Weekly Calibration #12

October 28, 2026 · MarketMania Research · Weekly series · window Oct 19-25, 2026 UTC

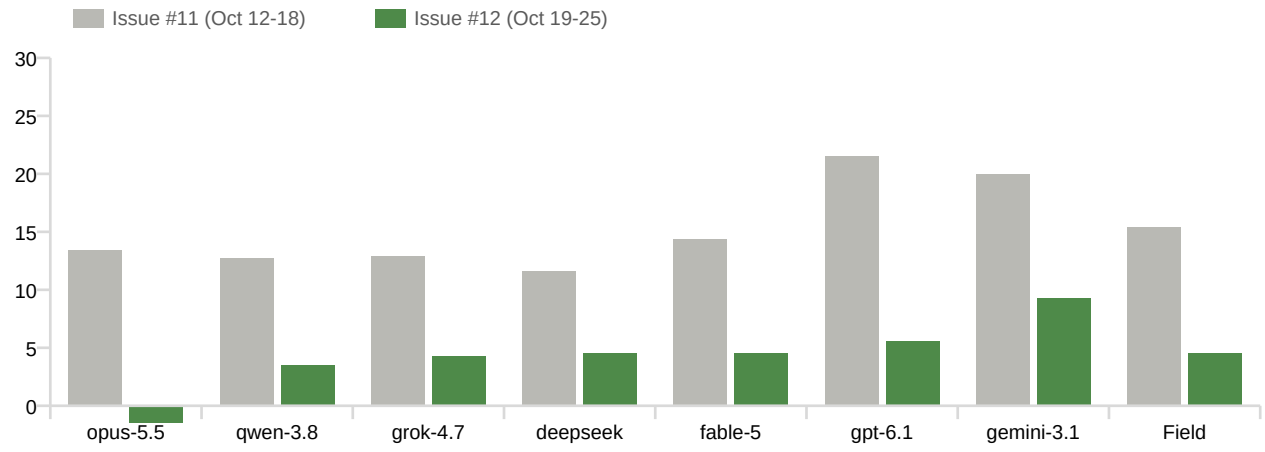
RESEARCH SNAPSHOT			
5,856 directional calls scored	**7** models (stable lineup)	of **9,171** mature forecasts	window **Oct 19-25, 2026 UTC**
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate **1h / 4h / 1d**	base field hit **57.0%** vs **61.5** stated	cutoff **Mon Oct 26, 2026, 16:00 UTC**
market: BTC net **-2.39%** (prior -1.98%)	ann vol **29.3%** (was 27.4%)	TOP5 volume **$14.02B**, -12% w/w	pairwise corr **0.93** (was 0.94)

KEY FINDING · OBSERVATION (one weekly window)

The field's overconfidence gap moved +15.4pp -> +4.5pp week over week, 7 of 7 model lines narrowed, and field Brier went 0.2759 -> 0.2484.

TL;DR

- **OBSERVATION -- The gap read +4.5pp this week.** Directional calls hit 57.0% against 61.5 stated mean confidence (issue #11: +15.4pp on 46.6% vs 61.9). Field Brier is 0.2484 (was 0.2759) and 6 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 gpt-6.1-sol's -15.9pp (+21.5pp -> +5.6pp) to deepseek-v4-pro's -7.1pp (+11.6pp -> +4.5pp); 7 of 7 model lines narrowed week over week. claude-opus-5.5 is the lowest-Brier line (Brier 0.2443); gemini-3.1-pro leads on hit-rate (58.3%). claude-opus-5.5 is the only line whose gap is negative (-1.5pp: stated confidence below its hit-rate).
- **The high-confidence buckets, N>=10 only.** gpt-6.1-sol 70-80 61.4% (n=44), deepseek-v4-pro 70-80 62.5% (n=16), gemini-3.1-pro 70-80 59.7% (n=342), gemini-3.1-pro 80-100 50.0% (n=26). Cells below N=10 are marked insufficient and rank nothing.
- **Trading, kept separate from every calibration table.** 3 of 7 lines finished net-positive; net PnL ran +$9.86 (qwen-3.8-max, best) to -$23.29 (gemini-3.1-pro, worst), field -$43.00 against issue #11's -$352.09.



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 this window's single-build line against issue #11's spliced line. "Was" values are the numbers published in issue #11 for the back-to-back window Oct 12-18.

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 #12 the series has twelve 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. Week A (Oct 12-18, the issue-#11 window) is the spliced week: the three lines flipped at Oct 17, 2026 09:01 UTC inside it (gpt-5.6-sol -> gpt-6.1-sol, claude-opus-5 -> claude-opus-5.5, grok-4.6 -> grok-4.7), so each week-over-week row for those lines joins this week's single-build line to last week's spliced line. This window (Oct 19-25) is the first full week on the new builds. In issue #11 those lines carried the labels "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)" and "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
claude-opus-5.5	700	52.6%	58.0%	56.5	-1.5	0.2443	54.3%-61.6%
gpt-6.1-sol	962	72.3%	57.6%	63.2	+5.6	0.2468	54.4%-60.7%
grok-4.7	836	69.4%	55.9%	60.1	+4.3	0.2472	52.5%-59.2%
qwen-3.8-max	829	62.4%	56.3%	59.9	+3.5	0.2480	52.9%-59.7%
deepseek-v4-pro	891	67.0%	56.5%	61.0	+4.5	0.2481	53.2%-59.7%
claude-fable-5	768	57.7%	56.4%	60.9	+4.5	0.2486	52.8%-59.9%
gemini-3.1-pro	870	66.1%	58.3%	67.5	+9.3	0.2551	55.0%-61.5%
Field (all models)	**5,856**	-	**57.0%**	**61.5**	**+4.5**	**0.2484**	-

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 week-over-week rows join this week's single-build line to issue #11's spliced line (the three lines flipped at Oct 17, 2026 09:01 UTC, inside week A) -- 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	52.5% (217)	58.0% (550)	0.0% (1) *	n/a (0)
claude-opus-5.5	58.9% (660)	42.5% (40)	n/a (0)	n/a (0)
deepseek-v4-pro	55.2% (257)	56.9% (617)	62.5% (16)	n/a (0)
gemini-3.1-pro	100.0% (1) *	57.7% (501)	59.7% (342)	50.0% (26)
gpt-6.1-sol	50.0% (154)	58.9% (764)	61.4% (44)	n/a (0)
grok-4.7	52.0% (325)	58.4% (510)	n/a (0)	n/a (0)
qwen-3.8-max	55.2% (353)	57.0% (468)	100.0% (1) *	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): grok-4.7 0.0% (n=1) *; qwen-3.8-max 57.1% (n=7) *; deepseek-v4-pro 0.0% (n=1) *. Mid-scale ranking: 5 of the 6 models with N>=10 in both 50-60 and 60-70 out-hit 50-60 from 60-70 this week (issue #11: 0 of 6).

Hit-rate by forecast horizon (FH)

Model	1h	4h	1d
claude-fable-5	55.4% (486)	57.3% (239)	62.8% (43)
claude-opus-5.5	54.3% (440)	62.7% (228)	75.0% (32)
deepseek-v4-pro	55.2% (563)	58.0% (286)	61.9% (42)
gemini-3.1-pro	58.4% (570)	57.4% (258)	61.9% (42)
gpt-6.1-sol	56.7% (598)	57.2% (313)	70.6% (51)
grok-4.7	54.1% (519)	55.9% (272)	75.6% (45)
qwen-3.8-max	52.9% (539)	61.0% (254)	75.0% (36)

n in parentheses; * = N<10 (insufficient). 7 of 7 model lines hit better on 1d than on 1h this window (issue #11: 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
qwen-3.8-max	829	45.6%	+$9.86	+$92.76	-$29.21
claude-opus-5.5	700	48.0%	+$7.89	+$77.89	-$33.38
deepseek-v4-pro	891	48.6%	+$0.54	+$89.64	-$36.55
grok-4.7	836	45.7%	-$2.57	+$81.03	-$43.89
gpt-6.1-sol	962	47.0%	-$13.57	+$82.63	-$44.75
claude-fable-5	768	45.2%	-$21.86	+$54.94	-$44.08
gemini-3.1-pro	870	46.3%	-$23.29	+$63.71	-$37.55

3 of 7 model lines finished net-positive and 7 gross-positive this week (issue #11: 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.

Market check: smaller daily moves on lower volume

OBSERVATION — mean |1d move| 2.05% -> 1.35% (-34% rel), BTC realized vol 34.9% -> 27.5% (ann., hourly); field directional accuracy 46.6% -> 57.0% (+10.4pp), 7 of 7 models improved, sim win-rate up for 7 of 7, field sim PnL -$352.09 -> -$43.00 (adjacent calendar weeks Oct 12-18 vs Oct 19-25; descriptive, one pair of weeks, not a claim).
Robustness: raw price-sign accuracy 44.1% -> 56.0% — the same direction as the trade-based hit rule.

Measure	Oct 12-18 (week A)	Oct 19-25 (week B)	Change
Mean |1d move| (5 assets)	2.05%	1.35%	-34% rel
BTC realized vol (ann., hourly)	34.9%	27.5%	-7.4pp
Field directional accuracy	46.6%	57.0%	+10.4pp
Models improving hit-rate	--	7 of 7	--
Raw price-sign accuracy	44.1%	56.0%	+11.9pp
Field sim win-rate	39.2%	46.6%	+7.5pp
Field sim net PnL	-$352.09	-$43.00	--

Week A is exactly the issue-#11 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, inside week A, 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 (2026-10-12..2026-10-18, the published issue-#11 window) holds the cutover: its old-id rows are aggregated under the heads (it reproduces the published issue #11); week B (2026-10-19..2026-10-25) is the first full week on the heads (0 old-id row(s) remapped in it). The grok 4.5 -> 4.6 flip (2026-08-24T09:22:00+00:00) sits eight 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 +4.5pp this week against +15.4pp in issue #11.
- Gap and regime move together. The field hit 57.0% against 46.6% in issue #11 and stated mean confidence read 61.5 against 61.9, so the gap came out at +4.5pp: a field that hits 57.0% instead of 46.6% closes or opens a confidence gap without a single stated number changing.
- Trading direction this week (3 of 7 lines net-positive) is a regime read, not a strategy result; the three net-positive lines were all net-negative one week earlier (issue #11: 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 5,856 scored calls sit inside one market regime. This window had trend on 1 of 7 days (issue #11: 3 of 7); every comparison with issue #11 is a comparison across regimes as well as across weeks. Twelve 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 12): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Oct 19-25) daily coverage is FULL, as it was in issues #4, #5, #6, #7, #8, #9, #10 and #11: 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 480, 1M 478). (2) This window is the first 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, and all 7 lines carry 1,470 rows each. Week A (the issue-#11 window) is the spliced week -- "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) -- so each of the three week-over-week rows joins this week's single-build line to last week's spliced line, and a remapped comparison is not the same model build. (3) No roll-out slot of the flip falls inside this window: the two hourly slots before it (Oct 17, 07:01 and 08:01 UTC) lie in week A, so this window's invalid counter holds none of their rows. (4) Past lineage flips: the Oct 17, 2026 09:01 UTC flip sits one window back (inside the issue-#11 window), the grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits eight 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 #11 after the audit that found the raw table mixes two exchanges; issue #11 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 26 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6, #7, #8, #9, #10 and #11, which printed 71,945 at the Oct 19 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 77,801 is an additive step under one definition; counter deltas against issue #10 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,171
OK in gate	9,171	-- of them directional	5,856
Out of gate (1w / 1M)	480 / 478	-- of them sideways	3,315
Invalid	161	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-10-19.json	Report cutoff	Mon Oct 26, 2026, 16:00 UTC
Generated at (pipeline)	2026-10-27 11:37 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 5,856 scored observations. MARKETMANIA RESEARCH TO DATE (as of Oct 26, 2026 cutoff): 77,801 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 60 published reports.

Issue #12. Weekly Calibration is a living series. Issue #11's open questions -- "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?" and "The best-Brier line changed hands again, to deepseek-v4-pro -- does it hold for a second issue, or change hands again?" -- read this issue as: field gap +15.4pp -> +4.5pp (7 of 7 lines narrowed) on a field base that moved +10.4pp, a seventh issue running in opposite directions, and the best-Brier line changed hands again: claude-opus-5.5 at 0.2443 (issue #11: deepseek-v4-pro at 0.2657). Also on the record: gpt-6.1-sol's 70-80 bucket read 61.4% (n=44) against its own 60-70 at 58.9% (n=764). The claude-opus-5.5, gpt-6.1-sol and grok-4.7 rows are the first full week on the Oct 17 builds; week-over-week rows compare against the spliced issue-#11 week, where those lines carried the labels "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)" and "grok-4.7 (incl. 4.6-era, Oct 12 00:00-Oct 17 09:01 UTC)". The grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits eight 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 -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?
- The best-Brier line changed hands again, to claude-opus-5.5 -- 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 #12	https://marketmania.ai/research/reports/consensus-watch-2026-10-19.pdf
Weekly Model Watch #12	https://marketmania.ai/research/reports/model-watch-2026-10-19.pdf
Weekly Calibration #11	https://marketmania.ai/research/reports/weekly-calibration-2026-10-12.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_2026w43,
  title  = {Weekly Calibration #12: confidence vs. direction-hit, Oct 19-25 2026},
  author = {{MarketMania Research}},
  year   = {2026}, month = {October}, day = {28},
  url    = {https://marketmania.ai/research/reports/weekly-calibration-2026-10-19.pdf},
  note   = {Methodology v1.1, hash e66c7e8c864a2233; source weekly_metrics_2026-10-19.json}
}
```