

Weekly Calibration #6

September 15, 2026 · MarketMania Research · Weekly series · window Sep 7-13, 2026 UTC

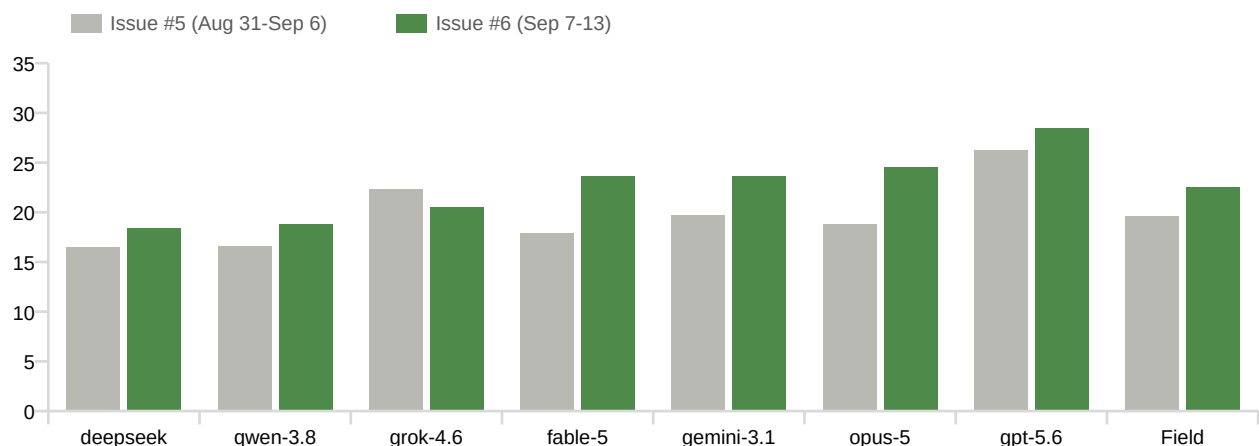
RESEARCH SNAPSHOT			
4,468 directional calls scored	7 models (stable lineup)	of 9,301 mature forecasts	window Sep 7-13, 2026 UTC
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate 1h / 4h / 1d	base field hit 39.6% vs 62.1 stated	cutoff Mon Sep 14, 2026, 16:00 UTC
market: BTC net -4.36% (prior +3.42%)	ann vol 19.7% (was 43.4%)	TOP5 volume \$16.05B , -1% w/w	pairwise corr 0.74 (was 0.78)

KEY FINDING · OBSERVATION (one weekly window)

The field's overconfidence gap moved +19.6pp -> +22.5pp week over week, 1 of 7 model lines narrowed, and field Brier went 0.2874 -> 0.2948.

TL;DR

- **OBSERVATION -- The gap read +22.5pp this week.** Directional calls hit 39.6% against 62.1 stated mean confidence (issue #5: +19.6pp on 42.8% vs 62.4). Field Brier is 0.2948 (was 0.2874) and 0 of 7 model lines scored below 0.25, the score of an uninformative always-50% predictor. One window, descriptive.
- **1 of 7 models narrowed their gap.** Gap moves ran from grok-4.6's -1.8pp (+22.3pp -> +20.5pp) to claude-fable-5's +5.7pp (+17.9pp -> +23.6pp); 1 of 7 model lines narrowed week over week. qwen-3.8-max is the best-calibrated line (Brier 0.2794); gemini-3.1-pro leads on hit-rate (41.8%).
- **The high-confidence buckets, N>=10 only.** gemini-3.1-pro 70-80 38.3% (n=193), gpt-5.6-sol 70-80 42.0% (n=269). 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 -\$114.37 (grok-4.6, best) to -\$163.27 (gpt-5.6-sol, worst), field -\$1018.93 against issue #5's -\$669.03.



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 two pure grok-4.6 weeks: the 4.5 -> 4.6 flip sits two windows back. "Was" values are the numbers published in issue #5 for the back-to-back window Aug 31-Sep 6.

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 #6 the series has six 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 two windows back (issue #4): both weeks of this issue are pure grok-4.6, with no 4.5-era rows in either window, so the grok week-over-week row is the first pure-4.6 against pure-4.6 comparison of the series.

Calibration by model (sorted by Brier, best first)

Model	N	Cov.	Hit rate	Mean conf	Gap pp	Brier	95% CI
qwen-3.8-max	707	53.4%	40.0%	58.8	+18.8	0.2794	36.5%-43.7%
deepseek-v4-pro	727	54.7%	41.4%	59.8	+18.4	0.2825	37.9%-45.0%
grok-4.6	518	39.0%	39.6%	60.0	+20.5	0.2847	35.5%-43.9%
claude-fable-5	624	46.9%	36.5%	60.1	+23.6	0.2921	32.9%-40.4%
claude-opus-5	510	38.4%	36.5%	61.0	+24.5	0.2932	32.4%-40.7%
gemini-3.1-pro	673	50.6%	41.8%	65.4	+23.6	0.3057	38.1%-45.5%
gpt-5.6-sol	709	53.4%	40.1%	68.5	+28.4	0.3230	36.5%-43.7%
Field (all models)	4,468	-	39.6%	62.1	+22.5	0.2948	-

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 two windows back (issue #4): both weeks of this issue are pure grok-4.6, with no 4.5-era rows in either window, so the grok week-over-week row is the first pure-4.6 against pure-4.6 comparison of the series.

Calibration curve: hit-rate by stated-confidence bucket

Model	50-60	60-70	70-80	80-100
claude-fable-5	41.0% (234)	33.9% (390)	n/a (0)	n/a (0)
claude-opus-5	36.8% (117)	36.4% (393)	n/a (0)	n/a (0)
deepseek-v4-pro	47.7% (308)	37.1% (410)	12.5% (8) *	n/a (0)
gemini-3.1-pro	16.7% (6) *	43.3% (473)	38.3% (193)	n/a (0)
gpt-5.6-sol	36.4% (11)	39.0% (428)	42.0% (269)	0.0% (1) *
grok-4.6	43.2% (199)	37.2% (317)	0.0% (1) *	n/a (0)
qwen-3.8-max	43.5% (386)	35.7% (308)	0.0% (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): qwen-3.8-max 41.7% (n=12); deepseek-v4-pro 100.0% (n=1) *; grok-4.6 100.0% (n=1) *; gemini-3.1-pro 100.0% (n=1) *. 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 #5: 0 of 5).

Hit-rate by forecast horizon (FH)

Model	1h	4h	1d
claude-fable-5	41.1% (387)	30.5% (203)	20.6% (34)
claude-opus-5	42.1% (299)	29.9% (177)	20.6% (34)
deepseek-v4-pro	45.3% (439)	39.7% (237)	15.7% (51)
gemini-3.1-pro	46.0% (437)	36.3% (201)	20.0% (35)
gpt-5.6-sol	44.5% (434)	34.9% (235)	22.5% (40)
grok-4.6	42.5% (327)	36.1% (169)	22.7% (22)
qwen-3.8-max	43.7% (439)	36.8% (231)	16.2% (37)

n in parentheses; * = N<10 (insufficient). 0 of 7 model lines hit better on 1d than on 1h this window (issue #5: 1 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-4.6	518	32.4%	-\$114.37	-\$62.57	-\$114.37
gemini-3.1-pro	673	33.7%	-\$131.95	-\$64.65	-\$134.16
claude-opus-5	510	29.0%	-\$137.47	-\$86.47	-\$137.47
deepseek-v4-pro	727	34.2%	-\$152.22	-\$79.52	-\$152.76
qwen-3.8-max	707	32.7%	-\$158.82	-\$88.12	-\$159.24
claude-fable-5	624	29.3%	-\$160.85	-\$98.45	-\$160.85
gpt-5.6-sol	709	32.3%	-\$163.27	-\$92.37	-\$163.27

0 of 7 model lines finished net-positive and 0 gross-positive this week (issue #5: 0 of 7 net-positive, 0 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: a down week on a quieter tape

OBSERVATION — mean |1d move| 2.10% -> 1.35% (-36% rel), BTC realized vol 35.2% -> 31.4% (ann., hourly); field directional accuracy 42.8% -> 39.6% (-3.3pp), 1 of 7 models improved, sim win-rate up for 1 of 7, field sim PnL -\$669.03 -> -\$1018.93 (adjacent calendar weeks Aug 31-Sep 6 vs Sep 7-13; descriptive, one pair of weeks, not a claim).

Robustness: raw price-sign accuracy 51.6% -> 38.7% — the same direction as the trade-based hit rule.

Measure	Aug 31-Sep 6 (week A)	Sep 7-13 (week B)	Change
Mean 1d move (5 assets)	2.10%	1.35%	-36% rel
BTC realized vol (ann., hourly)	35.2%	31.4%	-3.8pp
Field directional accuracy	42.8%	39.6%	-3.3pp
Models improving hit-rate	--	1 of 7	--
Raw price-sign accuracy	51.6%	38.7%	-12.8pp
Field sim win-rate	35.6%	32.1%	-3.5pp
Field sim net PnL	-\$669.03	-\$1018.93	--

Week A is exactly the issue-#5 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, two windows back, inside issue #4's window). Week A (2026-08-31..2026-09-06) is pure grok-4.6 (it reproduces the published issue #5); 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 +22.5pp this week against +19.6pp in issue #5.
- Gap and regime move together. The field hit 39.6% against 42.8% in issue #5 and stated mean confidence read 62.1 against 62.4, so the gap held at +22.5pp: a field that hits 39.6% instead of 42.8% 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 four 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 4,468 scored calls sit inside one market regime. This window ran 2 trend days of 7 (issue #5: 3 of 7); every comparison with issue #5 is a comparison across regimes as well as across weeks. Six 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 6): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Sep 7-13) daily coverage is FULL, as it was in issues #4 and #5: 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, two windows back (inside the issue-#4 window): this window carries 1,470 grok rows and none from the 4.5 era, and neither does week A (the issue-#5 window), so every grok number in this issue and in issue #5 is a pure grok-4.6 line -- the grok week-over-week row is the first pure-4.6 against pure-4.6 comparison of the series. (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 #5 after the audit that found the raw table mixes two exchanges; issue #5 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 14 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4 and #5, which printed 42,905 at the Sep 7 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 47,373 is an additive step under one definition; counter deltas against issue #4 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,301
OK in gate	9,301	-- of them directional	4,468
Out of gate (1w / 1M)	490 / 490	-- of them sideways	4,833
Invalid	9	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-09-07.json	Report cutoff	Mon Sep 14, 2026, 16:00 UTC
Generated at (pipeline)	2026-09-15 15:24 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 4,468 scored observations. MARKETMANIA RESEARCH TO DATE (as of Sep 14, 2026 cutoff): 47,373 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 28 published reports.

Issue #6. Weekly Calibration is a living series. Issue #5's open questions -- does the field gap hold near +19.6pp for a third issue or track the base rate one-for-one, and does the best-Brier line change hands again? -- read this issue as: field gap +19.6pp -> +22.5pp (1 of 7 lines narrowed) on a field base that moved -3.3pp, and the best-Brier line did not change hands: qwen-3.8-max holds it two issues running (0.2794, issue #5: qwen-3.8-max 0.2755). Also on the record: gpt-5.6-sol's 70-80 bucket read 42.0% (n=269) against its own 60-70 at 39.0% (n=428). The grok row is pure grok-4.6 this issue and was pure grok-4.6 in issue #5; the 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits two 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 +2.9pp while the field base moved -3.3pp -- does the gap keep tracking the base one-for-one, or does it hold a level of its own?
- Does qwen-3.8-max hold the best-Brier line for a third issue, or does it 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 #6	https://marketmania.ai/research/reports/consensus-watch-2026-09-07.pdf
Weekly Model Watch #6	https://marketmania.ai/research/reports/model-watch-2026-09-07.pdf
Weekly Calibration #5	https://marketmania.ai/research/reports/weekly-calibration-2026-08-31.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_2026w37,
  title = {Weekly Calibration #6: confidence vs. direction-hit, Sep 7-13 2026},
  author = {{MarketMania Research}},
  year  = {2026}, month = {September}, day = {15},
  url   = {https://marketmania.ai/research/reports/weekly-calibration-2026-09-07.pdf},
  note  = {Methodology v1.1, hash e66c7e8c864a2233; source weekly_metrics_2026-09-07.json}
}
```