

Weekly Calibration #3

August 26, 2026 · MarketMania Research · Weekly series · window Aug 17-23, 2026 UTC

RESEARCH SNAPSHOT

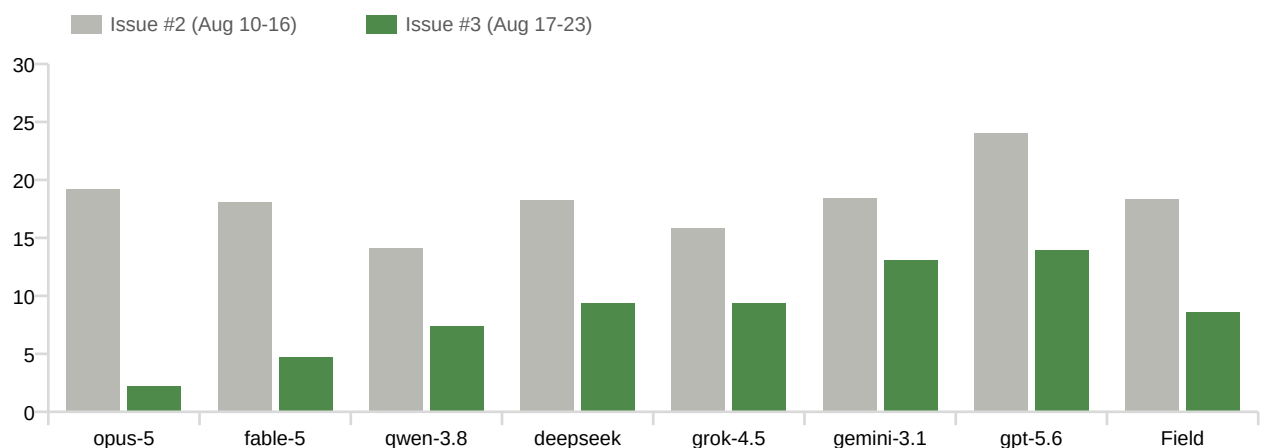
5,130 directional calls scored	7 models (stable lineup)	of 9,260 mature forecasts	window Aug 17-23, 2026 UTC
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate 1h / 4h / 1d	base field hit 54.1% vs 62.7 stated	cutoff Mon Aug 24, 2026, 16:00 UTC
market: BTC net +23.58% (prior -3.08%)	ann vol 71.1% (was 10.0%)	TOP5 volume \$25.97B , +235% w/w	pairwise corr 0.81 (was 0.52)

KEY FINDING · OBSERVATION (one weekly window)

The overconfidence gap halved -- field +18.3pp -> +8.6pp, with all 7 models narrowing for a second straight week -- and high confidence finally ranked something: gemini-3.1-pro's 70-80 bucket held at 55.1% on doubled volume while its 80-100 bucket opened at 84.6%.

TL;DR

- **OBSERVATION -- The gap halved in a live tape.** Directional calls hit 54.1% against 62.7 stated mean confidence -- a +8.6pp overconfidence gap (issue #2: +18.3pp). Field Brier improved to 0.2559 (was 0.2836) and, for the first time in the series, 2 of 7 models scored below 0.25 -- better than an uninformative always-50% predictor.
- **Every model narrowed its gap again** -- the second straight field-wide narrowing, from claude-opus-5's -17.0pp move (19.2 -> 2.2) to gemini-3.1-pro's -5.3pp (18.4 -> 13.1). claude-opus-5 is now both the best-calibrated model (Brier 0.2419) and the hit-rate leader (59.1%).
- **The high-confidence follow-up came back positive.** gemini-3.1-pro's 70-80 bucket held at 55.1% on n=307 (issue #2: 50.7% on n=152) and its 80-100 bucket opened at 84.6% (n=13, insufficient). deepseek-v4-pro's inversion un-inverted: 26.3% -> 64.7% (n=17, thin). The counter-example is gpt-5.6-sol, heavy in 70-80 at 48.1% (n=185) -- below its own 60-70 bucket.
- **Trading flipped with the tape.** All 7 models finished net-positive AND gross-positive (issue #2: all 7 negative on both); net PnL ran +\$197.44 (claude-opus-5, best) to +\$54.16 (grok-4.5, worst), field +\$889.37 against issue #2's -\$527.07.



Overconfidence gap in percentage points (mean stated confidence minus hit-rate), week over week, sorted by this week's gap. Every model narrowed for the second straight week; the same 7 models cover both columns. 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. "Was" values are the numbers published in issue #2 for the back-to-back window Aug 10-16.

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 #3 the series has three points on every model's gap -- enough to see a direction, not enough to claim one: this window was also the first live-tape week of the series, and calibration and market regime move together. 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	672	50.5%	59.1%	61.2	+2.2	0.2419	55.3%-62.7%
claude-fable-5	779	58.6%	56.5%	61.2	+4.7	0.2464	53.0%-59.9%
qwen-3.8-max	706	55.1%	52.3%	59.7	+7.4	0.2540	48.6%-55.9%
grok-4.5	710	53.4%	51.4%	60.8	+9.4	0.2557	47.7%-55.1%
deepseek-v4-pro	777	58.4%	51.6%	61.0	+9.4	0.2574	48.1%-55.1%
gemini-3.1-pro	798	60.1%	54.1%	67.2	+13.1	0.2646	50.7%-57.6%
gpt-5.6-sol	688	51.7%	53.6%	67.5	+13.9	0.2710	49.9%-57.3%
Field (all models)	5,130	-	54.1%	62.7	+8.6	0.2559	-

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. Brier improved for all 7 models, and the Brier and hit-rate orderings agree at the top for the first time in the series.

Calibration curve: hit-rate by stated-confidence bucket

Model	50-60	60-70	70-80	80-100
claude-fable-5	54.1% (183)	57.2% (596)	n/a (0)	n/a (0)
claude-opus-5	63.5% (126)	58.1% (546)	n/a (0)	n/a (0)
deepseek-v4-pro	45.8% (236)	53.8% (524)	64.7% (17)	n/a (0)
gemini-3.1-pro	50.0% (2) *	52.7% (476)	55.1% (307)	84.6% (13)
gpt-5.6-sol	50.0% (2) *	55.7% (501)	48.1% (185)	n/a (0)
grok-4.5	45.4% (315)	55.8% (389)	83.3% (6) *	n/a (0)
qwen-3.8-max	50.9% (318)	54.4% (379)	33.3% (3) *	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 0.0% (n=6) *; all other models had zero sub-50 calls. The cells worth the read: gemini-3.1-pro's 70-80 at 55.1% on n=307 with its 80-100 bucket open for the first time (84.6%, n=13, insufficient); gpt-5.6-sol's 70-80 at 48.1% (n=185), the field's only high-confidence bucket below its own 60-70; deepseek-v4-pro's 70-80 back above base at 64.7% on a thin n=17. Mid-scale ranking improved but did not resolve: 4 of the 5 models with N of 10 or more in both 50-60 and 60-70 out-hit 50-60 from 60-70 this week (issue #2: 1 of 5).

Hit-rate by forecast horizon (FH)

Model	1h	4h	1d
claude-fable-5	51.9% (472)	61.0% (254)	75.5% (53)
claude-opus-5	53.0% (385)	65.3% (236)	76.5% (51)
deepseek-v4-pro	47.9% (463)	56.1% (264)	62.0% (50)
gemini-3.1-pro	49.2% (492)	59.5% (257)	75.5% (49)
gpt-5.6-sol	48.8% (420)	62.0% (221)	57.5% (47)
grok-4.5	47.5% (425)	56.2% (242)	62.8% (43)
qwen-3.8-max	47.4% (432)	58.5% (219)	65.5% (55)

n in parentheses; no cell needed the N<10 or N<20 flag this week. The horizon ordering inverted against issue #2: 1d was the field's hardest horizon then and is its easiest now -- 7 of 7 models hit better on 1d than on 1h, led by claude-opus-5 (76.5%), with gemini-3.1-pro and claude-fable-5 both at 75.5% and gpt-5.6-sol weakest at 57.5%. In a week whose mean |1d| move was 4.28%, the daily horizon simply had more signal to read.

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	672	50.7%	+\$197.44	+\$264.64	-\$55.65
claude-fable-5	779	48.4%	+\$180.33	+\$258.23	-\$57.23
gemini-3.1-pro	798	46.9%	+\$168.37	+\$248.17	-\$84.80
qwen-3.8-max	706	44.5%	+\$111.62	+\$182.22	-\$39.89
gpt-5.6-sol	688	44.8%	+\$111.61	+\$180.41	-\$75.04
deepseek-v4-pro	777	46.2%	+\$65.84	+\$143.54	-\$45.00
grok-4.5	710	42.3%	+\$54.16	+\$125.16	-\$57.04

All seven models finished positive on BOTH gross and net PnL this week -- the exact mirror of issue #2, where all seven were negative on both. Win rate rose for 7 of 7 (field 29.7% -> 46.3%); max drawdown stayed real, -\$39.89 to -\$84.80. Ranked best-to-worst net PnL. Per methodology this table is never combined with the calibration tables above -- and a week when the tape moves is a week when a directional book makes money; that is regime, not 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 -- it went live inside this window and separating it from the regime change is not a weekly question. The toggle lives at marketmania.ai/indices.

Market check: the tape woke up

OBSERVATION — the tape sped up and accuracy followed: mean |1d move| 0.73% -> 4.28% (483% rel), BTC realized vol 19.3% -> 58.7% (ann., hourly); field directional accuracy 43.8% -> 54.1% (+10.3pp), 7 of 7 models improved, sim win-rate up for 7 of 7, field sim PnL -\$527 -> +\$889 (adjacent calendar weeks Aug 10-16 vs Aug 17-23; descriptive, one pair of weeks, not a claim).

Robustness: raw price-sign accuracy 40.1% -> 53.5% — consistent with the trade-based hit rule.

Measure	Aug 10-16 (week A)	Aug 17-23 (week B)	Change
Mean 1d move (5 assets)	0.73%	4.28%	+483% rel
BTC realized vol (ann., hourly)	19.3%	58.7%	+39.4pp
Field directional accuracy	43.8%	54.1%	+10.3pp
Models improving hit-rate	--	7 of 7	--
Raw price-sign accuracy	40.1%	53.5%	+13.4pp
Field sim win-rate	29.7%	46.3%	+16.6pp
Field sim net PnL	-\$527.07	+\$889.37	--

Week A is exactly the issue-#2 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 71.1% is the audited daily-candle figure -- different estimators, both reported as measured.

Practical implications

- Stated confidence moved from useless to weakly informative: 4 of 5 sufficient-N models out-ranked 50-60 with 60-70 this week (issue #2: 1 of 5), and the field's two heaviest 70-80 buckets split. It is still not a probability.
- The gap halved in the series' first live-tape week. The two facts are not separable on one week of data: a field that hits 54.1% instead of 43.8% closes a confidence gap without changing a single stated number.
- All 7 net-positive is a regime read, not a strategy result. The same seven were all net-negative one week earlier at the same stated confidences.

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 5,130 scored calls sit inside one market regime, and this window's regime is the opposite of issue #2's: 5 of 7 days were trend days against 1 of 7 last week. Three 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 3): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Aug 17-23) daily coverage is PARTIAL by design: the 1w series has the Mon Aug 17 anchor plus daily slots on Aug 22-23 only, and the 1M series has a daily slot on Aug 23 only; both sit outside this report's FH gate (1h/4h/1d) and appear only in the exclusions counter (1w 209, 1M 68). (2) After the report window — from Aug 24, 2026 — the grok line runs Grok 4.6; every grok forecast in this window and in the week-2 comparison is grok-4.5. (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 this issue, after an audit found the raw table mixes two exchanges; issue #2 row is as published.
- Research-to-date counter pinned from this issue: directional forecasts resolved inside the FH gate (1h/4h/1d) since Jul 11, counted at the issue cutoff (Aug 24 16:00 UTC). Issue #2 printed 23,458 under an earlier, unpinned definition; treat cross-issue counter deltas across the pin as definitional, not additive.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	9,590	Mature (scored pool)	9,260
OK in gate	9,260	-- of them directional	5,130
Out of gate (1w / 1M)	209 / 68	-- of them sideways	4,130
Invalid	53	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-08-17.json	Report cutoff	Mon Aug 24, 2026, 16:00 UTC
Generated at (pipeline)	2026-08-26 05:22 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 5,130 scored observations. MARKETMANIA RESEARCH TO DATE (as of Aug 24, 2026 cutoff): 33,809 directional forecasts resolved since Jul 11 · 11 models tracked (7 current + 4 archived legacy) · 5 assets · 5 horizons · hourly · 12 published reports.

Issue #3. Weekly Calibration is a living series. Issue #2's open questions -- does gemini-3.1-pro's 70-80 bucket keep outperforming at $n > 150$, does deepseek-v4-pro's inversion survive a third week, and does the field-wide gap keep narrowing? -- closed 'yes' (55.1% on $n=307$), 'no' (26.3% -> 64.7% on a thin $n=17$) and 'yes' (all 7 narrowed again, field 18.3 -> 8.6). After the report window — from Aug 24, 2026 — the grok line runs Grok 4.6; every grok forecast in this window and in the week-2 comparison is grok-4.5. Engine 1.1 has powered the sandbox since Aug 18, i.e. from day 2 of this window; no cross-engine PnL comparisons are claimed.

What we're testing next

- Next issue: does the halved gap survive a flat week, or was +8.6pp the tape rather than the models?
- Does gpt-5.6-sol's 70-80 bucket (48.1% on $n=185$) stay below its own 60-70 -- the field's new inversion candidate?
- Monthly test (Sep 2): is the overconfidence gap stable across market regimes (trend vs flat) at monthly n ?

Related research

Report	Direct PDF link
Consensus Watch #3	https://marketmania.ai/research/reports/consensus-watch-2026-08-17.pdf
Weekly Model Watch #3	https://marketmania.ai/research/reports/model-watch-2026-08-17.pdf
Weekly Calibration #2	https://marketmania.ai/research/reports/weekly-calibration-2026-08-10.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_2026w34,
  title = {Weekly Calibration #3: confidence vs. direction-hit, Aug 17-23 2026},
  author = {{MarketMania Research}},
  year  = {2026}, month = {August}, day = {26},
  url   = {https://marketmania.ai/research/reports/weekly-calibration-2026-08-17.pdf},
  note  = {Methodology v1.1, hash e66c7e8c864a2233; source weekly_metrics_2026-08-17.json}
}
```