

Weekly Calibration #2

August 19, 2026 · window Aug 10-16 · MarketMania Research · Weekly series

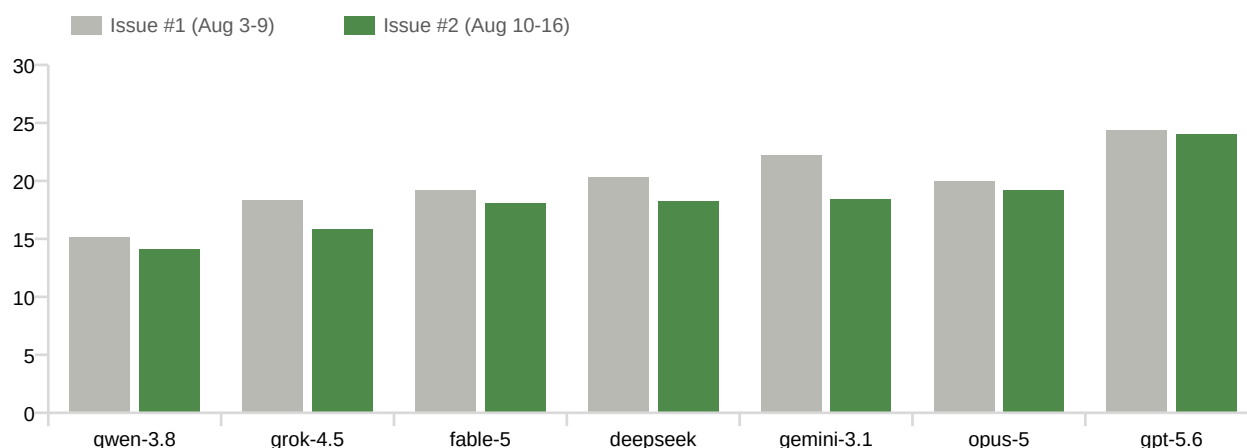
RESEARCH SNAPSHOT			
window Aug 10-16, 2026 (UTC)	7 models (stable lineup)	3,670 directional calls scored of 9,307 mature	FH gate 1h / 4h / 1d
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	cutoff Mon 16:00 UTC (frozen)	methodology v1.1 (2026-08-10)	hash e66c7e8c864a2233
market: BTC net -3.08% (prior +2.09%)	ann. vol 10.0% (was 11.4%)	TOP5 volume \$7.75B , -8.9% w/w	pairwise corr 0.52 (was 0.40)

KEY FINDING · OBSERVATION (one weekly window)

Overconfidence eased across the whole field -- every one of the 7 models narrowed its gap, 20.4pp -> 18.3pp pooled -- yet high confidence still ranked nothing, with one exception: gemini-3.1-pro's 70-80 bucket hit 50.7%, the field's first working high-confidence bucket.

TL;DR

- **OBSERVATION -- A second flat, hard week.** Directional calls hit **43.8%** against **62.0** stated mean confidence -- a **+18.3pp** overconfidence gap (issue #1: +20.4pp). Every model's Brier score again topped 0.25, worse than an uninformative always-50% predictor, for the second week running.
- **Every model narrowed its gap week-over-week** -- the series' first w/w delta, and it is field-wide: from gemini-3.1-pro's -3.8pp improvement to gpt-5.6-sol's -0.4pp. qwen-3.8-max is best-calibrated again (gap +14.1pp, Brier 0.2688), but the hit-rate lead moved: gemini-3.1-pro tops the field at 47.0% [42.9%, 51.2%].
- **Confidence still failed to rank outcomes in the middle of the scale:** 4 of the 5 models with sufficient N (≥ 10) in both 50-60 and 60-70 did not out-hit 50-60 from 60-70 (issue #1: 4 of 6). Higher up, issue #1's inversion flags split: gemini-3.1-pro un-inverted -- its 70-80 bucket hit **50.7%** (n=152), the only high-confidence cell in the field above base -- while deepseek-v4-pro's inversion deepened, 35.9% -> **26.3%** (n=38).
- **Trading stayed uniformly negative -- and got cleaner about it.** All 7 models finished net-negative AND gross-negative (issue #1 had one gross-positive outlier); net PnL ran -\$63.19 (claude-opus-5, best) to -\$86.17 (gpt-5.6-sol, worst).



Overconfidence gap (mean stated confidence minus hit-rate, percentage points) by model, week over week -- sorted by this week's gap, best (smallest) to worst. Every bar shrank: the field gap narrowed from +20.4pp to +18.3pp. gpt-5.6-sol remains the most overconfident model in the field at +24.0pp; the legacy qwen-3.7-max row retired with issue #1, so both weeks plot the same 7 models.

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. Weekly Calibration is a descriptive, single-week reading of that question -- and with issue #2 the series gains its first week-over-week deltas. Durability claims (does a model's calibration hold up over time) still need several weeks of data and belong in the Monthly series, not here.

Calibration by model (sorted by Brier, best first)

Model	N	Coverage	Hit rate	Mean conf	Gap pp	Brier	95% CI
qwen-3.8-max	572	43.0%	44.1%	58.2	+14.1	0.2688	40.0%-48.1%
grok-4.5	620	46.6%	44.0%	59.8	+15.8	0.2727	40.2%-48.0%
claude-fable-5	472	35.5%	41.7%	59.8	+18.1	0.2782	37.4%-46.2%
claude-opus-5	361	27.1%	41.5%	60.7	+19.2	0.2814	36.6%-46.7%
gemini-3.1-pro	555	41.7%	47.0%	65.4	+18.4	0.2848	42.9%-51.2%
deepseek-v4-pro	454	34.2%	42.5%	60.7	+18.2	0.2879	38.0%-47.1%
gpt-5.6-sol	636	47.8%	44.2%	68.2	+24.0	0.3090	40.4%-48.1%
Field (all models)	3,670	-	43.8%	62.0	+18.3	0.2836	-

N = directional calls scored this window. Coverage = scored / mature calls available to that model. Gap pp = mean stated confidence minus hit-rate, in percentage points (positive = overconfident). 95% CI is the descriptive Wilson interval on hit-rate (dependent observations -- see Limitations). Field row has no single-model coverage or CI. First issue with no legacy row: qwen-3.7-max retired Aug 5.

Calibration curve: hit-rate by stated-confidence bucket

Model	50-60	60-70	70-80	80-100
qwen-3.8-max	45.8% (360)	40.5% (200)	n/a (0)	n/a (0)
grok-4.5	43.0% (367)	45.8% (251)	0.0% (2) *	n/a (0)
claude-fable-5	43.5% (200)	40.4% (272)	n/a (0)	n/a (0)
claude-opus-5	43.5% (108)	40.7% (253)	n/a (0)	n/a (0)
gemini-3.1-pro	40.0% (5) *	45.7% (396)	50.7% (152)	50.0% (2) *
deepseek-v4-pro	45.3% (181)	42.2% (230)	26.3% (38)	0.0% (1) *
gpt-5.6-sol	0.0% (1) *	47.0% (419)	39.2% (214)	0.0% (2) *

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>0 only, footnote): qwen-3.8-max 50.0% (n=12); deepseek-v4-pro 100.0% (n=4) *. All other models had zero sub-50 calls this week. The two cells worth the read: gemini-3.1-pro's 70-80 at 50.7% (n=152) -- issue #1 had it inverted at 36.3% -- and deepseek-v4-pro's 70-80 at 26.3% (n=38), deeper inverted than last week's 35.9%.

Hit-rate by forecast horizon (FH)

Model	1h	4h	1d
qwen-3.8-max	45.5% (365)	42.2% (180)	37.0% (27)
grok-4.5	46.2% (400)	40.0% (190)	40.0% (30)
claude-fable-5	44.0% (311)	39.4% (132)	27.6% (29)
claude-opus-5	45.2% (228)	36.4% (107)	30.8% (26)
gemini-3.1-pro	48.2% (365)	44.7% (161)	44.8% (29)
deepseek-v4-pro	44.2% (276)	42.0% (157)	23.8% (21)
gpt-5.6-sol	45.4% (401)	41.4% (203)	46.9% (32)

n in parentheses. * = N<10 (insufficient); # = N<20, thin but above the N>=10 floor. No cell needed either flag this week -- every 1d cell cleared n=20 for the first time in the series. The 1d horizon stayed the field's hardest: 6 of 7 models hit worse on 1d than on 1h, with claude-fable-5 (27.6%) and deepseek-v4-pro (23.8%) at the bottom and gpt-5.6-sol the 1d leader at 46.9%.

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	361	29.1%	-\$63.19	-\$27.09	-\$64.58
deepseek-v4-pro	454	28.2%	-\$70.86	-\$25.46	-\$71.20
gemini-3.1-pro	555	31.2%	-\$71.64	-\$16.14	-\$72.62
qwen-3.8-max	572	29.7%	-\$72.62	-\$15.42	-\$77.53
claude-fable-5	472	28.2%	-\$77.13	-\$29.93	-\$77.13
grok-4.5	620	29.7%	-\$85.46	-\$23.46	-\$88.68
gpt-5.6-sol	636	30.7%	-\$86.17	-\$22.57	-\$88.98

All seven models finished negative on BOTH gross and net PnL this week. Issue #1 had one gross-positive outlier (qwen-3.8-max, +\$1.38 before costs); this week the field closed the question -- the spread ran from claude-opus-5's -\$63.19 net (best) to gpt-5.6-sol's -\$86.17 (worst). WR = trade win rate. PnL and max drawdown (Max DD) in USD, ranked best-to-worst net PnL. Per methodology, this table is never combined with the calibration/prediction tables above.

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. The toggle lives at marketmania.ai/indices.

Practical implications

- Stated confidence remains a style signal, not a probability: 4 of 5 sufficient-N models again failed to out-hit their 50-60 bucket from 60-70. Nothing this week changes issue #1's advice.
- The one working high-confidence cell -- gemini-3.1-pro's 70-80 at 50.7% -- is one week old and sits next to deepseek-v4-pro's 26.3% in the same bucket. Treat it as the thing to watch, not a filter to trade.
- The field-wide gap narrowing (all 7 models, first w/w delta of the series) is two data points, not a trend. Whether it is drift toward humility or just this week's regime is exactly what the monthly report will test.

Limitations & notes

- 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 (a single market wave can move many forecasts together), so read them as a range, not a formal coverage guarantee.
- All 3,670 scored calls sit inside one market regime -- 6 of 7 days this window were flat (|BTC daily move| < 1%). A single week cannot separate a calibration pattern from the week's specific conditions.
- Week-over-week deltas begin with this issue, but two points cannot separate drift from noise -- several issues can. No durability claim is made.
- Models report confidence at discrete levels, not a continuous scale; the 50-60 / 60-70 / 70-80 / 80-100 buckets reflect those natural breakpoints, not an arbitrary binning choice.
- Underlying price series are reconstructed from trade entry prices (median per symbol-slot), not an independent tick feed.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	9,380	Mature (scored pool)	9,307
OK in gate	9,307	-- of them directional	3,670
Out of gate (1w)	69	-- sideways	5,637
Invalid	4	Pending (next issue)	0
Uptime (slots seen/expected)	168/168 · 42/42 · 42/42 · 7/7 · 7/7	Late closes	0
Source file	weekly_metrics_2026-08-10.json	Report cutoff	Mon Aug 17, 16:00 UTC
Generated at (pipeline)	2026-08-18 08:37 UTC	Methodology	v1.1 · hash e66c7e8c864a2233

First perfect-uptime week of the series: every 1h, 4h and 1d slot grid ran full (issue #1: 165/168 on the hourly grid).

THIS REPORT: 3,670 scored observations. **MARKETMANIA RESEARCH TO DATE** (as of Aug 17, 2026 cutoff): 23,458 resolved forecasts since Jul 11 · 9 models tracked (7 current frontier + 2 archived legacy) · 5 assets · 5 forecast horizons · hourly cadence · 9 published research reports incl. this wave.

Issue #2. Weekly Calibration is a living series; week-over-week deltas begin with this issue. Issue #1's open questions -- does the confidence-bucket inversion persist, and does the 60-70 vs 50-60 non-ranking repeat? -- closed as 'split' and 'yes': gemini-3.1-pro un-inverted its 70-80 bucket, deepseek-v4-pro's inversion deepened, and 4 of 5 sufficient-N models again failed to rank. Engine 1.1 powers the sandbox since Aug 18 (after this window closed); no cross-engine PnL comparisons are claimed.

What we're testing next

- **Next issue:** does gemini-3.1-pro's 70-80 bucket keep outperforming at $n > 150$ -- and does deepseek-v4-pro's inversion survive a third week? Does the field-wide gap keep narrowing?
- **Monthly test (Sep 2):** is the overconfidence gap stable across market regimes (trend vs flat) at monthly n ?

Related research

Consensus Watch #2 -- <https://marketmania.ai/research/reports/consensus-watch-2026-08-10.pdf>

Weekly Model Watch #2 -- <https://marketmania.ai/research/reports/model-watch-2026-08-10.pdf>

Config Watch #2 -- <https://marketmania.ai/research/reports/config-watch-2026-08-12.pdf>

Weekly Calibration #1 -- <https://marketmania.ai/research/reports/weekly-calibration-2026-08-03.pdf>

The three weekly reports publish together as one issue each week; Config Watch follows on its own cycle.

CITE THIS REPORT

```
@misc{mm_weekly_calibration_2026w33,
  title = {Weekly Calibration #2: confidence vs. direction-hit, Aug 10-16 2026},
  author = {{MarketMania Research}},
  year = {2026}, month = {August}, day = {19},
  url = {https://marketmania.ai/research/reports/weekly-calibration-2026-08-10.pdf},
  note = {Methodology v1.1, hash e66c7e8c864a2233; source weekly_metrics_2026-08-10.json}
}
```