

Weekly Calibration #10

October 14, 2026 · MarketMania Research · Weekly series · window Oct 5-11, 2026 UTC

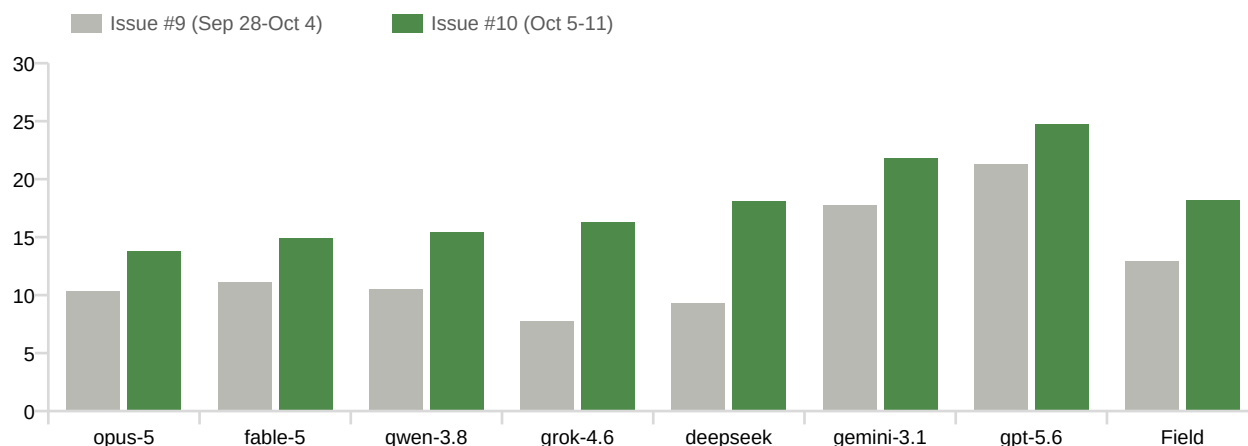
RESEARCH SNAPSHOT			
4,601 directional calls scored	7 models (stable lineup)	of 9,249 mature forecasts	window Oct 5-11, 2026 UTC
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate 1h / 4h / 1d	base field hit 43.8% vs 61.9 stated	cutoff Mon Oct 12, 2026, 16:00 UTC
market: BTC net -3.19% (prior +2.44%)	ann vol 27.2% (was 21.1%)	TOP5 volume \$15.97B , -11% w/w	pairwise corr 0.85 (was 0.57)

KEY FINDING · OBSERVATION (one weekly window)

The field's overconfidence gap moved +12.9pp -> +18.2pp week over week, 0 of 7 model lines narrowed, and field Brier went 0.2698 -> 0.2822.

TL;DR

- **OBSERVATION -- The gap read +18.2pp this week.** Directional calls hit 43.8% against 61.9 stated mean confidence (issue #9: +12.9pp on 49.6% vs 62.5). Field Brier is 0.2822 (was 0.2698) and 0 of 7 model lines scored below 0.25, the score of an uninformative always-50% predictor. One window, descriptive.
- **No model narrowed its gap.** Gap moves ran from gpt-5.6-sol's +3.4pp (+21.3pp -> +24.7pp) to deepseek-v4-pro's +8.8pp (+9.3pp -> +18.1pp); 0 of 7 model lines narrowed week over week. claude-opus-5 is both the lowest-Brier line (Brier 0.2685) and the hit-rate leader (47.0%).
- **The high-confidence buckets, N>=10 only.** deepseek-v4-pro 70-80 41.7% (n=12), gemini-3.1-pro 70-80 40.8% (n=179), gpt-5.6-sol 70-80 49.6% (n=244). 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 -\$40.48 (claude-opus-5, best) to -\$98.53 (deepseek-v4-pro, worst), field -\$520.82 against issue #9's -\$408.24.



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 six windows back. "Was" values are the numbers published in issue #9 for the back-to-back window Sep 28-Oct 4.

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 #10 the series has ten 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 six 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 fifth 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-opus-5	517	39.8%	47.0%	60.8	+13.8	0.2685	42.7%-51.3%
claude-fable-5	613	47.1%	45.0%	59.9	+14.9	0.2714	41.1%-49.0%
grok-4.6	547	41.1%	43.3%	59.7	+16.3	0.2717	39.2%-47.5%
qwen-3.8-max	707	53.2%	43.1%	58.5	+15.4	0.2724	39.5%-46.8%
deepseek-v4-pro	761	57.2%	42.0%	60.2	+18.1	0.2792	38.6%-45.6%
gemini-3.1-pro	722	54.3%	43.4%	65.2	+21.8	0.2979	39.8%-47.0%
gpt-5.6-sol	734	55.2%	43.5%	68.1	+24.7	0.3059	39.9%-47.1%
Field (all models)	4,601	-	43.8%	61.9	+18.2	0.2822	-

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 six 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 fifth 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	45.7% (256)	44.5% (357)	n/a (0)	n/a (0)
claude-opus-5	50.4% (119)	46.0% (398)	n/a (0)	n/a (0)
deepseek-v4-pro	42.9% (296)	41.5% (453)	41.7% (12)	n/a (0)
gemini-3.1-pro	80.0% (10)	43.5% (529)	40.8% (179)	50.0% (4) *
gpt-5.6-sol	54.5% (11)	39.6% (472)	49.6% (244)	71.4% (7) *
grok-4.6	38.9% (239)	46.9% (307)	n/a (0)	n/a (0)
qwen-3.8-max	44.5% (400)	41.4% (285)	0.0% (2) *	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): grok-4.6 0.0% (n=1) *; qwen-3.8-max 45.0% (n=20). Mid-scale ranking: 1 of the 7 models with N>=10 in both 50-60 and 60-70 out-hit 50-60 from 60-70 this week (issue #9: 4 of 5).

Hit-rate by forecast horizon (FH)

Model	1h	4h	1d
claude-fable-5	47.1% (380)	44.1% (202)	25.8% (31)
claude-opus-5	49.7% (314)	45.8% (168)	28.6% (35)
deepseek-v4-pro	46.1% (432)	40.1% (277)	19.2% (52)
gemini-3.1-pro	45.8% (443)	43.0% (237)	19.1% (42)
gpt-5.6-sol	46.7% (433)	42.4% (250)	21.6% (51)
grok-4.6	44.3% (314)	43.9% (198)	31.4% (35)
qwen-3.8-max	46.6% (423)	40.7% (243)	21.9% (41)

n in parentheses; * = N<10 (insufficient). 0 of 7 model lines hit better on 1d than on 1h this window (issue #9: 5 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	517	37.7%	-\$40.48	+\$11.22	-\$53.90
grok-4.6	547	35.3%	-\$52.78	+\$1.92	-\$79.76
claude-fable-5	613	35.2%	-\$64.62	-\$3.32	-\$69.05
qwen-3.8-max	707	33.7%	-\$79.87	-\$9.17	-\$83.77
gpt-5.6-sol	734	33.7%	-\$89.16	-\$15.76	-\$94.48
gemini-3.1-pro	722	33.8%	-\$95.39	-\$23.19	-\$96.94
deepseek-v4-pro	761	33.5%	-\$98.53	-\$22.43	-\$105.03

0 of 7 model lines finished net-positive and 2 gross-positive this week (issue #9: 0 of 7 net-positive, 4 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 with bigger daily moves

OBSERVATION — mean |1d move| 0.87% -> 1.54% (+77% rel), BTC realized vol 32.2% -> 30.9% (ann., hourly); field directional accuracy 49.6% -> 43.8% (-5.9pp), 0 of 7 models improved, sim win-rate up for 0 of 7, field sim PnL -\$408.24 -> -\$520.82 (adjacent calendar weeks Sep 28-Oct 4 vs Oct 5-11; descriptive, one pair of weeks, not a claim).

Robustness: raw price-sign accuracy 49.0% -> 45.5% — the same direction as the trade-based hit rule.

Measure	Sep 28-Oct 4 (week A)	Oct 5-11 (week B)	Change
Mean 1d move (5 assets)	0.87%	1.54%	+77% rel
BTC realized vol (ann., hourly)	32.2%	30.9%	-1.3pp
Field directional accuracy	49.6%	43.8%	-5.9pp
Models improving hit-rate	--	0 of 7	--
Raw price-sign accuracy	49.0%	45.5%	-3.5pp
Field sim win-rate	41.0%	34.5%	-6.5pp
Field sim net PnL	-\$408.24	-\$520.82	--

Week A is exactly the issue-#9 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, six windows back, inside issue #4's window). Week A (2026-09-28..2026-10-04) is pure grok-4.6 (it reproduces the published issue #9); 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 +18.2pp this week against +12.9pp in issue #9.
- Gap and regime move together. The field hit 43.8% against 49.6% in issue #9 and stated mean confidence read 61.9 against 62.5, so the gap came out at +18.2pp: a field that hits 43.8% instead of 49.6% 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 eight 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 4,601 scored calls sit inside one market regime. This window ran 2 trend days of 7 (issue #9: 2 of 7); every comparison with issue #9 is a comparison across regimes as well as across weeks. Ten 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 10): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Oct 5-11) daily coverage is FULL, as it was in issues #4, #5, #6, #7, #8 and #9: 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 479). (2) The grok line flipped 4.5 -> 4.6 at Aug 24, 2026 09:22 UTC, six 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-#9 window), so every grok number in this issue and in issue #9 is a pure grok-4.6 line -- the grok week-over-week row is the fifth 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 #9 after the audit that found the raw table mixes two exchanges; issue #9 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 12 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6, #7, #8 and #9, which printed 62,260 at the Oct 5 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 66,861 is an additive step under one definition; counter deltas against issue #8 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,290	Mature (scored pool)	9,249
OK in gate	9,249	-- of them directional	4,601
Out of gate (1w / 1M)	490 / 479	-- of them sideways	4,648
Invalid	72	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-10-05.json	Report cutoff	Mon Oct 12, 2026, 16:00 UTC
Generated at (pipeline)	2026-10-13 15:15 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 4,601 scored observations. MARKETMANIA RESEARCH TO DATE (as of Oct 12, 2026 cutoff): 66,861 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 52 published reports.

Issue #10. Weekly Calibration is a living series. Issue #9's open questions -- "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?" and "The best-Brier line changed hands again, to grok-4.6 -- does it hold for a second issue, or change hands again?" -- read this issue as: field gap +12.9pp -> +18.2pp (0 of 7 lines narrowed) on a field base that moved -5.9pp, a fifth issue running in opposite directions, and the best-Brier line changed hands again: claude-opus-5 at 0.2685 (issue #9: grok-4.6 at 0.2566). Also on the record: gpt-5.6-sol's 70-80 bucket read 49.6% (n=244) against its own 60-70 at 39.6% (n=472). The grok row is pure grok-4.6 this issue and was pure grok-4.6 in issue #9; the 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits six 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 +5.3pp while the field base moved -5.9pp -- the fifth issue running in which the two move nearly one-for-one in opposite directions; does that hold for a sixth?
- The best-Brier line changed hands again, to claude-opus-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 #10	https://marketmania.ai/research/reports/consensus-watch-2026-10-05.pdf
Weekly Model Watch #10	https://marketmania.ai/research/reports/model-watch-2026-10-05.pdf
Weekly Calibration #9	https://marketmania.ai/research/reports/weekly-calibration-2026-09-28.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_2026w41,
  title = {Weekly Calibration #10: confidence vs. direction-hit, Oct 5-11 2026},
  author = {{MarketMania Research}},
  year  = {2026}, month = {October}, day = {14},
  url   = {https://marketmania.ai/research/reports/weekly-calibration-2026-10-05.pdf},
  note  = {Methodology v1.1, hash e66c7e8c864a2233; source weekly_metrics_2026-10-05.json}
}
```