

# Weekly Calibration #8

September 30, 2026 · MarketMania Research · Weekly series · window Sep 21-27, 2026 UTC

## RESEARCH SNAPSHOT

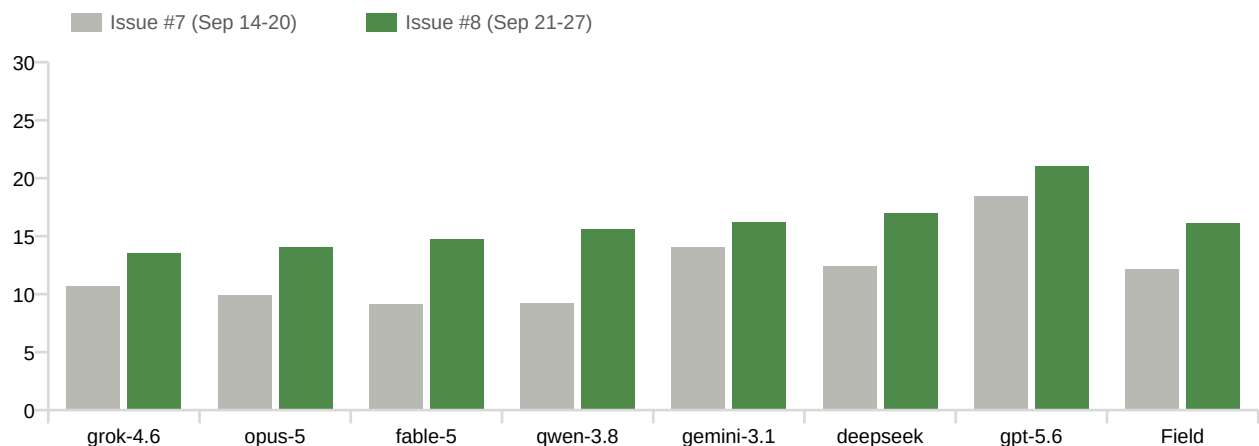
|                                                                               |                                  |                                                   |                                           |
|-------------------------------------------------------------------------------|----------------------------------|---------------------------------------------------|-------------------------------------------|
| 4,958 directional calls scored                                                | 7 models (stable lineup)         | of 9,273 mature forecasts                         | window <b>Sep 21-27, 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>46.8%</b> vs <b>63.0</b> stated | cutoff <b>Mon Sep 28, 2026, 16:00 UTC</b> |
| market: BTC net <b>+4.06%</b><br>(prior +5.64%)                               | ann vol <b>52.2%</b> (was 51.0%) | TOP5 volume <b>\$22.43B</b> ,<br>+20% w/w         | pairwise corr <b>0.71</b> (was 0.93)      |

KEY FINDING · OBSERVATION (one weekly window)

**The field's overconfidence gap moved +12.1pp -> +16.1pp week over week, 0 of 7 model lines narrowed, and field Brier went 0.2669 -> 0.2751.**

## TL;DR

- **OBSERVATION -- The gap read +16.1pp this week.** Directional calls hit 46.8% against 63.0 stated mean confidence (issue #7: +12.1pp on 50.7% vs 62.8). Field Brier is 0.2751 (was 0.2669) 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 gemini-3.1-pro's +2.2pp (+14.0pp -> +16.2pp) to qwen-3.8-max's +6.4pp (+9.2pp -> +15.6pp); 0 of 7 model lines narrowed week over week. qwen-3.8-max is the lowest-Brier line (Brier 0.2689); gemini-3.1-pro leads on hit-rate (51.4%).
- **The high-confidence buckets, N>=10 only.** deepseek-v4-pro 70-80 55.2% (n=29), gemini-3.1-pro 70-80 53.4% (n=320), gemini-3.1-pro 80-100 53.3% (n=15), gpt-5.6-sol 70-80 48.6% (n=220). 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 -\$5.94 (gemini-3.1-pro, best) to -\$86.60 (deepseek-v4-pro, worst), field -\$315.68 against issue #7's -\$337.72.



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 two pure grok-4.6 weeks: the 4.5 -> 4.6 flip sits four windows back. "Was" values are the numbers published in issue #7 for the back-to-back window Sep 14-20.

## 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 #8 the series has eight 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 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits four windows back (issue #4): both weeks of this issue are pure grok-4.6, with no 4.5-era rows in either window, so the grok week-over-week row is the third pure-4.6 against pure-4.6 comparison of the series.

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

| Model                     | N            | Cov.  | Hit rate     | Mean conf   | Gap pp       | Brier         | 95% CI      |
|---------------------------|--------------|-------|--------------|-------------|--------------|---------------|-------------|
| qwen-3.8-max              | 750          | 56.8% | 44.5%        | 60.1        | +15.6        | 0.2689        | 41.0%-48.1% |
| claude-fable-5            | 736          | 55.3% | 46.7%        | 61.4        | +14.7        | 0.2692        | 43.2%-50.3% |
| claude-opus-5             | 620          | 46.6% | 47.6%        | 61.5        | +14.0        | 0.2693        | 43.7%-51.5% |
| grok-4.6                  | 537          | 40.6% | 46.9%        | 60.4        | +13.5        | 0.2704        | 42.7%-51.2% |
| deepseek-v4-pro           | 777          | 59.3% | 43.9%        | 60.9        | +17.0        | 0.2740        | 40.4%-47.4% |
| gemini-3.1-pro            | 794          | 59.7% | 51.4%        | 67.6        | +16.2        | 0.2764        | 47.9%-54.9% |
| gpt-5.6-sol               | 744          | 55.9% | 46.8%        | 67.7        | +21.0        | 0.2951        | 43.2%-50.4% |
| <b>Field (all models)</b> | <b>4,958</b> | -     | <b>46.8%</b> | <b>63.0</b> | <b>+16.1</b> | <b>0.2751</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. The grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits four windows back (issue #4): both weeks of this issue are pure grok-4.6, with no 4.5-era rows in either window, so the grok week-over-week row is the third pure-4.6 against pure-4.6 comparison of the series.

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

| Model           | 50-60        | 60-70       | 70-80       | 80-100     |
|-----------------|--------------|-------------|-------------|------------|
| claude-fable-5  | 37.7% (199)  | 50.2% (536) | 0.0% (1) *  | n/a (0)    |
| claude-opus-5   | 45.9% (85)   | 47.9% (535) | n/a (0)     | n/a (0)    |
| deepseek-v4-pro | 41.3% (264)  | 44.7% (483) | 55.2% (29)  | n/a (0)    |
| gemini-3.1-pro  | 33.3% (3) *  | 50.0% (456) | 53.4% (320) | 53.3% (15) |
| gpt-5.6-sol     | 100.0% (5) * | 45.6% (517) | 48.6% (220) | 0.0% (2) * |
| grok-4.6        | 54.4% (182)  | 43.1% (355) | n/a (0)     | n/a (0)    |
| qwen-3.8-max    | 38.8% (299)  | 49.2% (437) | 20.0% (5) * | 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 22.2% (n=9) \*; deepseek-v4-pro 0.0% (n=1) \*. Mid-scale ranking: 4 of the 5 models with N>=10 in both 50-60 and 60-70 out-hit 50-60 from 60-70 this week (issue #7: 2 of 5).

## Hit-rate by forecast horizon (FH)

| Model           | 1h          | 4h          | 1d         |
|-----------------|-------------|-------------|------------|
| claude-fable-5  | 48.1% (464) | 44.1% (220) | 46.2% (52) |
| claude-opus-5   | 49.2% (378) | 44.8% (201) | 46.3% (41) |
| deepseek-v4-pro | 45.3% (470) | 40.1% (252) | 49.1% (55) |
| gemini-3.1-pro  | 51.6% (500) | 52.3% (237) | 45.6% (57) |
| gpt-5.6-sol     | 47.1% (448) | 46.5% (241) | 45.5% (55) |
| grok-4.6        | 48.9% (323) | 43.5% (170) | 45.5% (44) |
| qwen-3.8-max    | 45.1% (463) | 42.7% (234) | 47.2% (53) |

n in parentheses; \* = N<10 (insufficient). 2 of 7 model lines hit better on 1d than on 1h this window (issue #7: 0 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    |
|-----------------|--------|-------|----------|-----------|-----------|
| gemini-3.1-pro  | 794    | 43.0% | -\$5.94  | +\$73.46  | -\$99.60  |
| grok-4.6        | 537    | 39.9% | -\$29.75 | +\$23.95  | -\$87.02  |
| claude-opus-5   | 620    | 40.2% | -\$43.83 | +\$18.17  | -\$121.44 |
| claude-fable-5  | 736    | 39.4% | -\$45.60 | +\$28.00  | -\$127.29 |
| gpt-5.6-sol     | 744    | 40.5% | -\$48.68 | +\$25.72  | -\$129.10 |
| qwen-3.8-max    | 750    | 38.3% | -\$55.28 | +\$19.72  | -\$125.77 |
| deepseek-v4-pro | 777    | 38.6% | -\$86.60 | -\$8.90   | -\$154.88 |

0 of 7 model lines finished net-positive and 6 gross-positive this week (issue #7: 0 of 7 net-positive, 7 gross-positive). 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: an up week with smaller daily moves

OBSERVATION — mean |1d move| 2.69% -> 1.83% (-32% rel), BTC realized vol 35.1% -> 35.8% (ann., hourly); field directional accuracy 50.7% -> 46.8% (-3.9pp), 0 of 7 models improved, sim win-rate up for 0 of 7, field sim PnL -\$337.72 -> -\$315.68 (adjacent calendar weeks Sep 14-20 vs Sep 21-27; descriptive, one pair of weeks, not a claim).

Robustness: raw price-sign accuracy 54.2% -> 49.2% — the same direction as the trade-based hit rule.

| Measure                         | Sep 14-20 (week A) | Sep 21-27 (week B) | Change   |
|---------------------------------|--------------------|--------------------|----------|
| Mean  1d move  (5 assets)       | 2.69%              | 1.83%              | -32% rel |
| BTC realized vol (ann., hourly) | 35.1%              | 35.8%              | +0.7pp   |
| Field directional accuracy      | 50.7%              | 46.8%              | -3.9pp   |
| Models improving hit-rate       | --                 | 0 of 7             | --       |
| Raw price-sign accuracy         | 54.2%              | 49.2%              | -4.9pp   |
| Field sim win-rate              | 43.4%              | 40.0%              | -3.4pp   |
| Field sim net PnL               | -\$337.72          | -\$315.68          | --       |

Week A is exactly the issue-#7 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, four windows back, inside issue #4's window). Week A (2026-09-14..2026-09-20) is pure grok-4.6 (it reproduces the published issue #7); week B is pure grok-4.6.

## Practical implications

- Stated confidence is still a ranking hint at best, not a probability: the field's gap is +16.1pp this week against +12.1pp in issue #7.
- Gap and regime move together. The field hit 46.8% against 50.7% in issue #7 and stated mean confidence read 63.0 against 62.8, so the gap came out at +16.1pp: a field that hits 46.8% instead of 50.7% 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 six 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 4,958 scored calls sit inside one market regime. This window ran 2 trend days of 7 (issue #7: 3 of 7); every comparison with issue #7 is a comparison across regimes as well as across weeks. Eight 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 8): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Sep 21-27) daily coverage is FULL, as it was in issues #4, #5, #6 and #7: 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 485). (2) The grok line flipped 4.5 -> 4.6 at Aug 24, 2026 09:22 UTC, four windows back (inside the issue-#4 window): this window carries 1,470 grok rows and none from the 4.5 era, and neither does week A (the issue-#7 window), so every grok number in this issue and in issue #7 is a pure grok-4.6 line -- the grok week-over-week row is the third pure-4.6 against pure-4.6 comparison of the series. (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 #7 after the audit that found the raw table mixes two exchanges; issue #7 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 (Sep 28 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6 and #7, which printed 52,971 at the Sep 21 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 57,929 is an additive step under one definition; counter deltas against issue #6 and earlier remain definitional.

## Counters & lineage

| Counter                 | Value                          | Counter                | Value                                     |
|-------------------------|--------------------------------|------------------------|-------------------------------------------|
| Forecasts total         | 10,290                         | Mature (scored pool)   | 9,273                                     |
| OK in gate              | 9,273                          | -- of them directional | 4,958                                     |
| Out of gate (1w / 1M)   | 488 / 485                      | -- of them sideways    | 4,315                                     |
| Invalid                 | 44                             | Pending (next issue)   | 0                                         |
| Late closes             | 0                              | Uptime, all grids      | 266 / 266 slots                           |
| Source file             | weekly_metrics_2026-09-21.json | Report cutoff          | Mon Sep 28, 2026, 16:00 UTC               |
| Generated at (pipeline) | 2026-09-29 08:16 UTC           | Methodology            | v1.1 (2026-08-10) · hash e66c7e8c864a2233 |

**THIS REPORT: 4,958 scored observations.** MARKETMANIA RESEARCH TO DATE (as of Sep 28, 2026 cutoff): 57,929 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 36 published reports.

**Issue #8.** Weekly Calibration is a living series. Issue #7's open questions -- "the gap moved -10.4pp while the field base moved +11.2pp -- the second issue running in which the two move nearly one-for-one in opposite directions; does that hold for a third?" and "The best-Brier line changed hands to claude-fable-5 -- does it hold for a second issue, or change hands again?" -- read this issue as: field gap +12.1pp -> +16.1pp (0 of 7 lines narrowed) on a field base that moved -3.9pp, a third issue running in opposite directions, and the best-Brier line changed hands again: qwen-3.8-max at 0.2689 (issue #7: claude-fable-5 at 0.2593). Also on the record: gpt-5.6-sol's 70-80 bucket read 48.6% (n=220) against its own 60-70 at 45.6% (n=517). The grok row is pure grok-4.6 this issue and was pure grok-4.6 in issue #7; the 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits four windows back. 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: the gap moved +4.0pp while the field base moved -3.9pp -- the third issue running in which the two move nearly one-for-one in opposite directions; does that hold for a fourth?
- The best-Brier line changed hands again, to qwen-3.8-max -- does it hold for a second issue, or change hands again?
- Monthly series: is the overconfidence gap stable across market regimes (trend vs flat) at monthly n?

## Related research

| Report                | Direct PDF link                                                                                                                                                   |
|-----------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Consensus Watch #8    | <a href="https://marketmania.ai/research/reports/consensus-watch-2026-09-21.pdf">https://marketmania.ai/research/reports/consensus-watch-2026-09-21.pdf</a>       |
| Weekly Model Watch #8 | <a href="https://marketmania.ai/research/reports/model-watch-2026-09-21.pdf">https://marketmania.ai/research/reports/model-watch-2026-09-21.pdf</a>               |
| Weekly Calibration #7 | <a href="https://marketmania.ai/research/reports/weekly-calibration-2026-09-14.pdf">https://marketmania.ai/research/reports/weekly-calibration-2026-09-14.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_2026w39,
  title = {Weekly Calibration #8: confidence vs. direction-hit, Sep 21-27 2026},
  author = {{MarketMania Research}},
  year = {2026}, month = {September}, day = {30},
  url = {https://marketmania.ai/research/reports/weekly-calibration-2026-09-21.pdf},
  note = {Methodology v1.1, hash e66c7e8c864a2233; source weekly_metrics_2026-09-21.json}
}
```