$	$+0.139$ $[-0.070, +0.288]$
LoL: interaction − base	7	$+0.142$ $[+0.038, +0.250]$	$+0.145$ $[+0.040, +0.253]$
LoL: interaction − base	30	$+0.142$ $[+0.020, +0.265]$	$+0.145$ $[+0.023, +0.267]$
LoL: interaction − base	90	$+0.142$ $[+0.031, +0.280]$	$+0.145$ $[+0.033, +0.282]$
LoL: both − side	7	$+0.139$ $[+0.034, +0.247]$	$+0.142$ $[+0.039, +0.248]$
LoL: both − side	30	$+0.139$ $[+0.017, +0.262]$	$+0.142$ $[+0.023, +0.262]$
LoL: both − side	90	$+0.139$ $[+0.028, +0.277]$	$+0.142$ $[+0.034, +0.277]$
Chess: side − calendar	1	-0.119 $[-0.156, -0.082]$	$+0.013$ $[+0.003, +0.025]$
Chess: side − calendar	2	-0.119 $[-0.153, -0.084]$	$+0.013$ $[+0.003, +0.025]$
Chess: side − calendar	3	-0.119 $[-0.151, -0.086]$	$+0.013$ $[+0.004, +0.024]$
Chess: side − calendar	whole weeks	-0.119 $[-0.155, -0.074]$	$+0.013$ $[+0.002, +0.027]$
Chess: side − Elo	1	-0.223 $[-0.328, -0.113]$	$+0.134$ $[+0.086, +0.186]$
Chess: side − Elo	2	-0.223 $[-0.327, -0.118]$	$+0.134$ $[+0.085, +0.186]$
Chess: side − Elo	3	-0.223 $[-0.316, -0.125]$	$+0.134$ $[+0.085, +0.186]$
Chess: side − Elo	whole weeks	-0.223 $[-0.311, -0.126]$	$+0.134$ $[+0.075, +0.206]$

E ARTIFACT VERIFICATION AND REUSE

The artifact's verification scripts compare the compiled implementation with a Python reference within 1.5×10^{-12} on 3,000 events, check complementarity under mirrored histories with nonzero interaction offsets, and verify that changing same-day or future outcomes does not affect already-issued predictions. Zero interaction reduces exactly to the corresponding no-interaction path. Selected-model replay reconstructs all 1,411,445 new evaluation examples, with maximum probability discrepancy below 10^{-9} when an alternative summation order is used for interaction offsets. Refitting the saved readouts reproduces transformed probabilities within 5×10^{-10}. Independently evaluating log loss by the direct probability formula rather than the analysis code's logit formula reproduces all reported fold metrics within 2.3×10^{-16}.

The `results/` directory includes raw test predictions, labels, observed side, timestamps, selected parameters, full tuning grids, fitted maps, and optimizer diagnostics. Calibration arrays for the new experiments allow the two readouts to be refitted without replaying the ratings. Per-day paired loss sums support a lightweight bootstrap replay. The `scripts/` directory contains the analysis, numerical kernel, data contracts, and figure generator; `audit/` contains the protocol and review notes. Input-table hashes pin the snapshots used; preprocessing a later public dump is not silently assumed to produce the same table. The supplementary package does not require an external service for the fixed-stream analysis.

The hand-drawn appearance uses an openly licensed handwriting font and mildly sketched frame decoration. Data markers, lines, and table entries use the unmodified computed values. Source snapshots are retained to distinguish the historical implementation from the corrected experiments. Review notes are internal adversarial checks, not external peer-review outcomes.