Weekly Calibration #13

November 4, 2026 · MarketMania Research · Weekly series · window Oct 26-Nov 1, 2026 UTC

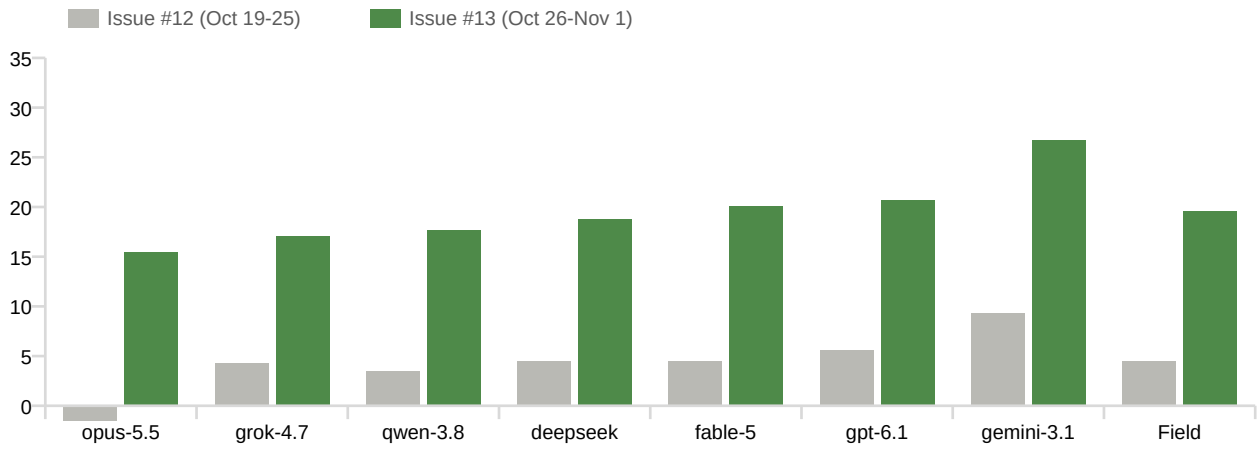
RESEARCH SNAPSHOT			
5,094 directional calls scored	**7** models (stable lineup)	of **9,304** mature forecasts	window **Oct 26-Nov 1, 2026 UTC**
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate **1h / 4h / 1d**	base field hit **41.6%** vs **61.2** stated	cutoff **Mon Nov 2, 2026, 16:00 UTC**
market: BTC net **-2.25%** (prior -2.39%)	ann vol **34.7%** (was 29.3%)	TOP5 volume **$14.86B**, +6% w/w	pairwise corr **0.93** (was 0.93)

KEY FINDING · OBSERVATION (one weekly window)

The field's overconfidence gap moved +4.5pp -> +19.6pp week over week, 0 of 7 model lines narrowed, and field Brier went 0.2484 -> 0.2858.

TL;DR

- **OBSERVATION -- The gap read +19.6pp this week.** Directional calls hit 41.6% against 61.2 stated mean confidence (issue #12: +4.5pp on 57.0% vs 61.5). Field Brier is 0.2858 (was 0.2484) 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 grok-4.7's +12.8pp (+4.3pp -> +17.1pp) to gemini-3.1-pro's +17.4pp (+9.3pp -> +26.7pp); 0 of 7 model lines narrowed week over week. claude-opus-5.5 is the lowest-Brier line (Brier 0.2676); grok-4.7 leads on hit-rate (42.5%). Every line is over-confident this window (positive gap); in issue #12 claude-opus-5.5 was the only line whose gap was negative (-1.5pp).
- **The high-confidence buckets, N>=10 only.** deepseek-v4-pro 70-80 30.0% (n=10), gpt-6.1-sol 70-80 26.9% (n=52), gemini-3.1-pro 70-80 41.4% (n=280), gemini-3.1-pro 80-100 36.4% (n=55). 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 -$93.22 (claude-opus-5.5, best) to -$161.50 (gemini-3.1-pro, worst), field -$869.65 against issue #12's -$43.00.



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 #12 for the back-to-back window Oct 19-25.

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 #13 the series has thirteen 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	591	44.4%	40.9%	56.4	+15.4	0.2676	37.0%-45.0%
grok-4.7	830	62.4%	42.5%	59.6	+17.1	0.2754	39.2%-45.9%
qwen-3.8-max	711	53.7%	41.3%	59.0	+17.7	0.2763	37.8%-45.0%
deepseek-v4-pro	745	56.0%	41.5%	60.2	+18.8	0.2805	38.0%-45.1%
claude-fable-5	633	47.6%	40.4%	60.5	+20.1	0.2835	36.7%-44.3%
gpt-6.1-sol	829	62.3%	42.5%	63.2	+20.7	0.2922	39.1%-45.9%
gemini-3.1-pro	755	56.8%	41.7%	68.4	+26.7	0.3205	38.3%-45.3%
Field (all models)	**5,094**	-	**41.6%**	**61.2**	**+19.6**	**0.2858**	-

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	43.0% (207)	39.2% (426)	n/a (0)	n/a (0)
claude-opus-5.5	42.2% (559)	18.8% (32)	n/a (0)	n/a (0)
deepseek-v4-pro	40.4% (287)	42.4% (448)	30.0% (10)	n/a (0)
gemini-3.1-pro	50.0% (2) *	42.6% (418)	41.4% (280)	36.4% (55)
gpt-6.1-sol	47.3% (131)	42.7% (646)	26.9% (52)	n/a (0)
grok-4.7	44.5% (402)	40.8% (426)	0.0% (2) *	n/a (0)
qwen-3.8-max	43.0% (388)	39.4% (310)	50.0% (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 36.4% (n=11). Mid-scale ranking: 1 of the 6 models with N>=10 in both 50-60 and 60-70 out-hit 50-60 from 60-70 this week (issue #12: 5 of 6).

Hit-rate by forecast horizon (FH)

Model	1h	4h	1d
claude-fable-5	43.8% (395)	34.5% (197)	36.6% (41)
claude-opus-5.5	44.3% (361)	33.3% (192)	47.4% (38)
deepseek-v4-pro	45.5% (442)	34.2% (260)	44.2% (43)
gemini-3.1-pro	44.5% (456)	36.0% (250)	44.9% (49)
gpt-6.1-sol	45.5% (501)	37.4% (273)	40.0% (55)
grok-4.7	45.4% (513)	36.0% (267)	48.0% (50)
qwen-3.8-max	45.5% (429)	33.2% (238)	45.5% (44)

n in parentheses; * = N<10 (insufficient). 3 of 7 model lines hit better on 1d than on 1h this window (issue #12: 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
claude-opus-5.5	591	33.2%	-$93.22	-$34.12	-$98.12
qwen-3.8-max	711	34.5%	-$105.90	-$34.80	-$112.94
grok-4.7	830	34.5%	-$109.74	-$26.74	-$114.44
claude-fable-5	633	31.3%	-$127.76	-$64.46	-$131.13
deepseek-v4-pro	745	34.8%	-$127.93	-$53.43	-$133.18
gpt-6.1-sol	829	33.7%	-$143.60	-$60.70	-$149.67
gemini-3.1-pro	755	33.0%	-$161.50	-$86.00	-$166.35

0 of 7 model lines finished net-positive and 0 gross-positive this week (issue #12: 3 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.

Market check: bigger daily moves on higher volume

OBSERVATION — mean |1d move| 1.35% -> 1.63% (+20% rel), BTC realized vol 27.5% -> 37.0% (ann., hourly); field directional accuracy 57.0% -> 41.6% (-15.3pp), 0 of 7 models improved, sim win-rate up for 0 of 7, field sim PnL -$43.00 -> -$869.65 (adjacent calendar weeks Oct 19-25 vs Oct 26-Nov 1; descriptive, one pair of weeks, not a claim).
Robustness: raw price-sign accuracy 56.0% -> 42.8% — the same direction as the trade-based hit rule.

Measure	Oct 19-25 (week A)	Oct 26-Nov 1 (week B)	Change
Mean |1d move| (5 assets)	1.35%	1.63%	+20% rel
BTC realized vol (ann., hourly)	27.5%	37.0%	+9.5pp
Field directional accuracy	57.0%	41.6%	-15.3pp
Models improving hit-rate	--	0 of 7	--
Raw price-sign accuracy	56.0%	42.8%	-13.2pp
Field sim win-rate	46.6%	33.6%	-13.0pp
Field sim net PnL	-$43.00	-$869.65	--

Week A is exactly the issue-#12 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 two 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 (2026-10-19..2026-10-25, the published issue-#12 window) and week B (2026-10-26..2026-11-01) are both full weeks on the heads (0 and 0 old-id row(s) remapped); week A reproduces the published issue #12. The grok 4.5 -> 4.6 flip (2026-08-24T09:22:00+00:00) sits nine 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 +19.6pp this week against +4.5pp in issue #12.
- Gap and regime move together. The field hit 41.6% against 57.0% in issue #12 and stated mean confidence read 61.2 against 61.5, so the gap came out at +19.6pp: a field that hits 41.6% instead of 57.0% 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 three lines that were net-positive one week earlier all finished net-negative (issue #12: 3 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 5,094 scored calls sit inside one market regime. This window had trend on 1 of 7 days (issue #12: 1 of 7); every comparison with issue #12 is a comparison across regimes as well as across weeks. Thirteen 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 13): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Oct 26-Nov 1) daily coverage is FULL, as it was in issues #4, #5, #6, #7, #8, #9, #10, #11 and #12: 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 488). (2) This window is the second 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 all 7 lines carry 1,470 rows each in both weeks, 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 two 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 two windows back (inside the issue-#11 window), the grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits nine 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 #12 after the audit that found the raw table mixes two exchanges; issue #12 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 2 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6, #7, #8, #9, #10, #11 and #12, which printed 77,801 at the Oct 26 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 82,895 is an additive step under one definition; counter deltas against issue #11 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,304
OK in gate	9,304	-- of them directional	5,094
Out of gate (1w / 1M)	490 / 488	-- of them sideways	4,210
Invalid	8	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-10-26.json	Report cutoff	Mon Nov 2, 2026, 16:00 UTC
Generated at (pipeline)	2026-11-03 11:35 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 5,094 scored observations. MARKETMANIA RESEARCH TO DATE (as of Nov 2, 2026 cutoff): 82,895 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 64 published reports.

Issue #13. Weekly Calibration is a living series. Issue #12's open questions -- "the gap moved -10.9pp while the field base moved +10.4pp -- the seventh issue running in which the two move nearly one-for-one in opposite directions; does that hold for an eighth?" and "The best-Brier line changed hands again, to claude-opus-5.5 -- does it hold for a second issue, or change hands again?" -- read this issue as: field gap +4.5pp -> +19.6pp (0 of 7 lines narrowed) on a field base that moved -15.3pp, an eighth issue running in opposite directions, and the lowest Brier score stayed with the same line for a second issue: claude-opus-5.5 at 0.2676 (issue #12: claude-opus-5.5 at 0.2443). Also on the record: gpt-6.1-sol's 70-80 bucket read 26.9% (n=52) against its own 60-70 at 42.7% (n=646). The claude-opus-5.5, gpt-6.1-sol and grok-4.7 rows are the second 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 nine 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 +15.1pp while the field base moved -15.3pp -- the eighth issue running in which the two move nearly one-for-one in opposite directions; does that hold for a ninth?
- claude-opus-5.5 kept the lowest Brier score for a second issue -- does it hold for a third 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 #13	https://marketmania.ai/research/reports/consensus-watch-2026-10-26.pdf
Weekly Model Watch #13	https://marketmania.ai/research/reports/model-watch-2026-10-26.pdf
Weekly Calibration #12	https://marketmania.ai/research/reports/weekly-calibration-2026-10-19.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_2026w44,
  title  = {Weekly Calibration #13: confidence vs. direction-hit, Oct 26-Nov 1 2026},
  author = {{MarketMania Research}},
  year   = {2026}, month = {November}, day = {4},
  url    = {https://marketmania.ai/research/reports/weekly-calibration-2026-10-26.pdf},
  note   = {Methodology v1.1, hash e66c7e8c864a2233; source weekly_metrics_2026-10-26.json}
}
```