

Weekly Calibration #5

September 8, 2026 · MarketMania Research · Weekly series · window Aug 31-Sep 6, 2026 UTC

RESEARCH SNAPSHOT

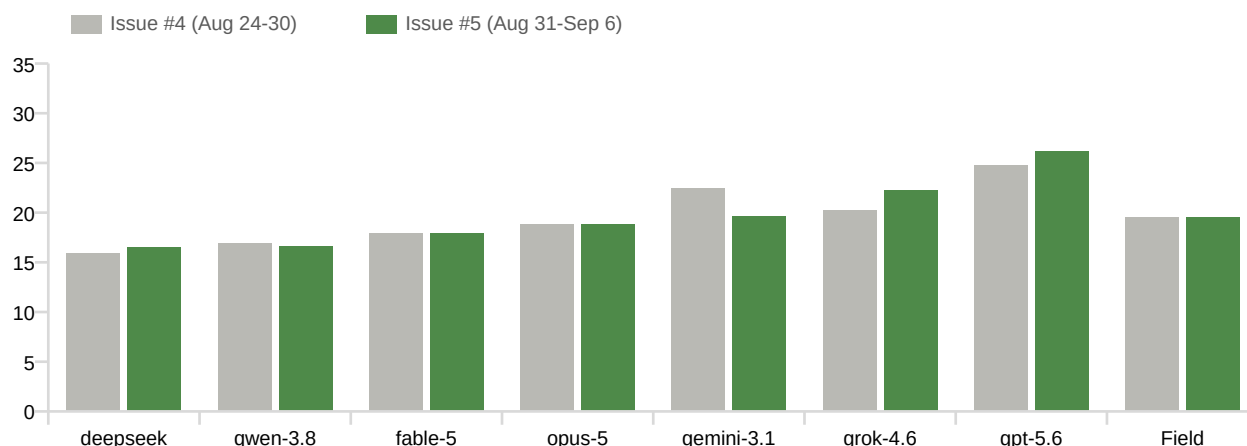
4,520 directional calls scored	7 models (stable lineup)	of 9,264 mature forecasts	window Aug 31-Sep 6, 2026 UTC
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate 1h / 4h / 1d	base field hit 42.8% vs 62.4 stated	cutoff Mon Sep 7, 2026, 16:00 UTC
market: BTC net +3.42% (prior -0.07%)	ann vol 43.4% (was 30.9%)	TOP5 volume \$16.24B , -18% w/w	pairwise corr 0.78 (was 0.85)

KEY FINDING · OBSERVATION (one weekly window)

The field's overconfidence gap moved +19.6pp -> +19.6pp week over week, 2 of 7 model lines narrowed, and field Brier went 0.2884 -> 0.2874.

TL;DR

- **OBSERVATION -- The gap read +19.6pp this week.** Directional calls hit 42.8% against 62.4 stated mean confidence (issue #4: +19.6pp on 42.9% vs 62.6). Field Brier is 0.2874 (was 0.2884) and 0 of 7 model lines scored below 0.25, the score of an uninformative always-50% predictor. One window, descriptive.
- **2 of 7 models narrowed their gap.** Gap moves ran from gemini-3.1-pro's -2.8pp (+22.5pp -> +19.7pp) to grok-4.6's +2.0pp (+20.3pp -> +22.3pp); 2 of 7 model lines narrowed week over week. qwen-3.8-max is the best-calibrated line (Brier 0.2755); gemini-3.1-pro leads on hit-rate (46.5%).
- **The high-confidence buckets, N>=10 only.** deepseek-v4-pro 70-80 52.6% (n=19), gemini-3.1-pro 70-80 42.2% (n=251), gpt-5.6-sol 70-80 35.0% (n=217). 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 -\$81.09 (gemini-3.1-pro, best) to -\$112.37 (gpt-5.6-sol, worst), field -\$669.03 against issue #4's -\$473.59.



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 this window's pure 4.6 line against issue #4's spliced 4.5+4.6 line. "Was" values are the numbers published in issue #4 for the back-to-back window Aug 24-30.

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 #5 the series has five 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 inside the PREVIOUS window (week A, issue #4): every issue-#4 grok number is the lineage splice "grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22)"; this window is pure grok-4.6, so the grok week-over-week row compares a pure-4.6 week with a spliced week.

Calibration by model (sorted by Brier, best first)

Model	N	Cov.	Hit rate	Mean conf	Gap pp	Brier	95% CI
qwen-3.8-max	697	53.5%	42.5%	59.1	+16.6	0.2755	38.9%-46.2%
deepseek-v4-pro	724	54.6%	44.2%	60.6	+16.5	0.2761	40.6%-47.8%
claude-fable-5	628	47.2%	42.8%	60.7	+17.9	0.2795	39.0%-46.7%
claude-opus-5	545	41.1%	42.4%	61.2	+18.8	0.2804	38.3%-46.6%
grok-4.6	476	36.0%	37.6%	59.9	+22.3	0.2883	33.4%-42.0%
gemini-3.1-pro	761	57.3%	46.5%	66.2	+19.7	0.2922	43.0%-50.1%
gpt-5.6-sol	689	51.8%	41.6%	67.8	+26.2	0.3183	38.0%-45.4%
Field (all models)	4,520	-	42.8%	62.4	+19.6	0.2874	-

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 inside the PREVIOUS window (week A, issue #4): every issue-#4 grok number is the lineage splice "grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22)"; this window is pure grok-4.6, so the grok week-over-week row compares a pure-4.6 week with a spliced week.

Calibration curve: hit-rate by stated-confidence bucket

Model	50-60	60-70	70-80	80-100
claude-fable-5	43.2% (199)	42.7% (429)	n/a (0)	n/a (0)
claude-opus-5	46.0% (113)	41.4% (432)	n/a (0)	n/a (0)
deepseek-v4-pro	49.2% (244)	41.2% (461)	52.6% (19)	n/a (0)
gemini-3.1-pro	40.0% (5) *	48.6% (498)	42.2% (251)	57.1% (7) *
gpt-5.6-sol	40.0% (5) *	44.9% (466)	35.0% (217)	0.0% (1) *
grok-4.6	46.2% (186)	32.1% (290)	n/a (0)	n/a (0)
qwen-3.8-max	44.4% (363)	40.1% (324)	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 50.0% (n=8) *. Mid-scale ranking: 0 of the 5 models with N>=10 in both 50-60 and 60-70 out-hit 50-60 from 60-70 this week (issue #4: 1 of 5).

Hit-rate by forecast horizon (FH)

Model	1h	4h	1d
claude-fable-5	45.1% (410)	38.1% (181)	40.5% (37)
claude-opus-5	45.0% (342)	38.5% (174)	34.5% (29)
deepseek-v4-pro	46.9% (469)	38.3% (209)	43.5% (46)
gemini-3.1-pro	48.4% (514)	42.6% (209)	42.1% (38)
gpt-5.6-sol	43.5% (430)	38.5% (221)	39.5% (38)
grok-4.6	42.7% (314)	29.3% (140)	18.2% (22)
qwen-3.8-max	43.5% (446)	39.2% (209)	47.6% (42)

n in parentheses; * = N<10 (insufficient). 1 of 7 model lines hit better on 1d than on 1h this window (issue #4: 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
gemini-3.1-pro	761	37.5%	-\$81.09	-\$4.99	-\$89.79
claude-fable-5	628	36.3%	-\$83.56	-\$20.76	-\$87.66
claude-opus-5	545	34.7%	-\$89.04	-\$34.54	-\$92.14
qwen-3.8-max	697	36.0%	-\$92.91	-\$23.21	-\$98.00
deepseek-v4-pro	724	36.9%	-\$104.50	-\$32.10	-\$109.34
grok-4.6	476	32.1%	-\$105.56	-\$57.96	-\$106.14
gpt-5.6-sol	689	34.5%	-\$112.37	-\$43.47	-\$117.14

0 of 7 model lines finished net-positive and 0 gross-positive this week (issue #4: 0 of 7 net-positive, 3 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: an up week on lower volume

OBSERVATION — mean |1d move| 2.35% -> 2.10% (-11% rel), BTC realized vol 37.9% -> 35.2% (ann., hourly); field directional accuracy 42.9% -> 42.8% (-0.1pp), 1 of 7 models improved, sim win-rate up for 1 of 7, field sim PnL -\$473.59 -> -\$669.03 (adjacent calendar weeks Aug 24-30 vs Aug 31-Sep 6; descriptive, one pair of weeks, not a claim).

Robustness: raw price-sign accuracy 44.2% -> 51.6% — the opposite direction to the trade-based hit rule.

Measure	Aug 24-30 (week A)	Aug 31-Sep 6 (week B)	Change
Mean 1d move (5 assets)	2.35%	2.10%	-11% rel
BTC realized vol (ann., hourly)	37.9%	35.2%	-2.7pp
Field directional accuracy	42.9%	42.8%	-0.1pp
Models improving hit-rate	--	1 of 7	--
Raw price-sign accuracy	44.2%	51.6%	+7.4pp
Field sim win-rate	37.0%	35.6%	-1.4pp
Field sim net PnL	-\$473.59	-\$669.03	--

Week A is exactly the issue-#4 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, inside week A). Week A (2026-08-24..2026-08-30) mixes the 4.5-era (Aug 24 00:00-09:22) with 4.6 and reproduces the published issue #4; 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 +19.6pp this week against +19.6pp in issue #4.
- Gap and regime move together. The field hit 42.8% against 42.9% in issue #4 and stated mean confidence read 62.4 against 62.6, so the gap held at +19.6pp: a field that hits 42.8% instead of 42.9% 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 three 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 4,520 scored calls sit inside one market regime. This window ran 3 trend days of 7 (issue #4: 3 of 7); every comparison with issue #4 is a comparison across regimes as well as across weeks. Five 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 5): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Aug 31-Sep 6) daily coverage is FULL, as it was in issue #4 (the first full week of the series): 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 490). (2) The grok line flipped 4.5 -> 4.6 at Aug 24, 2026 09:22 UTC, inside the PREVIOUS window (week A, issue #4): this window carries 1,470 grok rows and none from the 4.5 era, so every grok number in this issue is a pure grok-4.6 line, while every issue-#4 grok number is the lineage splice "grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22)" -- the grok week-over-week row compares a pure-4.6 week with a spliced week. (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 #4 after the audit that found the raw table mixes two exchanges; issue #4 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 7 16:00 UTC) -- the definition pinned in issue #3 and carried by issue #4, which printed 38,385 at the Aug 31 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 42,905 is an additive step under one definition; counter deltas against issue #3 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,264
OK in gate	9,264	-- of them directional	4,520
Out of gate (1w / 1M)	490 / 490	-- of them sideways	4,744
Invalid	46	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-08-31.json	Report cutoff	Mon Sep 7, 2026, 16:00 UTC
Generated at (pipeline)	2026-09-08 08:08 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 4,520 scored observations. MARKETMANIA RESEARCH TO DATE (as of Sep 7, 2026 cutoff): 42,905 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 24 published reports.

Issue #5. Weekly Calibration is a living series. Issue #4's open questions -- does the field gap stay inside one band or track the base rate one-for-one, and do any two models keep the same Brier ordering for a third issue running? -- read this issue as: field gap +19.6pp -> +19.6pp (2 of 7 lines narrowed) on a field base that moved -0.1pp, and a new best-Brier line (qwen-3.8-max at 0.2755; issue #4: deepseek-v4-pro at 0.2760), so the three-issue ordering question stays open. Also on the record: gpt-5.6-sol's 70-80 bucket read 35.0% (n=217) against its own 60-70 at 44.9% (n=466). The grok row is pure grok-4.6 this issue; issue #4's grok row was the lineage splice "grok-4.6 (incl. 4.5-era, Aug 24 00:00-09:22)". 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: does the field gap hold near +19.6pp for a third issue, or does it track the base rate one-for-one?
- Does the best-Brier line change hands again, or does one model hold it two issues running?
- Monthly series: is the overconfidence gap stable across market regimes (trend vs flat) at monthly n?

Related research

Report	Direct PDF link
Consensus Watch #5	https://marketmania.ai/research/reports/consensus-watch-2026-08-31.pdf
Weekly Model Watch #5	https://marketmania.ai/research/reports/model-watch-2026-08-31.pdf
Weekly Calibration #4	https://marketmania.ai/research/reports/weekly-calibration-2026-08-24.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_2026w36,
  title = {Weekly Calibration #5: confidence vs. direction-hit, Aug 31-Sep 6 2026},
  author = {{MarketMania Research}},
  year  = {2026}, month = {September}, day = {8},
  url   = {https://marketmania.ai/research/reports/weekly-calibration-2026-08-31.pdf},
  note  = {Methodology v1.1, hash e66c7e8c864a2233; source weekly_metrics_2026-08-31.json}
}
```