

# Weekly Calibration #4

September 2, 2026 · MarketMania Research · Weekly series · window Aug 24-30, 2026 UTC

## RESEARCH SNAPSHOT

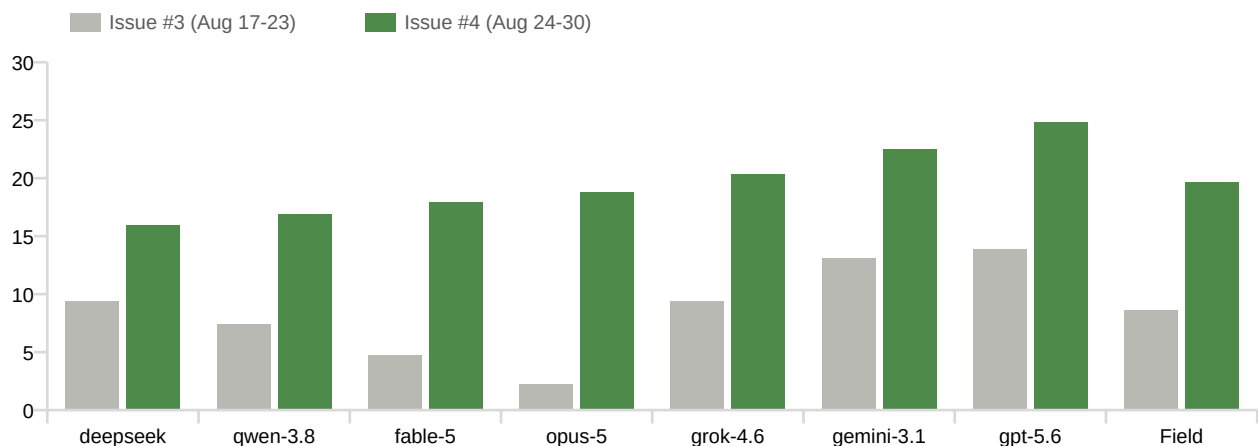
|                                                                               |                                  |                                                      |                                           |
|-------------------------------------------------------------------------------|----------------------------------|------------------------------------------------------|-------------------------------------------|
| 4,576 directional calls scored                                                | 7 models (grok line spliced)     | of 9,301 mature forecasts                            | window <b>Aug 24-30, 2026 UTC</b>         |
| hit = direction rule:<br>tp1/tp2->hit, sl->miss,<br>expiry->sign of gross pnl | FH gate <b>1h / 4h / 1d</b>      | base field hit <b>42.9%</b> vs <b>62.6</b><br>stated | cutoff <b>Mon Aug 31, 2026, 16:00 UTC</b> |
| market: BTC net <b>-0.07%</b><br>(prior +23.58%)                              | ann vol <b>30.9%</b> (was 71.1%) | TOP5 volume <b>\$19.72B</b> ,<br>-24% w/w            | pairwise corr <b>0.85</b> (was 0.81)      |

## KEY FINDING · OBSERVATION (one weekly window)

**The field's overconfidence gap moved +8.6pp -> +19.6pp week over week, 0 of 7 model lines narrowed, and field Brier went 0.2559 -> 0.2884.**

## TL;DR

- **OBSERVATION -- The gap read +19.6pp this week.** Directional calls hit 42.9% against 62.6 stated mean confidence (issue #3: +8.6pp on 54.0% vs 62.7). Field Brier is 0.2884 (was 0.2559) and 0 of 7 model lines scored below 0.25, the score of an uninformative always-50% predictor. One window, descriptive.
- **No model narrowed its gap.** Gap moves ran from deepseek-v4-pro's +6.5pp (+9.4pp -> +15.9pp) to claude-opus-5's +16.6pp (+2.2pp -> +18.8pp); 0 of 7 model lines narrowed week over week. deepseek-v4-pro is both the best-calibrated line (Brier 0.2760) and the hit-rate leader (44.6%).
- **The high-confidence buckets, N>=10 only.** deepseek-v4-pro 70-80 27.3% (n=11), gemini-3.1-pro 70-80 43.3% (n=261), gemini-3.1-pro 80-100 36.4% (n=11), gpt-5.6-sol 70-80 35.2% (n=210). Cells below N=10 are marked insufficient and rank nothing.
- **Trading, kept separate from every calibration table.** 0 of 7 lines finished net-positive; net PnL ran -\$58.75 (grok-4.6 (spliced), best) to -\$88.36 (gpt-5.6-sol, worst), field -\$473.59 against issue #3's +\$889.37.



Overconfidence gap in percentage points (mean stated confidence minus hit-rate), week over week, sorted by this week's gap. Short names: opus-5 = claude-opus-5, fable-5 = claude-fable-5, gemini-3.1 = gemini-3.1-pro, gpt-5.6 = gpt-5.6-sol, qwen-3.8 = qwen-3.8-max, deepseek = deepseek-v4-pro. The grok column compares the spliced 4.5+4.6 line against issue #3's pure grok-4.5. "Was" values are the numbers published in issue #3 for the back-to-back window Aug 17-23.

## Why it matters

Every MarketMania forecast carries a model-stated confidence from 0 to 100 alongside its direction call. This report checks whether that number tracks reality: on a well-calibrated forecaster, calls made at 70% confidence should hit their direction about 70% of the time. With issue #4 the series has four points on every model's gap -- enough to see movement, not enough to claim a trend: calibration and market regime move together, and durability claims belong in the Monthly series, not here.

## How to read this

Gap pp = mean stated confidence minus hit-rate, in percentage points; positive means overconfident. Brier is the mean squared error of the stated probability against the realised outcome, so lower is better and 0.25 is what an uninformative always-50% predictor scores. Coverage = calls scored divided by the mature calls available to that model. Confidence buckets are the models' own natural breakpoints, not an arbitrary binning; cells below N=10 are marked insufficient and never used to rank anything. Prediction and trading metrics live in separate sections and are never combined. The grok line is the lineage splice grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22).

## Calibration by model (sorted by Brier, best first)

| Model                     | N            | Cov.  | Hit rate     | Mean conf   | Gap pp       | Brier         | 95% CI      |
|---------------------------|--------------|-------|--------------|-------------|--------------|---------------|-------------|
| deepseek-v4-pro           | 729          | 54.8% | 44.6%        | 60.5        | +15.9        | 0.2760        | 41.0%-48.2% |
| qwen-3.8-max              | 672          | 50.5% | 42.6%        | 59.4        | +16.9        | 0.2772        | 38.9%-46.3% |
| claude-fable-5            | 651          | 49.2% | 43.0%        | 60.9        | +17.9        | 0.2799        | 39.3%-46.8% |
| claude-opus-5             | 561          | 42.2% | 42.4%        | 61.2        | +18.8        | 0.2816        | 38.4%-46.6% |
| grok-4.6 (spliced)        | 479          | 36.0% | 39.9%        | 60.2        | +20.3        | 0.2825        | 35.6%-44.3% |
| gemini-3.1-pro            | 755          | 56.9% | 44.2%        | 66.7        | +22.5        | 0.3025        | 40.7%-47.8% |
| gpt-5.6-sol               | 729          | 54.8% | 42.7%        | 67.5        | +24.8        | 0.3133        | 39.1%-46.3% |
| <b>Field (all models)</b> | <b>4,576</b> | -     | <b>42.9%</b> | <b>62.6</b> | <b>+19.6</b> | <b>0.2884</b> | -           |

N = directional calls scored this window. Cov. = coverage, scored / mature calls available to that model. Gap pp = mean stated confidence minus hit-rate (positive = overconfident). 95% CI is the descriptive Wilson interval on hit-rate. grok-4.6 (spliced) = grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22).

## Calibration curve: hit-rate by stated-confidence bucket

| Model              | 50-60       | 60-70       | 70-80       | 80-100       |
|--------------------|-------------|-------------|-------------|--------------|
| claude-fable-5     | 44.7% (199) | 42.3% (452) | n/a (0)     | n/a (0)      |
| claude-opus-5      | 46.0% (100) | 41.6% (461) | n/a (0)     | n/a (0)      |
| deepseek-v4-pro    | 48.9% (264) | 42.5% (454) | 27.3% (11)  | n/a (0)      |
| gemini-3.1-pro     | 50.0% (6) * | 44.9% (477) | 43.3% (261) | 36.4% (11)   |
| gpt-5.6-sol        | 75.0% (4) * | 45.3% (514) | 35.2% (210) | 100.0% (1) * |
| grok-4.6 (spliced) | 39.8% (181) | 40.1% (297) | 0.0% (1) *  | n/a (0)      |
| qwen-3.8-max       | 45.9% (325) | 39.2% (339) | 0.0% (2) *  | n/a (0)      |

n in parentheses. \* = N<10 (insufficient) -- no conclusions drawn from these cells. "n/a (0)" = no calls landed in that bucket this week. Sub-50 bucket (n greater than 0 only): qwen-3.8-max 66.7% (n=6) \*. Mid-scale ranking: 1 of the 5 models with N>=10 in both 50-60 and 60-70 out-hit 50-60 from 60-70 this week (issue #3: 4 of 5).

## Hit-rate by forecast horizon (FH)

| Model              | 1h          | 4h          | 1d         |
|--------------------|-------------|-------------|------------|
| claude-fable-5     | 42.3% (416) | 42.5% (188) | 51.1% (47) |
| claude-opus-5      | 42.1% (330) | 42.5% (186) | 44.4% (45) |
| deepseek-v4-pro    | 43.9% (435) | 44.1% (236) | 51.7% (58) |
| gemini-3.1-pro     | 44.0% (493) | 42.4% (212) | 54.0% (50) |
| gpt-5.6-sol        | 43.9% (458) | 39.9% (233) | 44.7% (38) |
| grok-4.6 (spliced) | 38.3% (300) | 41.1% (151) | 50.0% (28) |
| qwen-3.8-max       | 43.0% (430) | 40.6% (197) | 46.7% (45) |

n in parentheses; \* = N<10 (insufficient). 7 of 7 model lines hit better on 1d than on 1h this window (issue #3: 7 of 7).

## Trading result (a different question)

A different question: is any of this profitable to trade. "Accurate" and "profitable" are not the same thing, and per methodology this section is never mixed into the prediction/calibration metrics above.

| Model              | Trades | WR    | Net PnL  | Gross PnL | Max DD   |
|--------------------|--------|-------|----------|-----------|----------|
| grok-4.6 (spliced) | 479    | 35.1% | -\$58.75 | -\$10.85  | -\$63.22 |
| deepseek-v4-pro    | 729    | 39.4% | -\$59.07 | +\$13.83  | -\$67.18 |
| claude-opus-5      | 561    | 36.9% | -\$59.66 | -\$3.56   | -\$65.63 |
| claude-fable-5     | 651    | 36.9% | -\$62.85 | +\$2.25   | -\$76.89 |
| gemini-3.1-pro     | 755    | 37.2% | -\$67.73 | +\$7.77   | -\$78.95 |
| qwen-3.8-max       | 672    | 36.5% | -\$77.17 | -\$9.97   | -\$77.17 |
| gpt-5.6-sol        | 729    | 36.5% | -\$88.36 | -\$15.46  | -\$88.36 |

0 of 7 model lines finished net-positive and 3 gross-positive this week (issue #3: 7 of 7 on both). Ranked best-to-worst net PnL. Per methodology this table is never combined with the calibration tables above; a week's PnL is regime as much as skill.

## Platform note: TP/SL calibration

**Context, not a finding.** Since Aug 13 the public sandbox can trade with model-specific TP/SL multipliers learned from this same weekly history (v1); since Aug 19, v2 adds per-ticker and confidence-bucket (50-70 / 70-100) resolution. That feature consumes calibration history; it does not feed back into any table in this report, which measures stated confidence vs. direction-hit only. No effect size is claimed for v2. The toggle lives at [marketmania.ai/indices](https://marketmania.ai/indices).

## Market check: the tape cooled off

OBSERVATION — mean |1d move| 4.28% -> 2.35% (-45% rel), BTC realized vol 58.7% -> 37.9% (ann., hourly); field directional accuracy 54.0% -> 42.9% (-11.1pp), 0 of 7 models improved, sim win-rate up for 0 of 7, field sim PnL +\$889.37 -> -\$473.59 (adjacent calendar weeks Aug 17-23 vs Aug 24-30; descriptive, one pair of weeks, not a claim).

Robustness: raw price-sign accuracy 53.5% -> 44.2% — the same direction as the trade-based hit rule.

| Measure                         | Aug 17-23 (week A) | Aug 24-30 (week B) | Change   |
|---------------------------------|--------------------|--------------------|----------|
| Mean  1d move  (5 assets)       | 4.28%              | 2.35%              | -45% rel |
| BTC realized vol (ann., hourly) | 58.7%              | 37.9%              | -20.8pp  |
| Field directional accuracy      | 54.0%              | 42.9%              | -11.1pp  |
| Models improving hit-rate       | --                 | 0 of 7             | --       |
| Raw price-sign accuracy         | 53.5%              | 44.2%              | -9.3pp   |
| Field sim win-rate              | 46.3%              | 37.0%              | -9.2pp   |
| Field sim net PnL               | +\$889.37          | -\$473.59          | --       |

Week A is exactly the issue-#3 window, so this slice and every week-over-week column in this report compare the same two back-to-back weeks. Descriptive, one pair of weeks; the hit rule is methodology v1.1 (trade-based) and the raw price-sign row is the

robustness check (hourly closes at :00 against slots at :01). BTC realized vol here is the hourly-return series from the alive slice; the snapshot row's figure is the audited daily-candle estimate -- different estimators, both reported as measured. Lineage splice ACTIVE: grok-4.5 rows are aggregated into grok-4.6 (flip 2026-08-24T09:22:00+00:00, inside week B). Week A (2026-08-17..2026-08-23) is pure grok-4.5 and reproduces the published issue #3; week B mixes the 4.5-era (Aug 24 00:00-09:22) with 4.6.

## Practical implications

- Stated confidence is still a ranking hint at best, not a probability: the field's gap is +19.6pp this week against +8.6pp in issue #3.
- Gap and regime move together. A field that hits 42.9% instead of 54.0% closes or opens a confidence gap without a single stated number changing.
- Trading direction this week (0 of 7 lines net-positive) is a regime read, not a strategy result; the same lines printed the opposite sign inside three weeks.

## Limitations

- Prediction metrics (this report) and trading metrics are kept in separate sections per methodology -- they are never combined into a single score.
- 95% Wilson CIs shown are descriptive, not inferential: observations inside one window are dependent, so read them as a range, not a formal coverage guarantee.
- All 4,576 scored calls sit inside one market regime. This window ran 3 trend days of 7 (issue #3: 5 of 7); every comparison with issue #3 is a comparison across regimes as well as across weeks. Four week-over-week points cannot separate drift from regime; no durability claim is made.
- Models report confidence at discrete levels, not a continuous scale; the buckets reflect those natural breakpoints. Cells below N=10 are marked insufficient.
- Underlying price series are reconstructed from trade entry prices (median per symbol-slot), not an independent tick feed. Series density and lineage notes (wave 4): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Aug 24-30) daily coverage is FULL for the first time in the series: the 1w series has slots on 7 of 7 days and the 1M series on 7 of 7 days; both sit outside this report's FH gate (1h/4h/1d) and appear only in the exclusions counter (1w 488, 1M 490). (2) The grok line flipped 4.5 -> 4.6 at Aug 24, 2026 09:22 UTC, INSIDE this window: every grok row in this issue is the lineage splice "grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22)" (110 rows from the 4.5 era, 1,360 from 4.6); issue #3 was pure grok-4.5, so the grok week-over-week row joins a spliced line to a single-model line. (3) qwen-3.8-max succeeded qwen-3.7-max on Aug 5, 2026 (fully before this window).
- Market-state row is computed on a single-exchange (binance) daily candle series, as in issue #3 after the audit that found the raw table mixes two exchanges; issue #3 values are as published.
- Research-to-date counter: directional forecasts resolved inside the FH gate (1h/4h/1d) since Jul 11, counted at the issue cutoff (Aug 31 16:00 UTC) -- the definition pinned in issue #3, which printed 33,809 at the Aug 24 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 38,385 is an additive step under one definition; counter deltas against issue #2 and earlier remain definitional.

## Counters & lineage

| Counter                 | Value                          | Counter                | Value                                     |
|-------------------------|--------------------------------|------------------------|-------------------------------------------|
| Forecasts total         | 10,290                         | Mature (scored pool)   | 9,301                                     |
| OK in gate              | 9,301                          | -- of them directional | 4,576                                     |
| Out of gate (1w / 1M)   | 488 / 490                      | -- of them sideways    | 4,725                                     |
| Invalid                 | 11                             | Pending (next issue)   | 0                                         |
| Late closes             | 0                              | Uptime, all grids      | 266 / 266 slots                           |
| Source file             | weekly_metrics_2026-08-24.json | Report cutoff          | Mon Aug 31, 2026, 16:00 UTC               |
| Generated at (pipeline) | 2026-09-02 06:14 UTC           | Methodology            | v1.1 (2026-08-10) · hash e66c7e8c864a2233 |

**THIS REPORT: 4,576 scored observations.** MARKETMANIA RESEARCH TO DATE (as of Aug 31, 2026 cutoff): 38,385 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 16 published reports.

**Issue #4.** Weekly Calibration is a living series. Issue #3's open questions -- does the halved gap survive another week, and does gpt-5.6-sol's 70-80 bucket stay below its own 60-70? -- read this issue as: field gap +8.6pp -> +19.6pp (0 of 7 lines narrowed), and gpt-5.6-sol's 70-80 bucket read 35.2% (n=210) against its own 60-70 at 45.3% (n=514). The grok line is spliced this issue (grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22)) because the 4.5 -> 4.6 flip happened inside the window. Engine 1.1 has powered the sandbox since Aug 18, i.e. before this window; no cross-engine PnL comparisons are claimed.

## What we're testing next

- Next issue: does the field gap stay inside one band across two regimes, or does it track the base rate one-for-one?
- Do any two models keep the same Brier ordering for a third issue running?
- Monthly series: is the overconfidence gap stable across market regimes (trend vs flat) at monthly n?

## Related research

| Report                | Direct PDF link                                                                                                                                                   |
|-----------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Consensus Watch #4    | <a href="https://marketmania.ai/research/reports/consensus-watch-2026-08-24.pdf">https://marketmania.ai/research/reports/consensus-watch-2026-08-24.pdf</a>       |
| Weekly Model Watch #4 | <a href="https://marketmania.ai/research/reports/model-watch-2026-08-24.pdf">https://marketmania.ai/research/reports/model-watch-2026-08-24.pdf</a>               |
| Weekly Calibration #3 | <a href="https://marketmania.ai/research/reports/weekly-calibration-2026-08-17.pdf">https://marketmania.ai/research/reports/weekly-calibration-2026-08-17.pdf</a> |

The three weekly reports publish together as one issue each week; each links straight to the others' PDF and to its own previous issue. Direct links are the posting rule from wave 2 on.

## CITE THIS REPORT

```
@misc{mm_calibration_2026w35,
  title = {Weekly Calibration #4: confidence vs. direction-hit, Aug 24-30 2026},
  author = {{MarketMania Research}},
  year = {2026}, month = {September}, day = {2},
  url = {https://marketmania.ai/research/reports/weekly-calibration-2026-08-24.pdf},
  note = {Methodology v1.1, hash e66c7e8c864a2233; source weekly_metrics_2026-08-24.json}
}
```