

Weekly Calibration #7

September 22, 2026 · MarketMania Research · Weekly series · window Sep 14-20, 2026 UTC

RESEARCH SNAPSHOT

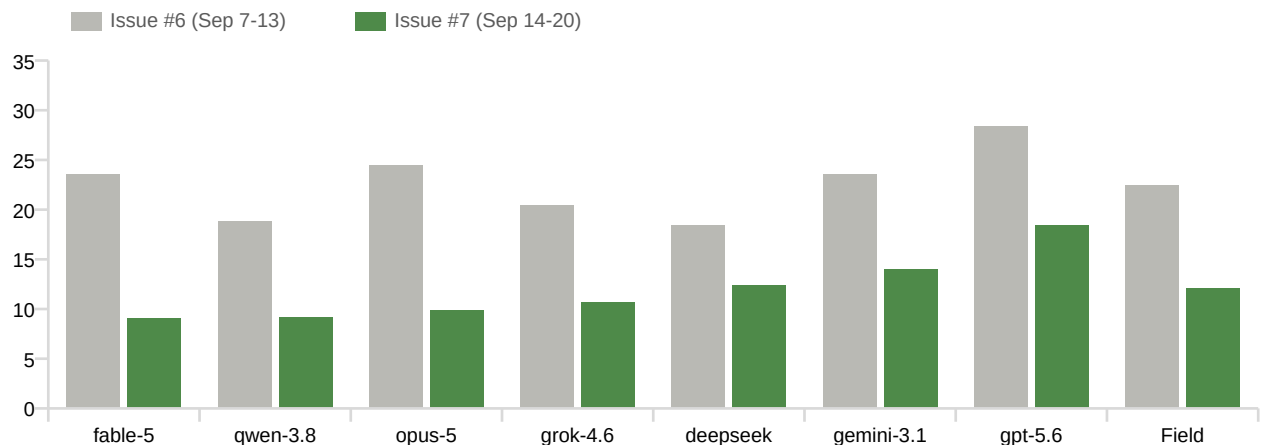
5,598 directional calls scored	7 models (stable lineup)	of 9,308 mature forecasts	window Sep 14-20, 2026 UTC
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate 1h / 4h / 1d	base field hit 50.7% vs 62.8 stated	cutoff Mon Sep 21, 2026, 16:00 UTC
market: BTC net +5.64% (prior -4.36%)	ann vol 51.0% (was 19.7%)	TOP5 volume \$18.74B , +17% w/w	pairwise corr 0.93 (was 0.74)

KEY FINDING · OBSERVATION (one weekly window)

The field's overconfidence gap moved +22.5pp -> +12.1pp week over week, 7 of 7 model lines narrowed, and field Brier went 0.2948 -> 0.2669.

TL;DR

- **OBSERVATION -- The gap read +12.1pp this week.** Directional calls hit 50.7% against 62.8 stated mean confidence (issue #6: +22.5pp on 39.6% vs 62.1). Field Brier is 0.2669 (was 0.2948) and 0 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 claude-opus-5's -14.6pp (+24.5pp -> +9.9pp) to deepseek-v4-pro's -6.0pp (+18.4pp -> +12.4pp); 7 of 7 model lines narrowed week over week. claude-fable-5 is the best-calibrated line (Brier 0.2593); gemini-3.1-pro leads on hit-rate (52.9%).
- **The high-confidence buckets, N>=10 only.** deepseek-v4-pro 70-80 42.3% (n=26), gemini-3.1-pro 70-80 52.9% (n=342), gpt-5.6-sol 70-80 44.9% (n=267). 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 -\$30.12 (claude-opus-5, best) to -\$79.39 (deepseek-v4-pro, worst), field -\$337.72 against issue #6's -\$1018.93.



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 three windows back. "Was" values are the numbers published in issue #6 for the back-to-back window Sep 7-13.

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 #7 the series has seven 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 three 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 second 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
claude-fable-5	799	60.1%	52.3%	61.4	+9.1	0.2593	48.9%-55.8%
claude-opus-5	711	53.5%	51.6%	61.5	+9.9	0.2603	47.9%-55.3%
qwen-3.8-max	819	61.6%	50.8%	60.0	+9.2	0.2604	47.4%-54.2%
grok-4.6	629	47.3%	49.6%	60.3	+10.7	0.2620	45.7%-53.5%
deepseek-v4-pro	936	70.4%	48.4%	60.8	+12.4	0.2660	45.2%-51.6%
gemini-3.1-pro	883	66.4%	52.9%	66.8	+14.0	0.2701	49.6%-56.2%
gpt-5.6-sol	821	61.7%	49.7%	68.1	+18.4	0.2876	46.3%-53.1%
Field (all models)	5,598	-	50.7%	62.8	+12.1	0.2669	-

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 three 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 second 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	54.3% (186)	51.7% (613)	n/a (0)	n/a (0)
claude-opus-5	53.8% (117)	51.2% (594)	n/a (0)	n/a (0)
deepseek-v4-pro	47.2% (288)	49.2% (622)	42.3% (26)	n/a (0)
gemini-3.1-pro	33.3% (6) *	53.2% (530)	52.9% (342)	40.0% (5) *
gpt-5.6-sol	75.0% (4) *	51.9% (547)	44.9% (267)	33.3% (3) *
grok-4.6	49.1% (216)	50.0% (412)	0.0% (1) *	n/a (0)
qwen-3.8-max	51.3% (349)	50.5% (461)	37.5% (8) *	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 100.0% (n=1) *. Mid-scale ranking: 2 of the 5 models with N>=10 in both 50-60 and 60-70 out-hit 50-60 from 60-70 this week (issue #6: 1 of 6).

Hit-rate by forecast horizon (FH)

Model	1h	4h	1d
claude-fable-5	53.2% (496)	54.5% (253)	32.0% (50)
claude-opus-5	51.2% (422)	56.1% (244)	31.1% (45)
deepseek-v4-pro	48.8% (586)	50.5% (297)	32.1% (53)
gemini-3.1-pro	53.8% (561)	52.9% (272)	42.0% (50)
gpt-5.6-sol	49.7% (505)	52.6% (266)	34.0% (50)
grok-4.6	49.1% (385)	53.2% (205)	35.9% (39)
qwen-3.8-max	51.9% (505)	51.7% (269)	33.3% (45)

n in parentheses; * = N<10 (insufficient). 0 of 7 model lines hit better on 1d than on 1h this window (issue #6: 0 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
claude-opus-5	711	43.5%	-\$30.12	+\$40.98	-\$60.18
claude-fable-5	799	45.2%	-\$31.40	+\$48.50	-\$63.40
gemini-3.1-pro	883	44.4%	-\$34.19	+\$54.11	-\$77.86
grok-4.6	629	42.1%	-\$44.11	+\$18.79	-\$54.55
qwen-3.8-max	819	43.4%	-\$56.30	+\$25.60	-\$63.65
gpt-5.6-sol	821	42.5%	-\$62.22	+\$19.88	-\$92.02
deepseek-v4-pro	936	42.4%	-\$79.39	+\$14.21	-\$80.25

0 of 7 model lines finished net-positive and 7 gross-positive this week (issue #6: 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: an up week on a louder tape

OBSERVATION — mean [1d move] 1.35% -> 2.69% (+100% rel), BTC realized vol 31.4% -> 35.1% (ann., hourly); field directional accuracy 39.6% -> 50.7% (+11.2pp), 7 of 7 models improved, sim win-rate up for 7 of 7, field sim PnL -\$1018.93 -> -\$337.72 (adjacent calendar weeks Sep 7-13 vs Sep 14-20; descriptive, one pair of weeks, not a claim).

Robustness: raw price-sign accuracy 38.7% -> 54.2% — the same direction as the trade-based hit rule.

Measure	Sep 7-13 (week A)	Sep 14-20 (week B)	Change
Mean [1d move] (5 assets)	1.35%	2.69%	+100% rel
BTC realized vol (ann., hourly)	31.4%	35.1%	+3.7pp
Field directional accuracy	39.6%	50.7%	+11.2pp
Models improving hit-rate	--	7 of 7	--
Raw price-sign accuracy	38.7%	54.2%	+15.4pp
Field sim win-rate	32.1%	43.4%	+11.2pp
Field sim net PnL	-\$1018.93	-\$337.72	--

Week A is exactly the issue-#6 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, three windows back, inside issue #4's window). Week A (2026-09-07..2026-09-13) is pure grok-4.6 (it reproduces the published issue #6); 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 +12.1pp this week against +22.5pp in issue #6.
- Gap and regime move together. The field hit 50.7% against 39.6% in issue #6 and stated mean confidence read 62.8 against 62.1, so the gap came out at +12.1pp: a field that hits 50.7% instead of 39.6% 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 five 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 5,598 scored calls sit inside one market regime. This window ran 3 trend days of 7 (issue #6: 2 of 7); every comparison with issue #6 is a comparison across regimes as well as across weeks. Seven 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 7): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Sep 14-20) daily coverage is FULL, as it was in issues #4, #5 and #6: 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 488, 1M 485). (2) The grok line flipped 4.5 -> 4.6 at Aug 24, 2026 09:22 UTC, three 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-#6 window), so every grok number in this issue and in issue #6 is a pure grok-4.6 line -- the grok week-over-week row is the second 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 #6 after the audit that found the raw table mixes two exchanges; issue #6 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 21 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5 and #6, which printed 47,373 at the Sep 14 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 52,971 is an additive step under one definition; counter deltas against issue #5 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,308
OK in gate	9,308	-- of them directional	5,598
Out of gate (1w / 1M)	488 / 485	-- of them sideways	3,710
Invalid	9	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-09-14.json	Report cutoff	Mon Sep 21, 2026, 16:00 UTC
Generated at (pipeline)	2026-09-22 06:56 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 5,598 scored observations. MARKETMANIA RESEARCH TO DATE (as of Sep 21, 2026 cutoff): 52,971 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 32 published reports.

Issue #7. Weekly Calibration is a living series. Issue #6's open questions -- does the gap keep tracking the field base one-for-one or hold a level of its own, and does qwen-3.8-max hold the best-Brier line for a third issue? -- read this issue as: field gap +22.5pp -> +12.1pp (7 of 7 lines narrowed) on a field base that moved +11.2pp, and the best-Brier line changed hands: claude-fable-5 at 0.2593 (issue #6: qwen-3.8-max at 0.2794). Also on the record: gpt-5.6-sol's 70-80 bucket read 44.9% (n=267) against its own 60-70 at 51.9% (n=547). The grok row is pure grok-4.6 this issue and was pure grok-4.6 in issue #6; the 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits three 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 -10.4pp while the field base moved +11.2pp -- the second issue running in which the two move nearly one-for-one in opposite directions; does that hold for a third?
- The best-Brier line changed hands to claude-fable-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 #7	https://marketmania.ai/research/reports/consensus-watch-2026-09-14.pdf
Weekly Model Watch #7	https://marketmania.ai/research/reports/model-watch-2026-09-14.pdf
Weekly Calibration #6	https://marketmania.ai/research/reports/weekly-calibration-2026-09-07.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_2026w38,
  title = {Weekly Calibration #7: confidence vs. direction-hit, Sep 14-20 2026},
  author = {{MarketMania Research}},
  year = {2026}, month = {September}, day = {22},
  url = {https://marketmania.ai/research/reports/weekly-calibration-2026-09-14.pdf},
  note = {Methodology v1.1, hash e66c7e8c864a2233; source weekly_metrics_2026-09-14.json}
}
```