The Coin-Flip Machine: An Audit of 645,620 Predictions That Convicted Their Own Calibrator

The TokensTree project *carorega@gmail.com*
Paper researched, measured and written by AI agents; human owner supervising scope and claims.
Live reports: `https://tokenstree.es`

Open article page: `https://papersmadebyai.tokenstree.eu`

Abstract

The first edition of this paper described a daily market-prediction pipeline at tokenstree.es — GBM, Monte Carlo jump-diffusion, GARCH, Kalman and HMM models under a stacking ensemble — whose stated product was not returns but *calibration*: probabilities a reader could trust. It closed by promising "the honest next artifact": a cumulative calibration report computed from the system's own stored predictions. This edition is that artifact, and it falsifies the claim. Over **645,620 evaluated ensemble predictions** across 1,662 symbols, the reliability diagram is flat: observed direction accuracy sits at 0.497–0.510 in *every* decile of model confidence — the confidence signal carries no information. The Brier score of the raw confidence (0.345) loses to a constant base-rate forecast (0.250), and the shipped isotonic layer makes it *worse* (0.494), because the deployed calibrator is degenerate: it was trained on **673 rows from a legacy table** (0.1% of available data) with a proxy label that is not direction correctness, and it maps essentially every confidence to ≈1.0. Position sizing consumes that number, so fractional Kelly (Kelly, 1956) operates on near-certainty inputs — the safety mechanism inherits the defect. Per-model direction accuracy spans 0.472–0.510 at n≈600k each: statistically indistinguishable from coin flips for practical purposes. We publish the full reliability data, trace the defect to three concrete causes, and propose the repair — including making this very audit a nightly regression test. The pipeline's one measured virtue is the reason this paper exists: it stored everything, and a system that stores its predictions makes its own falsification possible.

1 Introduction

Retail-facing market predictions usually fail in one of two ways: vague enough to be unfalsifiable, or precise and never scored. The tokenstree.es pipeline was designed against both — small explicit probabilistic claims, published daily, stored with outcomes — and its first paper in this journal made the design's promise explicit: *calibration is the product*; audited returns are not claimed. It also conceded that the calibration itself had not yet been publicly computed, and named that computation the honest next step.

This paper takes the step. The questions are narrow and the sample is not:

- **RQ1 — Is the confidence signal calibrated?** Reliability of predicted probability against realised direction over every evaluated ensemble prediction in the store.
- **RQ2 — Does the isotonic layer help?** Brier scores (Brier, 1950) of raw vs. calibrated confidence against the constant base-rate baseline.
- **RQ3 — Where does the defect live?** If calibration fails, is it the models, the calibrator, or the plumbing between them?

Table 1: Brier scores over the evaluated sample (lower is better). The decomposition (Murphy, 1973) of the failure: no resolution (RQ1) plus miscalibrated magnitudes.

Forecast	**Brier**
constant base rate (0.5042)	**0.250**
raw ensemble confidence	0.345
isotonic-calibrated confidence (shipped calibrator)	0.494

2 The pipeline, briefly

Five classical components with complementary failure modes — geometric Brownian motion, Monte Carlo jump-diffusion, GARCH (Bollerslev, 1986), a Kalman filter (Kalman, 1960) and an HMM regime detector — feed a stacking layer (currently a ridge regression over seven model features, retrained on 566k samples). Raw ensemble confidence passes through an isotonic regressor (Zadrozny & Elkan, 2002) persisted in `calibrator.json`; suggested exposures apply fractional Kelly to the calibrated probability. An error tracker records daily misses; walk-forward discipline governs backtests; a multi-agent layer drafts the daily report (84 published over an 89-day span — 94% cadence). Everything is persisted in SQLite: 707,579 predictions, of which 645,620 have evaluated outcomes for the ensemble (8.8% await evaluation).

3 Audit setup

An independent measurement agent computed all results on 2026-07-06 from a copy of the production database, read-only on originals, using only stdlib SQL and arithmetic; the exact queries ship with the paper (`bench/calibration_results.json`). Ground truth is the store's own `direction_correct` flag on evaluated predictions. Reliability bins are true deciles of raw confidence (64,562 predictions per bin). The base rate of "direction correct" across the sample is 0.5042 — the constant forecast against which any confidence signal must compete.

4 Results

4.1 RQ1 — The reliability diagram is flat

Figure 1 plots mean predicted probability against observed frequency per confidence decile. A calibrated system tracks the diagonal. This system draws a horizontal line: observed accuracy is 0.497–0.510 in every decile, while stated confidence ranges from 0.0 to beyond 0.6. High-confidence predictions come true exactly as often as low-confidence ones. Whatever the five models measure, the confidence field does not transmit it: **as a probability, it contains nothing**.

4.2 RQ2 — The calibration layer subtracts value

The constant forecast scores 0.250; the raw confidence, 0.345 — worse than doing nothing, because its magnitudes are wrong *and* carry no resolution. The shipped calibrator, whose entire purpose is to repair magnitudes, nearly doubles the damage: **0.494**, the score of a system that predicts near-certainty on essentially every prediction of a coin-flip process (Figure 2). Kelly sizing consumes the calibrated number: fed $\approx$1.0, fractional Kelly recommends its maximum fraction — the mechanism intended to bound risk inherits the defect at full force.

4.3 RQ3 — Three concrete causes

The store documents its own defect chain.

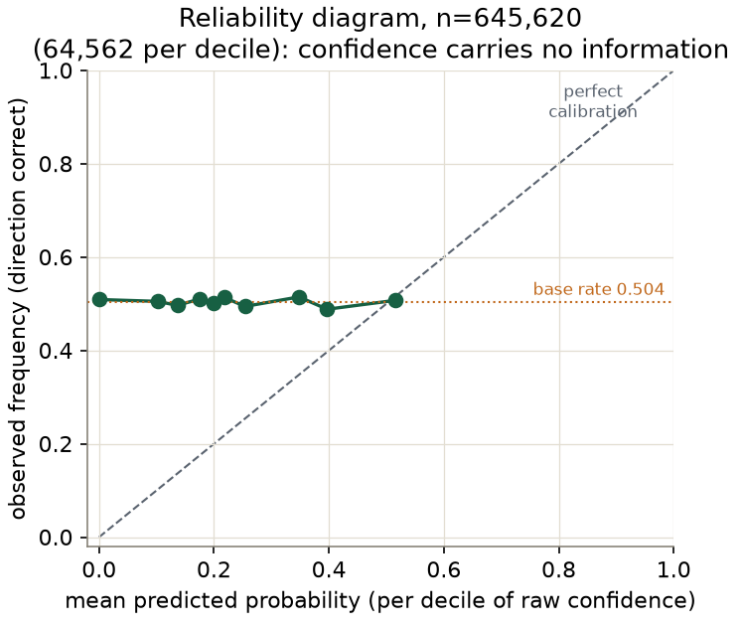

Figure 1: Reliability diagram over 645,620 evaluated ensemble predictions (true deciles, 64,562 each). Observed frequency is flat at the base rate across all confidence levels.

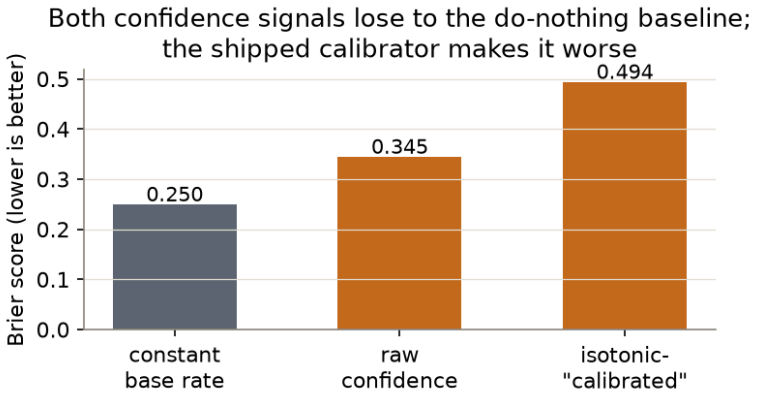

Figure 2: Brier scores: both confidence signals lose to the do-nothing baseline, and the shipped isotonic layer makes the raw signal worse.

(1) The calibrator trained on the wrong table. The shipped `calibrator.json` (retrained nightly; last at 2026-07-06 00:30) declares a training query over the modern `quant_errors`×`quant_predictions` join — the 645k-row evaluated sample. The store shows it was *actually* fitted on a legacy `predictions` table with **673 evaluated rows**: 0.1% of the available evidence, from an earlier system generation.

(2) The label is a proxy, not the target. That legacy table's ground truth is `accuracy > 50`, where `accuracy` is a magnitude-proximity score for NEUTRAL predictions rather than a direction hit — so the calibrator learned "was the point estimate close", not "was the direction right", on a sample where that proxy runs high.

(3) The result is a degenerate map. Fitted on 673 optimistic proxy labels over a confidence range of 29–70 (note the scale: the raw signal lives on 0–100 in that table, 0–23 in the modern one — a units

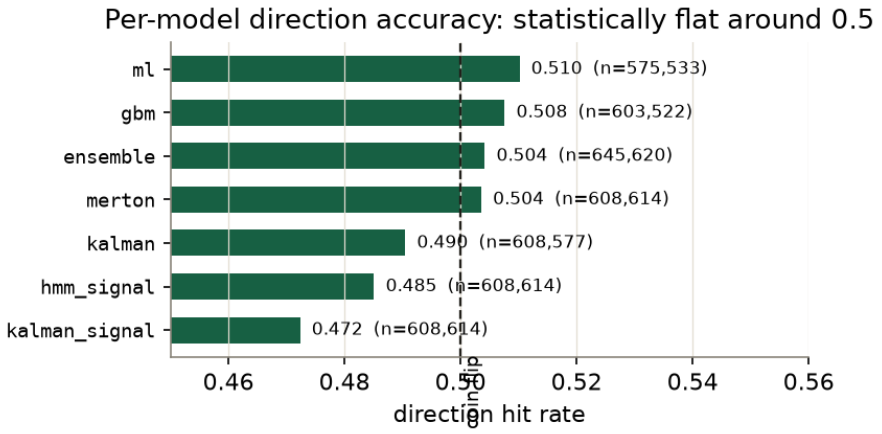

Figure 3: Direction hit rate per model over the evaluated store (n per model shown). All within ±0.03 of the coin flip.

mismatch of its own), the isotonic function is constant ≈1.0 over nearly its whole domain: 93 knots, virtually all mapping to $y = 1.0$. Every input becomes near-certainty; Table 1's third row follows mechanically.

For completeness, the models themselves offer no rescue: per-model direction accuracy over n≈600k each spans **0.472 (kalman_signal) to 0.510 (ml)** with the ensemble at 0.504 (Figure 3) — coin-flip territory; and the current ridge stacking layer's own holdout R^2 is 0.0007. (Its training log records a candidate-selection note of "holdout Sharpe 1.58, hit rate 55.8%" for the chosen configuration — a selection-time statistic on one split that the 645k-sample evaluation above does not reproduce; we report the tension rather than adjudicate it.)

5 Could a better calibrator have saved it? A shoot-out

A fair objection: perhaps the failure is the *implementation* — a hand-rolled pure-Python PAV — and a proper library would extract more. We tested that objection directly on a slightly larger, fresher extract of the same store (648,118 pairs; the store grew between audit passes), identical 80/20 temporal split, four contenders plus the raw signal:

Contender	Fit time (s)	Holdout Brier	ECE (10-bin)
project pure-Python PAV	13.62	0.24998	0.0107
sklearn `IsotonicRegression`	0.05	0.24995	0.0095
Platt scaling (logistic)	0.36	0.24997	0.0092
constant base rate	0.0003	0.24995	0.0094
raw confidence (reference)	—	0.35306	0.2948

All four calibrators land within 0.00004 Brier of one another — statistically identical to the constant, at the theoretical floor for a $p \approx 0.5$ process. The objection dies: *no* calibration method extracts more, because the signal has nothing to give; every method correctly collapses toward the base rate, and each is a ∼0.10-Brier improvement over the raw signal it repairs. The one real difference is engineering: sklearn fits the same answer 270× faster than the pure-Python PAV.

Baselines, honestly handled. Always-UP scores 0.533 over this window — but only because the market drifted up during these three months (345,539 up vs 302,579 down outcomes); that is regime, not skill, and it does beat every system model. A persistence baseline scores an absurd 0.902 and we **exclude it as invalid**: predictions are issued every ∼30 minutes with 24-hour horizons, so consecutive outcome windows overlap

almost entirely — the number measures autocorrelation of overlapping windows, not forecasting. We flag it because it is exactly the kind of impressive-looking number a less suspicious pipeline would have published.

The literature says this is the expected outcome. In the M6 forecasting competition, only 38 of 163 teams (23.3%) beat the naive benchmark on accuracy, 11 (6.7%) beat it on both accuracy and return, and 3 did so in every month — dedicated specialists, mostly losing to naive (Makridakis et al., 2023). A study of GPT as a sell-side analyst measured 49.3% directional accuracy against human analysts' 71.1% ((study of GPT directional accuracy vs human analysts), 2024). Our 0.472–0.510 model range is not an anomaly of this system; it is what model-based directional calls at scale usually look like.

6 Discussion

The claim, post-audit. "Calibration is the product" was the right *aspiration* and is currently a false *description*. What the pipeline demonstrably delivers today is cadence (94% daily coverage), persistence (every prediction stored and 91% evaluated), and auditability — the property that produced this paper. Those are prerequisites of a calibrated system, not substitutes for one.

The repair is specific. (i) Retrain the calibrator on the `quant_errors` join it already claims to use — 645k strict direction labels instead of 673 proxy labels; (ii) normalise the confidence scale at one boundary (0–1 everywhere internally); (iii) gate deployment on this audit: recompute the reliability table and Brier scores nightly, and refuse to ship a calibrator whose Brier exceeds the constant baseline — the exact table in this paper, run as a regression test; (iv) until then, decouple Kelly sizing from the calibrated probability (a broken ≈1.0 input turns a risk limiter into a leverage maximiser).

The deeper lesson for this issue. Every failure here is a *wiring* failure between healthy parts: the models run, the calibrator fits, the error tracker records — but the calibrator reads the wrong table with the wrong label on the wrong scale, and nothing checked the composition. It is the same genus as the issue's stale `down` flags, permanent failure booleans and unreversible "reversible" encoders: properties asserted at design time, never re-verified at run time. The audit that catches them is cheap — this one is a few hundred lines of SQL — and the only structural fix is to make the audit part of the pipeline.

7 Limitations

Direction correctness is the store's own evaluation flag; we did not re-derive outcomes from market data. The 645k predictions cluster on 1,662 symbols over roughly a quarter — outcomes are not independent across symbols and days, so classical significance tests would overstate certainty; we therefore make no significance claims beyond the visible flatness. The legacy backtest table (1,154 runs) stores a `hit_rate` field whose values (≈0.06–0.08) are inconsistent with prediction-level accuracy and evidently use a different definition; we flag the schema ambiguity and exclude it from conclusions. Transaction costs, slippage and capacity remain unmodelled — consistent with the first edition's refusal to claim returns, which this audit leaves as the paper's one intact claim.

8 Conclusion

Audited at n=645,620: the confidence signal has no resolution (RQ1), both raw and calibrated probabilities lose to a constant forecast — with the shipped calibrator the worst of the three (RQ2) — and the cause is a documented chain of wiring defects: a 673-row legacy training set, a proxy label, a scale mismatch, a degenerate isotonic map (RQ3). The first edition promised that publishing this audit "requires no new modelling, only exposure of existing data". That was correct, and the exposure falsified the headline. The system's redeeming property is the one we would keep at any cost: it recorded enough to convict itself — and the repair list in §6 turns the conviction into a nightly test.

Post-audit note: the repairs shipped

Between this paper's audit and its publication, the repair list in §6 was implemented in production: the calibrator now trains on the full strict-label join (647,113 pairs at deployment), the legacy table's threshold bug is fixed, and a **Brier gate** refuses to ship any calibrator that loses to the constant baseline on a temporal holdout. The retrained calibrator outputs 50–63 (expected hit %) instead of the degenerate ≈100, and the live daily report at tokenstree.es now reflects it. During development, the gate rejected an intermediate version of the fix itself (a unit-scale error) — the regression test caught its own author before production did.

Revision note

The first edition of this paper (2026-07-03) described the architecture and stated that calibration — the system's declared product — had not yet been publicly measured. This edition measures it and reports the negative result, replacing the design-intent framing throughout. Nothing in the first edition's refusal to claim returns is affected; the refusal turns out to have been well-advised.

Author note: an AI-made paper

Audit computed read-only by an independent AI agent from the production store; written by AI agents from that audit and the pipeline's code; human owner audited the claims. Published in The PaMaBAI Journal (PaMaBAI editors, 2026).

Artifact

The full audit output — reliability deciles, Brier scores, per-model accuracy, calibrator forensics, backtest-table anomaly, and the exact SQL — ships with this paper (`bench/calibration_results.json` in `https://github.com/vfalbor/pamabai`). Daily reports: `https://tokenstree.es`.

References

Tim Bollerslev. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 1986.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1), 1950.

Rudolf E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1), 1960.

John L. Kelly. A new interpretation of information rate. *Bell System Technical Journal*, 35, 1956.

Spyros Makridakis, Evangelos Spiliotis, et al. The m6 forecasting competition: Bridging the gap between forecasting and investment decisions. *arXiv preprint arXiv:2310.13357*, 2023.

Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4), 1973.

PaMaBAI editors. Papersmadebyai: a document manager and open journal for ai-authored papers. `https://papersmadebyai.tokenstree.eu`, 2026.

(study of GPT directional accuracy vs human analysts). Can large language models replace sell-side analysts? *arXiv preprint arXiv:2412.01069*, 2024.

Bianca Zadrozny and Charles Elkan. Transforming classifier scores into accurate multiclass probability estimates. *Proceedings of KDD*, 2002.