Weekly Calibration #15

November 18, 2026 · MarketMania Research · Weekly series · window Nov 9-15, 2026 UTC

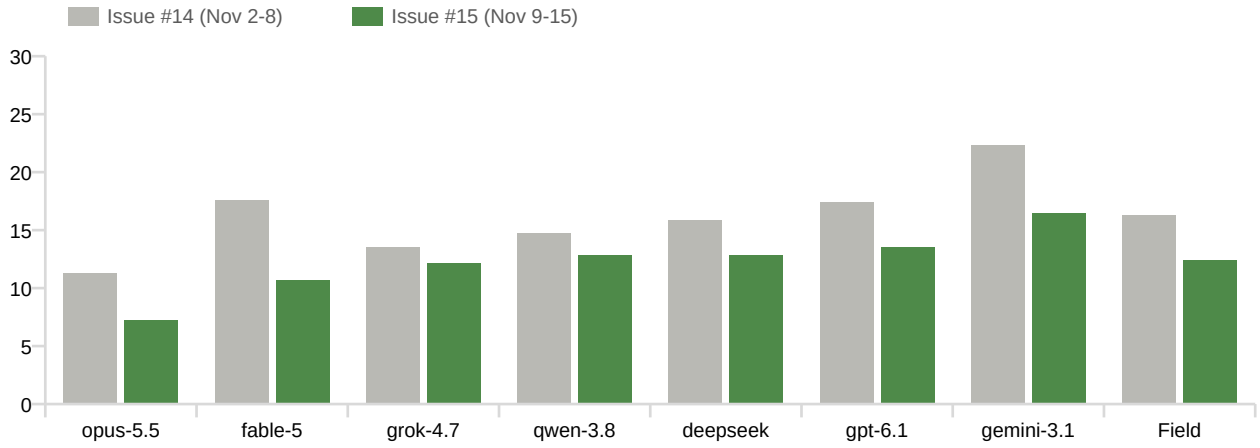
RESEARCH SNAPSHOT			
5,669 directional calls scored	**7** models (stable lineup)	of **9,298** mature forecasts	window **Nov 9-15, 2026 UTC**
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate **1h / 4h / 1d**	base field hit **48.5%** vs **60.9** stated	cutoff **Mon Nov 16, 2026, 16:00 UTC**
market: BTC net **-12.69%** (prior -4.47%)	ann vol **57.5%** (was 55.3%)	TOP5 volume **$18.77B**, +32% w/w	pairwise corr **0.95** (was 0.94)

KEY FINDING · OBSERVATION (one weekly window)

The field's overconfidence gap moved +16.3pp -> +12.4pp week over week, 7 of 7 model lines narrowed, and field Brier went 0.2772 -> 0.2677.

TL;DR

- **OBSERVATION -- The gap read +12.4pp this week.** Directional calls hit 48.5% against 60.9 stated mean confidence (issue #14: +16.3pp on 44.8% vs 61.1). Field Brier is 0.2677 (was 0.2772) 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 claude-fable-5's -6.9pp (+17.6pp -> +10.7pp) to grok-4.7's -1.4pp (+13.5pp -> +12.1pp); 7 of 7 model lines narrowed week over week. claude-opus-5.5 is the lowest-Brier line (Brier 0.2555); gpt-6.1-sol leads on hit-rate (51.1%).
- **The high-confidence buckets, N>=10 only.** gpt-6.1-sol 70-80 43.8% (n=144), gemini-3.1-pro 70-80 47.7% (n=239). Cells below N=10 are marked insufficient and rank nothing.
- **Trading, kept separate from every calibration table.** 5 of 7 lines finished net-positive; net PnL ran +$46.68 (claude-fable-5, best) to -$7.53 (qwen-3.8-max, worst), field +$160.79 against issue #14's -$464.89.



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. "Was" values are the numbers published in issue #14 for the back-to-back window Nov 2-8.

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 #15 the series has fifteen 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.

Calibration by model (sorted by Brier, best first)

Model	N	Cov.	Hit rate	Mean conf	Gap pp	Brier	95% CI
claude-opus-5.5	613	46.1%	48.6%	55.8	+7.2	0.2555	44.7%-52.6%
claude-fable-5	705	53.0%	48.5%	59.2	+10.7	0.2622	44.8%-52.2%
qwen-3.8-max	725	54.5%	45.7%	58.4	+12.8	0.2650	42.1%-49.3%
grok-4.7	963	72.9%	47.6%	59.6	+12.1	0.2658	44.4%-50.7%
deepseek-v4-pro	839	63.1%	47.4%	60.2	+12.8	0.2679	44.1%-50.8%
gpt-6.1-sol	1,034	77.7%	51.1%	64.6	+13.5	0.2719	48.0%-54.1%
gemini-3.1-pro	790	59.5%	49.8%	66.2	+16.4	0.2811	46.3%-53.2%
Field (all models)	**5,669**	-	**48.5%**	**60.9**	**+12.4**	**0.2677**	-

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.

Calibration curve: hit-rate by stated-confidence bucket

Model	50-60	60-70	70-80	80-100
claude-fable-5	47.6% (313)	49.2% (392)	n/a (0)	n/a (0)
claude-opus-5.5	48.6% (611)	50.0% (2) *	n/a (0)	n/a (0)
deepseek-v4-pro	46.6% (307)	48.5% (526)	0.0% (6) *	n/a (0)
gemini-3.1-pro	20.0% (5) *	51.4% (537)	47.7% (239)	22.2% (9) *
gpt-6.1-sol	48.8% (127)	52.8% (763)	43.8% (144)	n/a (0)
grok-4.7	49.4% (431)	46.1% (532)	n/a (0)	n/a (0)
qwen-3.8-max	45.9% (458)	45.4% (262)	n/a (0)	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 40.0% (n=5) *. Mid-scale ranking: 3 of the 5 models with N>=10 in both 50-60 and 60-70 out-hit 50-60 from 60-70 this week (issue #14: 2 of 6).

Hit-rate by forecast horizon (FH)

Model	1h	4h	1d
claude-fable-5	49.1% (407)	42.7% (234)	65.6% (64)
claude-opus-5.5	48.8% (361)	42.9% (205)	72.3% (47)
deepseek-v4-pro	48.7% (472)	41.6% (308)	67.8% (59)
gemini-3.1-pro	52.0% (492)	43.1% (248)	60.0% (50)
gpt-6.1-sol	51.9% (597)	46.6% (367)	67.1% (70)
grok-4.7	47.4% (572)	43.1% (325)	71.2% (66)
qwen-3.8-max	44.8% (415)	42.6% (249)	63.9% (61)

n in parentheses; * = N<10 (insufficient). 7 of 7 model lines hit better on 1d than on 1h this window (issue #14: 5 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
claude-fable-5	705	44.1%	+$46.68	+$117.18	-$59.80
gpt-6.1-sol	1,034	46.4%	+$41.66	+$145.06	-$79.41
gemini-3.1-pro	790	43.8%	+$38.07	+$117.07	-$59.99
grok-4.7	963	43.4%	+$32.32	+$128.62	-$78.14
claude-opus-5.5	613	43.9%	+$12.79	+$74.09	-$56.34
deepseek-v4-pro	839	43.5%	-$3.19	+$80.71	-$74.52
qwen-3.8-max	725	41.0%	-$7.53	+$64.97	-$77.97

5 of 7 model lines finished net-positive and 7 gross-positive this week (issue #14: 0 of 7 net-positive, 4 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 on higher volume

OBSERVATION — mean |1d move| 2.16% -> 3.40% (+57% rel), BTC realized vol 41.0% -> 55.8% (ann., hourly); field directional accuracy 44.8% -> 48.5% (+3.7pp), 7 of 7 models improved, sim win-rate up for 7 of 7, field sim PnL -$464.89 -> +$160.79 (adjacent calendar weeks Nov 2-8 vs Nov 9-15; descriptive, one pair of weeks, not a claim).
Robustness: raw price-sign accuracy 48.9% -> 54.4% — the same direction as the trade-based hit rule.

Measure	Nov 2-8 (week A)	Nov 9-15 (week B)	Change
Mean |1d move| (5 assets)	2.16%	3.40%	+57% rel
BTC realized vol (ann., hourly)	41.0%	55.8%	+14.8pp
Field directional accuracy	44.8%	48.5%	+3.7pp
Models improving hit-rate	--	7 of 7	--
Raw price-sign accuracy	48.9%	54.4%	+5.6pp
Field sim win-rate	37.4%	43.9%	+6.4pp
Field sim net PnL	-$464.89	+$160.79	--

Week A is exactly the issue-#14 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 note: the model cutover (2026-10-17T09:01:00+00:00, day 6 of 7 of its Mon-Sun week) sits four windows back, inside the published issue-#11 window, where it spliced 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 (2026-11-02..2026-11-08, the published issue-#14 window) and week B (2026-11-09..2026-11-15) are both full weeks on the heads (0 and 0 old-id row(s) remapped); week A reproduces the published issue #14. The grok 4.5 -> 4.6 flip (2026-08-24T09:22:00+00:00) sits eleven 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 +12.4pp this week against +16.3pp in issue #14.
- Gap and regime move together. The field hit 48.5% against 44.8% in issue #14 and stated mean confidence read 60.9 against 61.1, so the gap came out at +12.4pp: a field that hits 48.5% instead of 44.8% closes or opens a confidence gap without a single stated number changing.
- Trading direction this week (5 of 7 lines net-positive) is a regime read, not a strategy result; it follows two issues in which no line finished net-positive (issue #14: 0 of 7 net-positive).

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 5,669 scored calls sit inside one market regime. This window had trend on 5 of 7 days (issue #14: 4 of 7); every comparison with issue #14 is a comparison across regimes as well as across weeks. Fifteen 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 15): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Nov 9-15) daily coverage is FULL, as it was in issues #4, #5, #6, #7, #8, #9, #10, #11, #12, #13 and #14: 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 489, 1M 487). (2) This window is the fourth full week on the builds that took over at Oct 17, 2026 09:01 UTC: claude-opus-5.5, gpt-6.1-sol and grok-4.7 carry no rows from the old builds in either week, and the 7 lines carry 1,470 rows in this window and 1,461 / 1,470 in week A, so every week-over-week row compares two full weeks on the same builds. (3) No roll-out slot of the flip falls inside either week: the two hourly slots before it (Oct 17, 07:01 and 08:01 UTC) lie four windows back, so this window's invalid counter holds none of their rows. (4) Past lineage flips: the Oct 17, 2026 09:01 UTC flip sits four windows back (inside the issue-#11 window), the grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits eleven 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 #14 after the audit that found the raw table mixes two exchanges; issue #14 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 (Nov 16 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6, #7, #8, #9, #10, #11, #12, #13 and #14, which printed 87,760 at the Nov 9 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 93,429 is an additive step under one definition; counter deltas against issue #13 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,298
OK in gate	9,298	-- of them directional	5,669
Out of gate (1w / 1M)	489 / 487	-- of them sideways	3,629
Invalid	16	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-11-09.json	Report cutoff	Mon Nov 16, 2026, 16:00 UTC
Generated at (pipeline)	2026-11-18 06:18 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

Provider note: grok-4.7 returned no valid answer for 13 of its 1,470 forecasts in this window (8 on Nov 9, 3 on Nov 10, 1 on Nov 11, 1 on Nov 15); those rows are counted in Invalid. The engine retries a call the provider rejects on input moderation without the news block (since Nov 6); 4 forecasts in this window (qwen-3.8-max: 1 on Nov 10, 2 on Nov 11, 1 on Nov 13) were produced that way and are flagged `news_stripped` in the open data. Hit rates in this report use the rows that resolved.

THIS REPORT: 5,669 scored observations. MARKETMANIA RESEARCH TO DATE (as of Nov 16, 2026 cutoff): 93,429 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 76 published reports.

Issue #15. Weekly Calibration is a living series. Issue #14's open questions -- "the gap moved -3.3pp while the field base moved +3.1pp -- the ninth issue running in which the two move nearly one-for-one in opposite directions; does that hold for a tenth?" and "claude-opus-5.5 kept the lowest Brier score for a third issue -- does it hold for a fourth issue, or change hands?" -- read this issue as: field gap +16.3pp -> +12.4pp (7 of 7 lines narrowed) on a field base that moved +3.7pp, a tenth issue running in opposite directions, and the lowest Brier score stayed with the same line for a fourth issue: claude-opus-5.5 at 0.2555 (issue #14: claude-opus-5.5 at 0.2609). Also on the record: gpt-6.1-sol's 70-80 bucket read 43.8% (n=144) against its own 60-70 at 52.8% (n=763). The claude-opus-5.5, gpt-6.1-sol and grok-4.7 rows are the fourth full week on the Oct 17 builds; week-over-week rows compare two full weeks on the same builds. The grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits eleven 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 -3.9pp while the field base moved +3.7pp -- the tenth issue running in which the two move nearly one-for-one in opposite directions; does that hold for an eleventh?
- claude-opus-5.5 kept the lowest Brier score for a fourth issue -- does it hold for a fifth issue, or change hands?
- Monthly series: is the overconfidence gap stable across market regimes (trend vs flat) at monthly n?

Related research

Report	Direct PDF link
Consensus Watch #15	https://marketmania.ai/research/reports/consensus-watch-2026-11-09.pdf
Weekly Model Watch #15	https://marketmania.ai/research/reports/model-watch-2026-11-09.pdf
Weekly Calibration #14	https://marketmania.ai/research/reports/weekly-calibration-2026-11-02.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_2026w46,
  title  = {Weekly Calibration #15: confidence vs. direction-hit, Nov 9-15 2026},
  author = {{MarketMania Research}},
  year   = {2026}, month = {November}, day = {18},
  url    = {https://marketmania.ai/research/reports/weekly-calibration-2026-11-09.pdf},
  note   = {Methodology v1.1, hash e66c7e8c864a2233; source weekly_metrics_2026-11-09.json}
}
```