

Weekly Calibration #9

October 7, 2026 · MarketMania Research · Weekly series · window Sep 28-Oct 4, 2026 UTC

RESEARCH SNAPSHOT

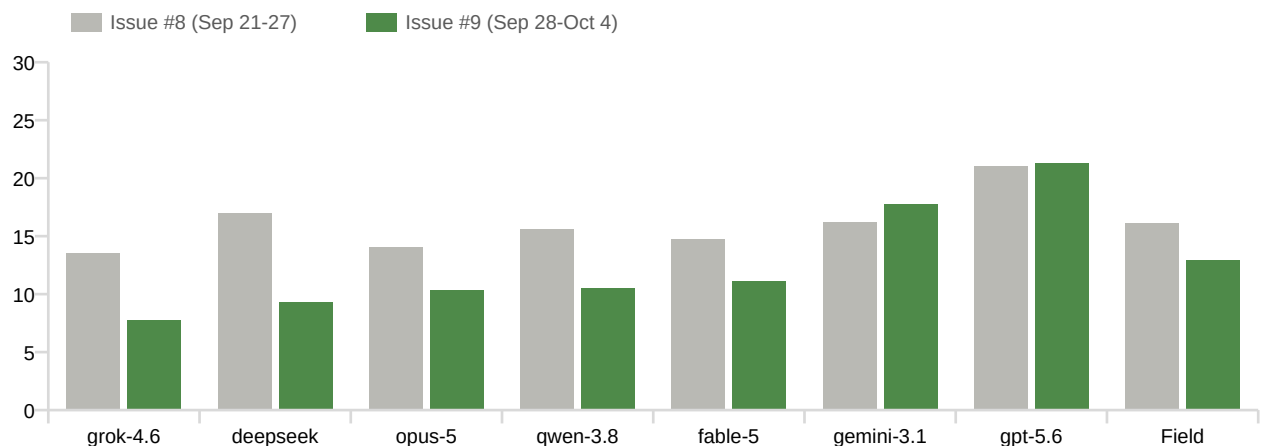
4,331 directional calls scored	7 models (stable lineup)	of 9,286 mature forecasts	window Sep 28-Oct 4, 2026 UTC
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate 1h / 4h / 1d	base field hit 49.6% vs 62.5 stated	cutoff Mon Oct 5, 2026, 16:00 UTC
market: BTC net +2.44% (prior +4.06%)	ann vol 21.1% (was 52.2%)	TOP5 volume \$17.96B , -20% w/w	pairwise corr 0.57 (was 0.71)

KEY FINDING · OBSERVATION (one weekly window)

The field's overconfidence gap moved +16.1pp -> +12.9pp week over week, 5 of 7 model lines narrowed, and field Brier went 0.2751 -> 0.2698.

TL;DR

- **OBSERVATION -- The gap read +12.9pp this week.** Directional calls hit 49.6% against 62.5 stated mean confidence (issue #8: +16.1pp on 46.8% vs 63.0). Field Brier is 0.2698 (was 0.2751) and 0 of 7 model lines scored below 0.25, the score of an uninformative always-50% predictor. One window, descriptive.
- **5 of 7 models narrowed their gap.** Gap moves ran from deepseek-v4-pro's -7.7pp (+17.0pp -> +9.3pp) to gemini-3.1-pro's +1.5pp (+16.2pp -> +17.7pp); 5 of 7 model lines narrowed week over week. grok-4.6 is both the lowest-Brier line (Brier 0.2566) and the hit-rate leader (52.7%).
- **The high-confidence buckets, N>=10 only.** deepseek-v4-pro 70-80 56.2% (n=16), gemini-3.1-pro 70-80 46.4% (n=209), gemini-3.1-pro 80-100 54.5% (n=22), gpt-5.6-sol 70-80 47.2% (n=216). 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 -\$30.94 (grok-4.6, best) to -\$91.49 (gemini-3.1-pro, worst), field -\$408.24 against issue #8's -\$315.68.



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 five windows back. "Was" values are the numbers published in issue #8 for the back-to-back window Sep 21-27.

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 #9 the series has nine 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 five 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 fourth 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
grok-4.6	431	32.6%	52.7%	60.4	+7.7	0.2566	47.9%-57.3%
deepseek-v4-pro	687	51.6%	50.9%	60.3	+9.3	0.2594	47.2%-54.7%
qwen-3.8-max	711	53.5%	48.9%	59.4	+10.5	0.2608	45.3%-52.6%
claude-opus-5	534	40.2%	50.7%	61.1	+10.3	0.2617	46.5%-55.0%
claude-fable-5	623	46.8%	49.9%	61.0	+11.1	0.2630	46.0%-53.8%
gemini-3.1-pro	694	52.2%	49.0%	66.6	+17.7	0.2858	45.3%-52.7%
gpt-5.6-sol	651	49.5%	46.5%	67.9	+21.3	0.2951	42.7%-50.4%
Field (all models)	4,331	-	49.6%	62.5	+12.9	0.2698	-

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 five 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 fourth 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	47.3% (201)	51.3% (421)	0.0% (1) *	n/a (0)
claude-opus-5	53.4% (103)	50.1% (431)	n/a (0)	n/a (0)
deepseek-v4-pro	49.3% (270)	51.9% (401)	56.2% (16)	n/a (0)
gemini-3.1-pro	60.0% (5) *	49.8% (458)	46.4% (209)	54.5% (22)
gpt-5.6-sol	50.0% (6) *	46.2% (429)	47.2% (216)	n/a (0)
grok-4.6	50.7% (138)	53.9% (291)	0.0% (2) *	n/a (0)
qwen-3.8-max	46.4% (345)	50.8% (352)	75.0% (4) *	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 60.0% (n=10). Mid-scale ranking: 4 of the 5 models with N>=10 in both 50-60 and 60-70 out-hit 50-60 from 60-70 this week (issue #8: 4 of 5).

Hit-rate by forecast horizon (FH)

Model	1h	4h	1d
claude-fable-5	50.2% (434)	48.6% (177)	58.3% (12)
claude-opus-5	51.5% (375)	49.3% (150)	44.4% (9) *
deepseek-v4-pro	50.3% (459)	49.5% (192)	66.7% (36)
gemini-3.1-pro	49.7% (477)	47.4% (192)	48.0% (25)
gpt-5.6-sol	46.2% (429)	46.2% (199)	56.5% (23)
grok-4.6	52.8% (301)	52.0% (123)	57.1% (7) *
qwen-3.8-max	48.6% (469)	48.4% (213)	58.6% (29)

n in parentheses; * = N<10 (insufficient). 5 of 7 model lines hit better on 1d than on 1h this window (issue #8: 2 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	431	45.5%	-\$30.94	+\$12.16	-\$55.18
deepseek-v4-pro	687	43.4%	-\$38.00	+\$30.70	-\$63.69
qwen-3.8-max	711	41.6%	-\$49.80	+\$21.30	-\$77.40
claude-opus-5	534	40.6%	-\$56.43	-\$3.03	-\$71.19
claude-fable-5	623	41.6%	-\$60.76	+\$1.54	-\$71.65
gpt-5.6-sol	651	37.6%	-\$80.82	-\$15.72	-\$100.20
gemini-3.1-pro	694	38.3%	-\$91.49	-\$22.09	-\$109.15

0 of 7 model lines finished net-positive and 4 gross-positive this week (issue #8: 0 of 7 net-positive, 6 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 volatility

OBSERVATION — mean |1d move| 1.83% -> 0.87% (-52% rel), BTC realized vol 35.8% -> 32.2% (ann., hourly); field directional accuracy 46.8% -> 49.6% (+2.8pp), 5 of 7 models improved, sim win-rate up for 5 of 7, field sim PnL -\$315.68 -> -\$408.24 (adjacent calendar weeks Sep 21-27 vs Sep 28-Oct 4; descriptive, one pair of weeks, not a claim).

Robustness: raw price-sign accuracy 49.2% -> 49.0% — the opposite direction to the trade-based hit rule.

Measure	Sep 21-27 (week A)	Sep 28-Oct 4 (week B)	Change
Mean 1d move (5 assets)	1.83%	0.87%	-52% rel
BTC realized vol (ann., hourly)	35.8%	32.2%	-3.6pp
Field directional accuracy	46.8%	49.6%	+2.8pp
Models improving hit-rate	--	5 of 7	--
Raw price-sign accuracy	49.2%	49.0%	-0.2pp
Field sim win-rate	40.0%	41.0%	+1.1pp
Field sim net PnL	-\$315.68	-\$408.24	--

Week A is exactly the issue-#8 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, five windows back, inside issue #4's window). Week A (2026-09-21..2026-09-27) is pure grok-4.6 (it reproduces the published issue #8); 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 +12.9pp this week against +16.1pp in issue #8.
- Gap and regime move together. The field hit 49.6% against 46.8% in issue #8 and stated mean confidence read 62.5 against 63.0, so the gap came out at +12.9pp: a field that hits 49.6% instead of 46.8% 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 seven 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 4,331 scored calls sit inside one market regime. This window ran 2 trend days of 7 (issue #8: 2 of 7); every comparison with issue #8 is a comparison across regimes as well as across weeks. Nine 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 9): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Sep 28-Oct 4) daily coverage is FULL, as it was in issues #4, #5, #6, #7 and #8: 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 486). (2) The grok line flipped 4.5 -> 4.6 at Aug 24, 2026 09:22 UTC, five 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-#8 window), so every grok number in this issue and in issue #8 is a pure grok-4.6 line -- the grok week-over-week row is the fourth 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 #8 after the audit that found the raw table mixes two exchanges; issue #8 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 (Oct 5 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6, #7 and #8, which printed 57,929 at the Sep 28 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 62,260 is an additive step under one definition; counter deltas against issue #7 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,286
OK in gate	9,286	-- of them directional	4,331
Out of gate (1w / 1M)	490 / 486	-- of them sideways	4,955
Invalid	28	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-09-28.json	Report cutoff	Mon Oct 5, 2026, 16:00 UTC
Generated at (pipeline)	2026-10-07 06:16 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 4,331 scored observations. MARKETMANIA RESEARCH TO DATE (as of Oct 5, 2026 cutoff): 62,260 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 44 published reports.

Issue #9. Weekly Calibration is a living series. Issue #8's open questions -- "the gap moved +4.0pp while the field base moved -3.9pp -- the third issue running in which the two move nearly one-for-one in opposite directions; does that hold for a fourth?" and "The best-Brier line changed hands again, to qwen-3.8-max -- does it hold for a second issue, or change hands again?" -- read this issue as: field gap +16.1pp -> +12.9pp (5 of 7 lines narrowed) on a field base that moved +2.8pp, a fourth issue running in opposite directions, and the best-Brier line changed hands again: grok-4.6 at 0.2566 (issue #8: qwen-3.8-max at 0.2689). Also on the record: gpt-5.6-sol's 70-80 bucket read 47.2% (n=216) against its own 60-70 at 46.2% (n=429). The grok row is pure grok-4.6 this issue and was pure grok-4.6 in issue #8; the 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits five 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 -3.2pp while the field base moved +2.8pp -- the fourth issue running in which the two move nearly one-for-one in opposite directions; does that hold for a fifth?
- The best-Brier line changed hands again, to grok-4.6 -- 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 #9	https://marketmania.ai/research/reports/consensus-watch-2026-09-28.pdf
Weekly Model Watch #9	https://marketmania.ai/research/reports/model-watch-2026-09-28.pdf
Weekly Calibration #8	https://marketmania.ai/research/reports/weekly-calibration-2026-09-21.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_2026w40,
  title = {Weekly Calibration #9: confidence vs. direction-hit, Sep 28-Oct 4 2026},
  author = {{MarketMania Research}},
  year = {2026}, month = {October}, day = {7},
  url = {https://marketmania.ai/research/reports/weekly-calibration-2026-09-28.pdf},
  note = {Methodology v1.1, hash e66c7e8c864a2233; source weekly_metrics_2026-09-28.json}
}
```