# Weekly Calibration #11

October 21, 2026 · MarketMania Research · Weekly series · window Oct 12-18, 2026 UTC

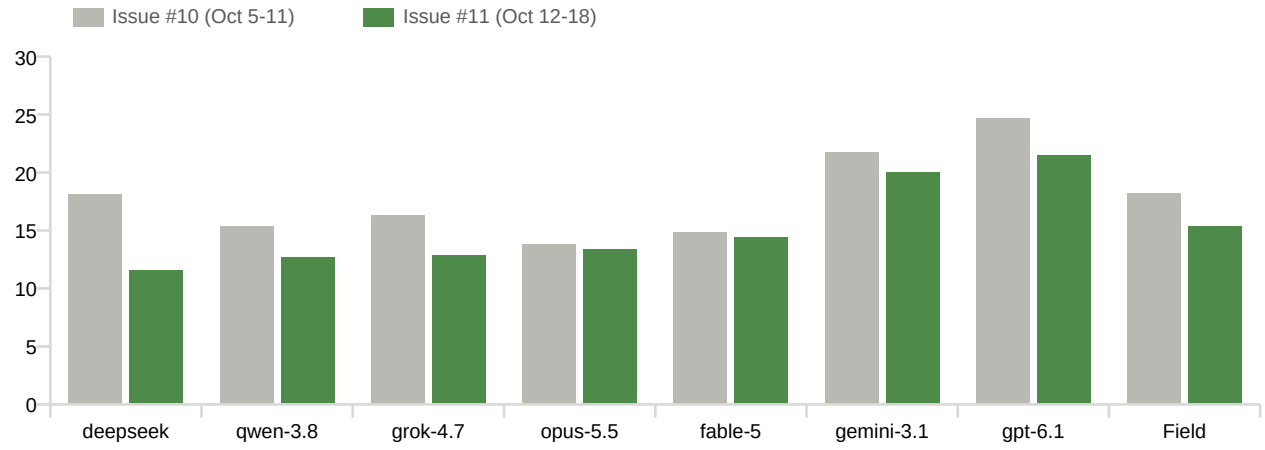
| RESEARCH SNAPSHOT | | | |
|---|---|---|---|
| **5,084** directional calls scored | **7** models (3 lines spliced) | of **9,198** mature forecasts | window **Oct 12-18, 2026 UTC** |
| hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl | FH gate **1h / 4h / 1d** | base field hit **46.6%** vs **61.9** stated | cutoff **Mon Oct 19, 2026, 16:00 UTC** |
| market: BTC net **-1.98%** (prior -3.19%) | ann vol **27.4%** (was 27.2%) | TOP5 volume **$15.84B**, -1% w/w | pairwise corr **0.94** (was 0.85) |

KEY FINDING · OBSERVATION (one weekly window)

**The field's overconfidence gap moved +18.2pp -> +15.4pp week over week, 7 of 7 model lines narrowed, and field Brier went 0.2822 -> 0.2759.**

## TL;DR

- **OBSERVATION -- The gap read +15.4pp this week.** Directional calls hit 46.6% against 61.9 stated mean confidence (issue #10: +18.2pp on 43.8% vs 61.9). Field Brier is 0.2759 (was 0.2822) and 0 of 7 model lines scored below 0.25, the score of an uninformative always-50% predictor. One window, descriptive.
- **7 of 7 models narrowed their gap.** Gap moves ran from deepseek-v4-pro's -6.5pp (+18.1pp -> +11.6pp) to claude-opus-5.5 (spliced)'s -0.4pp (+13.8pp -> +13.4pp); 7 of 7 model lines narrowed week over week. deepseek-v4-pro is both the lowest-Brier line (Brier 0.2657) and the hit-rate leader (48.2%).
- **The high-confidence buckets, N>=10 only.** gemini-3.1-pro 70-80 47.2% (n=233), gpt-6.1-sol (spliced) 70-80 46.5% (n=282). 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 -$24.12 (grok-4.7 (spliced), best) to -$71.98 (gemini-3.1-pro, worst), field -$352.09 against issue #10's -$520.82.




Overconfidence gap in percentage points (mean stated confidence minus hit-rate), week over week, sorted by this week's gap. Short names: opus-5.5 = claude-opus-5.5, fable-5 = claude-fable-5, gemini-3.1 = gemini-3.1-pro, gpt-6.1 = gpt-6.1-sol, qwen-3.8 = qwen-3.8-max, deepseek = deepseek-v4-pro. The opus-5.5, gpt-6.1 and grok-4.7 columns compare the spliced line against issue #10's pure claude-opus-5 / gpt-5.6-sol / grok-4.6. "Was" values are the numbers published in issue #10 for the back-to-back window Oct 5-11.

## 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 #11 the series has eleven 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. Three model lines flipped at Oct 17, 2026 09:01 UTC, INSIDE this window: gpt-5.6-sol -> gpt-6.1-sol, claude-opus-5 -> claude-opus-5.5 and grok-4.6 -> grok-4.7. Every row of those three lines in this issue is a lineage splice (old-era rows Oct 12 00:00-Oct 17 09:01, new-era rows after), week A (the issue-#10 window) is entirely old-era, so each of the three week-over-week rows joins a spliced line to a single-build line. Engine ids are kept; the vendors' names are Claude Opus 5.5, GPT-6.1 Sol and Grok 4.7. claude-opus-5.5 (spliced) = claude-opus-5.5 (incl. opus-5-era, Oct 12 00:00-Oct 17 09:01 UTC); gpt-6.1-sol (spliced) = gpt-6.1-sol (incl. 5.6-era, Oct 12 00:00-Oct 17 09:01 UTC); grok-4.7 (spliced) = grok-4.7 (incl. 4.6-era, Oct 12 00:00-Oct 17 09:01 UTC).

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

| Model | N | Cov. | Hit rate | Mean conf | Gap pp | Brier | 95% CI |
|---|---|---|---|---|---|---|---|
| deepseek-v4-pro | 767 | 59.0% | 48.2% | 59.9 | +11.6 | 0.2657 | 44.7%-51.8% |
| qwen-3.8-max | 766 | 57.6% | 46.3% | 59.1 | +12.7 | 0.2664 | 42.8%-49.9% |
| grok-4.7 (spliced) | 695 | 53.0% | 47.0% | 60.0 | +12.9 | 0.2682 | 43.4%-50.8% |
| opus-5.5 (spliced) | 572 | 43.7% | 46.3% | 59.7 | +13.4 | 0.2688 | 42.3%-50.4% |
| claude-fable-5 | 704 | 52.9% | 45.6% | 60.0 | +14.4 | 0.2714 | 41.9%-49.3% |
| gemini-3.1-pro | 765 | 58.5% | 45.8% | 65.8 | +20.0 | 0.2898 | 42.2%-49.3% |
| gpt-6.1 (spliced) | 815 | 62.2% | 46.5% | 68.0 | +21.5 | 0.2971 | 43.1%-49.9% |
| **Field (all models)** | **5,084** | - | **46.6%** | **61.9** | **+15.4** | **0.2759** | - |

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 claude-opus-5.5, gpt-6.1-sol and grok-4.7 rows are lineage splices (the three lines flipped at Oct 17, 2026 09:01 UTC, inside this window; week A is entirely old-era) -- see How to read this.

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

| Model | 50-60 | 60-70 | 70-80 | 80-100 |
|---|---|---|---|---|
| claude-fable-5 | 48.6% (247) | 44.0% (457) | n/a (0) | n/a (0) |
| claude-opus-5.5 (spliced) | 48.4% (219) | 45.0% (353) | n/a (0) | n/a (0) |
| deepseek-v4-pro | 50.3% (324) | 46.7% (435) | 42.9% (7) * | n/a (0) |
| gemini-3.1-pro | 0.0% (2) * | 45.5% (528) | 47.2% (233) | 0.0% (2) * |
| gpt-6.1-sol (spliced) | 47.2% (36) | 46.4% (489) | 46.5% (282) | 50.0% (8) * |
| grok-4.7 (spliced) | 51.6% (275) | 44.0% (420) | n/a (0) | n/a (0) |
| qwen-3.8-max | 46.5% (415) | 45.9% (338) | 60.0% (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): deepseek-v4-pro 100.0% (n=1) *; qwen-3.8-max 50.0% (n=8) *. Mid-scale ranking: 0 of the 6 models with N>=10 in both 50-60 and 60-70 out-hit 50-60 from 60-70 this week (issue #10: 1 of 7).

## Hit-rate by forecast horizon (FH)

| Model | 1h | 4h | 1d |
|---|---|---|---|
| claude-fable-5 | 47.4% (447) | 42.0% (212) | 44.4% (45) |
| claude-opus-5.5 (spliced) | 50.9% (336) | 36.7% (196) | 55.0% (40) |
| deepseek-v4-pro | 49.7% (445) | 45.3% (274) | 52.1% (48) |
| gemini-3.1-pro | 49.1% (481) | 38.5% (234) | 48.0% (50) |
| gpt-6.1-sol (spliced) | 45.9% (471) | 45.2% (281) | 57.1% (63) |
| grok-4.7 (spliced) | 46.8% (408) | 45.0% (238) | 59.2% (49) |
| qwen-3.8-max | 46.7% (473) | 44.1% (247) | 54.4% (46) |

n in parentheses; * = N<10 (insufficient). 5 of 7 model lines hit better on 1d than on 1h this window (issue #10: 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 |
|---|---|---|---|---|---|
| grok-4.7 (spliced) | 695 | 39.7% | -$24.12 | +$45.38 | -$62.19 |
| deepseek-v4-pro | 767 | 41.7% | -$42.95 | +$33.75 | -$76.15 |
| qwen-3.8-max | 766 | 38.6% | -$45.59 | +$31.01 | -$77.30 |
| gpt-6.1-sol (spliced) | 815 | 39.5% | -$50.53 | +$30.97 | -$72.48 |
| claude-opus-5.5 (spliced) | 572 | 39.3% | -$50.97 | +$6.23 | -$71.66 |
| claude-fable-5 | 704 | 38.4% | -$65.96 | +$4.44 | -$84.75 |
| gemini-3.1-pro | 765 | 37.0% | -$71.98 | +$4.52 | -$89.35 |

0 of 7 model lines finished net-positive and 7 gross-positive this week (issue #10: 0 of 7 net-positive, 2 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.

## Market check: bigger daily moves and higher BTC volatility

OBSERVATION — mean |1d move| 1.54% -> 2.05% (+33% rel), BTC realized vol 30.9% -> 34.9% (ann., hourly); field directional accuracy 43.8% -> 46.6% (+2.8pp), 6 of 7 models improved, sim win-rate up for 7 of 7, field sim PnL -$520.82 -> -$352.09 (adjacent calendar weeks Oct 5-11 vs Oct 12-18; descriptive, one pair of weeks, not a claim).
Robustness: raw price-sign accuracy 45.5% -> 44.1% — the opposite direction to the trade-based hit rule.

| Measure | Oct 5-11 (week A) | Oct 12-18 (week B) | Change |
|---|---|---|---|
| Mean \|1d move\| (5 assets) | 1.54% | 2.05% | +33% rel |
| BTC realized vol (ann., hourly) | 30.9% | 34.9% | +3.9pp |
| Field directional accuracy | 43.8% | 46.6% | +2.8pp |
| Models improving hit-rate | -- | 6 of 7 | -- |
| Raw price-sign accuracy | 45.5% | 44.1% | -1.4pp |
| Field sim win-rate | 34.5% | 39.2% | +4.7pp |
| Field sim net PnL | -$520.82 | -$352.09 | -- |

Week A is exactly the issue-#10 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: the model cutover (2026-10-17T09:01:00+00:00, inside week B, day 6 of 7) splices three lines -- gpt-5.6-sol -> gpt-6.1-sol, claude-opus-5 -> claude-opus-5.5, grok-4.6 -> grok-4.7. Week A (2026-10-05..2026-10-11) is entirely before the cutover: its old-id rows are aggregated under the heads (it reproduces the published issue #10); week B mixes the old-id era with the new ids. The grok 4.5 -> 4.6 flip (2026-08-24T09:22:00+00:00) sits seven windows back (inside issue #4's window).

## Practical implications

- Stated confidence is still a ranking hint at best, not a probability: the field's gap is +15.4pp this week against +18.2pp in issue #10.
- Gap and regime move together. The field hit 46.6% against 43.8% in issue #10 and stated mean confidence read 61.9 against 61.9, so the gap came out at +15.4pp: a field that hits 46.6% instead of 43.8% 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 nine 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 5,084 scored calls sit inside one market regime. This window ran 3 trend days of 7 (issue #10: 2 of 7); every comparison with issue #10 is a comparison across regimes as well as across weeks. Eleven 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 11): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Oct 12-18) daily coverage is FULL, as it was in issues #4, #5, #6, #7, #8, #9 and #10: 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 490, 1M 483). (2) Model lineage is a splice. Three lines flipped at Oct 17, 2026 09:01 UTC, INSIDE this window (day 6 of 7): "claude-opus-5.5 (incl. opus-5-era, Oct 12 00:00-Oct 17 09:01 UTC)" (1,155 rows from the claude-opus-5 era, 315 from claude-opus-5.5); "gpt-6.1-sol (incl. 5.6-era, Oct 12 00:00-Oct 17 09:01 UTC)" (1,155 rows from the gpt-5.6-sol era, 315 from gpt-6.1-sol); "grok-4.7 (incl. 4.6-era, Oct 12 00:00-Oct 17 09:01 UTC)" (1,155 rows from the grok-4.6 era, 315 from grok-4.7). Week A (the issue-#10 window) carries no new-build rows, so each of the three week-over-week rows joins a spliced line to a single-build line, and a remapped comparison is not the same model build. (3) The two hourly slots before the flip (Oct 17, 07:01 and 08:01 UTC) hold 60 invalid rows (claude-opus-5 20, gpt-5.6-sol 20, grok-4.6 20), counted in the invalid counter as they are; in those two hours the other 4 lines (claude-fable-5, deepseek-v4-pro, gemini-3.1-pro, qwen-3.8-max) carry ok rows. (4) deepseek-v4-pro carries 1,440 rows in this window against 1,470 for every other line (Oct 15: 180 rows instead of 210). (5) The grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits seven windows back (inside the issue-#4 window), and 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 #10 after the audit that found the raw table mixes two exchanges; issue #10 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 (Oct 19 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6, #7, #8, #9 and #10, which printed 66,861 at the Oct 12 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 71,945 is an additive step under one definition; counter deltas against issue #9 and earlier remain definitional.

## Counters & lineage

| Counter | Value | Counter | Value |
|---|---|---|---|
| Forecasts total | 10,260 | Mature (scored pool) | 9,198 |
| OK in gate | 9,198 | -- of them directional | 5,084 |
| Out of gate (1w / 1M) | 490 / 483 | -- of them sideways | 4,114 |
| Invalid | 89 | Pending (next issue) | 0 |
| Late closes | 0 | Uptime, all grids | 266 / 266 slots |
| Source file | weekly_metrics_2026-10-12.json | Report cutoff | Mon Oct 19, 2026, 16:00 UTC |
| Generated at (pipeline) | 2026-10-21 03:36 UTC | Methodology | v1.1 (2026-08-10) · hash e66c7e8c864a2233 |

**THIS REPORT: 5,084 scored observations.** MARKETMANIA RESEARCH TO DATE (as of Oct 19, 2026 cutoff): 71,945 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 56 published reports.

**Issue #11.** Weekly Calibration is a living series. Issue #10's open questions -- "the gap moved +5.3pp while the field base moved -5.9pp -- the fifth issue running in which the two move nearly one-for-one in opposite directions; does that hold for a sixth?" and "The best-Brier line changed hands again, to claude-opus-5 -- does it hold for a second issue, or change hands again?" -- read this issue as: field gap +18.2pp -> +15.4pp (7 of 7 lines narrowed) on a field base that moved +2.8pp, a sixth issue running in opposite directions, and the best-Brier line changed hands again: deepseek-v4-pro at 0.2657 (issue #10: claude-opus-5 at 0.2685). Also on the record: gpt-6.1-sol's 70-80 bucket read 46.5% (n=282) against its own 60-70 at 46.4% (n=489). Three rows are lineage splices this issue (claude-opus-5.5 (incl. opus-5-era, Oct 12 00:00-Oct 17 09:01 UTC); gpt-6.1-sol (incl. 5.6-era, Oct 12 00:00-Oct 17 09:01 UTC); grok-4.7 (incl. 4.6-era, Oct 12 00:00-Oct 17 09:01 UTC)) because the flip (Oct 17, 2026 09:01 UTC) fell inside the window; in issue #10 they were pure claude-opus-5, gpt-5.6-sol and grok-4.6. The grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits seven 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 -2.8pp while the field base moved +2.8pp -- the sixth issue running in which the two move nearly one-for-one in opposite directions; does that hold for a seventh?
- The best-Brier line changed hands again, to deepseek-v4-pro -- 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 #11 | https://marketmania.ai/research/reports/consensus-watch-2026-10-12.pdf |
| Weekly Model Watch #11 | https://marketmania.ai/research/reports/model-watch-2026-10-12.pdf |
| Weekly Calibration #10 | https://marketmania.ai/research/reports/weekly-calibration-2026-10-05.pdf |

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_2026w42,
  title  = {Weekly Calibration #11: confidence vs. direction-hit, Oct 12-18 2026},
  author = {{MarketMania Research}},
  year   = {2026}, month = {October}, day = {21},
  url    = {https://marketmania.ai/research/reports/weekly-calibration-2026-10-12.pdf},
  note   = {Methodology v1.1, hash e66c7e8c864a2233; source weekly_metrics_2026-10-12.json}
}
```