Weekly Calibration #1

August 10, 2026 · MarketMania Research · Weekly series

RESEARCH SNAPSHOT			
window **Aug 3-9, 2026** (UTC)	**8** models (7 current + qwen-3.7-max legacy to Aug 5)	**4,042** directional calls scored of **9,121** 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**

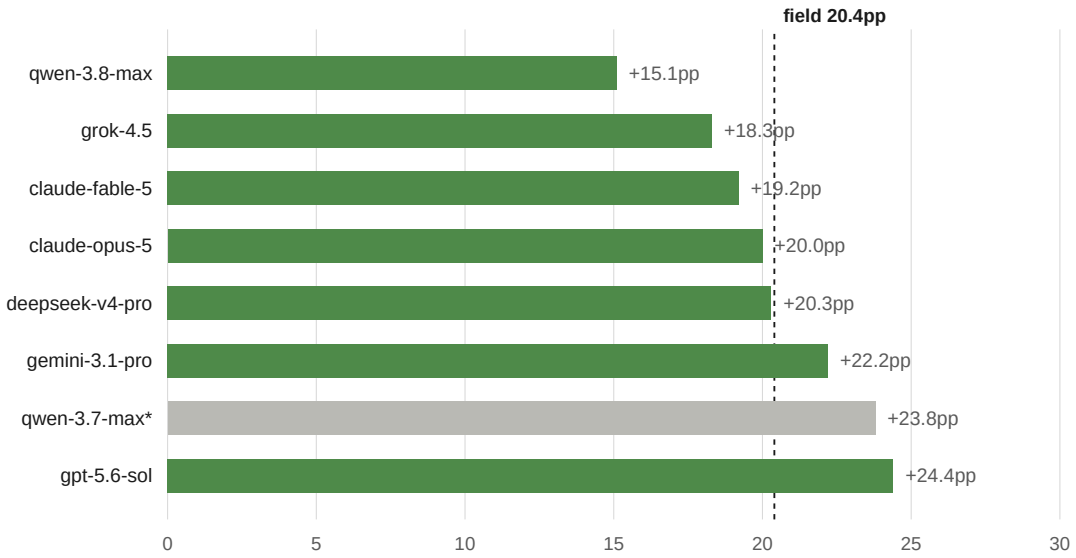
KEY FINDING · OBSERVATION (one weekly window)

Frontier models were systematically overconfident this week: 62.8% stated confidence vs 42.4% realized directional accuracy.

TL;DR

- **OBSERVATION -- A flat, hard week for the whole field.** Directional calls hit **42.4%** of the time against **62.8** stated mean confidence -- a **+20.4pp** overconfidence gap. Every model's Brier score topped 0.25 this week, worse than an uninformative always-50% predictor.
- **qwen-3.8-max is this week's best-calibrated model:** lowest Brier (**0.2713**), smallest gap (**+15.1pp**) and highest hit-rate (**44.1%**) of the 8-model field -- in line with the caution it showed in Stability Index Run 2 (a different metric, different sample; read as context, not confirmation).
- **Confidence mostly failed to rank outcomes this week:** of the 6 models with sufficient N (>=10) in both the 50-60 and 60-70 buckets, 4 did not out-hit 50-60 from 60-70. Higher up, deepseek-v4-pro, gemini-3.1-pro, qwen-3.7-max all hit meaningfully less in 70-80 than 60-70 (inverted); gpt-5.6-sol alone stayed roughly flat.
- **Brier and hit-rate rankings reward different things:** gemini-3.1-pro is 2nd of 8 by hit-rate (43.7%) but only 6th by Brier (0.3038), with the field's third-largest gap (22.2pp). gpt-5.6-sol posts the single largest gap this week, at 24.4pp.

Overconfidence gap by model



Overconfidence gap (mean stated confidence minus hit-rate, percentage points) by model, sorted best (smallest gap) to worst. Dashed line = field average (20.4 pp). Gray bar = legacy model (qwen-3.7-max, replaced Aug 5). Bucket- and horizon-level detail follows on the next pages.

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 for the Aug 3-9 window -- it establishes the baseline for the series. Durability claims (does a model's calibration hold up over time) 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	447	53.9%	44.1%	59.2	+15.1	0.2713	39.5%-48.7%
grok-4.5	607	46.7%	42.2%	60.5	+18.3	0.2811	38.3%-46.1%
claude-fable-5	642	49.2%	41.4%	60.7	+19.2	0.2811	37.7%-45.3%
claude-opus-5	465	35.6%	40.9%	60.8	+20.0	0.2825	36.5%-45.4%
deepseek-v4-pro	377	28.9%	41.6%	62.0	+20.3	0.2915	36.8%-46.7%
gemini-3.1-pro	629	48.3%	43.7%	65.9	+22.2	0.3038	39.9%-47.6%
gpt-5.6-sol	612	46.9%	43.0%	67.4	+24.4	0.3076	39.1%-46.9%
qwen-3.7-max *legacy, to Aug 5*	263	56.0%	42.2%	66.0	+23.8	0.3129	36.4%-48.2%
Field (all models)	**4,042**	-	**42.4%**	**62.8**	**+20.4**	**0.2908**	-

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 (see Limitations -- in-window observations are dependent). Field row has no single-model coverage or CI.

Calibration curve: hit-rate by stated-confidence bucket

Model	50-60	60-70	70-80	80-100
qwen-3.8-max	48.4% (223)	40.3% (221)	*n/a (0)*	*n/a (0)*
grok-4.5	44.0% (323)	40.0% (280)	*50.0% (4) **	*n/a (0)*
claude-fable-5	42.8% (208)	40.8% (434)	*n/a (0)*	*n/a (0)*
claude-opus-5	39.5% (114)	41.3% (351)	*n/a (0)*	*n/a (0)*
deepseek-v4-pro	36.6% (93)	45.0% (218)	35.9% (53)	*33.3% (3) **
gemini-3.1-pro	*50.0% (6) **	47.3% (421)	36.3% (201)	*0.0% (1) **
gpt-5.6-sol	*33.3% (6) **	43.2% (444)	42.2% (161)	*100.0% (1) **
qwen-3.7-max*	66.7% (12)	45.3% (159)	34.2% (79)	*25.0% (4) **

Sub-50 bucket (n>0 only, footnote): qwen-3.8-max: 0.0% (n=3) *; deepseek-v4-pro: 50.0% (n=10); qwen-3.7-max: 33.3% (n=9) *. All other models had zero sub-50 calls this week.
n in parentheses. * = N<10 (insufficient, greyed) -- no conclusions drawn from these cells. "n/a (0)" = no calls landed in that bucket this week. * after a model name marks the legacy row (qwen-3.7-max, to Aug 5).

Hit-rate by forecast horizon (FH)

Model	1h	4h	1d
qwen-3.8-max	43.4% (279)	47.0% (149)	*31.6% (19) #*
grok-4.5	42.1% (404)	42.8% (180)	39.1% (23)
claude-fable-5	42.3% (397)	39.4% (216)	44.8% (29)
claude-opus-5	41.1% (275)	39.4% (165)	48.0% (25)
deepseek-v4-pro	43.2% (250)	41.5% (118)	*0.0% (9) **
gemini-3.1-pro	45.9% (405)	40.4% (198)	34.6% (26)
gpt-5.6-sol	44.6% (386)	40.3% (201)	40.0% (25)
qwen-3.7-max*	44.9% (165)	41.2% (85)	*15.4% (13) #*

n in parentheses. * after value = N<10 (insufficient); # after value = N<20, thin but above the N>=10 floor (low, read cautiously). deepseek-v4-pro's 1d cell is insufficient (n=9); qwen-3.7-max's 1d cell is low (n=13). * after a model name marks the legacy row.

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
qwen-3.8-max	447	29.8%	-$43.32	+$1.38	-$43.63
qwen-3.7-max*	263	27.4%	-$50.32	-$24.02	-$56.40
deepseek-v4-pro	377	26.3%	-$53.93	-$16.23	-$57.41
claude-opus-5	465	28.0%	-$61.79	-$15.29	-$61.79
gpt-5.6-sol	612	28.8%	-$78.59	-$17.39	-$80.26
gemini-3.1-pro	629	27.3%	-$78.71	-$15.81	-$79.73
claude-fable-5	642	26.6%	-$87.15	-$22.95	-$87.15
grok-4.5	607	27.2%	-$89.08	-$28.38	-$91.01

All eight models finished net-negative this week. qwen-3.8-max was the only model with a positive gross PnL (**+$1.38**) before fees and slippage; net of those, it closed at **-$43.32**. The other seven models were negative on both gross and net PnL.

WR = trade win rate. PnL and max drawdown (Max DD) in USD. Ranked best-to-worst net PnL. * marks the legacy model. Per methodology, this table is never combined with the calibration/prediction tables above.

Practical implications

- Treat stated confidence as a style signal this week, not a probability -- the 60-70 bucket did not reliably out-hit 50-60, and three models inverted between 60-70 and 70-80.
- Ignore the sub-50 and N<10 cells entirely this issue. There are too few calls in most high-confidence buckets (70-80 / 80-100) and in a couple of 1d horizons to support any conclusion from them.
- Watch, don't conclude, on the inverted 70-80 bucket for deepseek-v4-pro, gemini-3.1-pro and qwen-3.7-max. One week of data cannot separate a real pattern from noise -- this is a flag for next issue, not a claim.

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 4,042 scored calls this week sit inside one market regime (Aug 3-9). A single week cannot separate a persistent calibration pattern from this week's specific conditions.

- This is issue #1 of Weekly Calibration -- there is no prior issue to compare against, so no week-over-week deltas are reported. They begin with issue #2.
- 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

Metric	Value
Forecasts total	9,194
OK in gate	9,121
Invalid	3
Out of gate (1w)	70
Mature (scored pool)	9,121
- directional	4,042
- sideways	5,079
Pending (next issue)	0
Uptime, 1h slots (tf=1h)	165 / 168
Uptime, 4h slots (tf=4h)	41 / 42
Uptime, 4h slots (tf=1h)	41 / 42
Uptime, 1d slots (tf=1d)	7 / 7
Uptime, 1d slots (tf=4h)	7 / 7
Source file	weekly_metrics_2026-08-03.json
Generated at (metrics pipeline)	2026-08-11T14:24:23.984020+00:00
Methodology	v1.1 (2026-08-10), hash e66c7e8c864a2233
Report cutoff	2026-08-10 16:00 UTC (frozen)

THIS REPORT	4,042 scored observations
MARKETMANIA RESEARCH TO DATE (as of Aug 10, 2026 cutoff)	14,151 resolved forecasts since Jul 11, 2026 · 9 models tracked (7 current frontier + 2 archived legacy generations) · 5 assets · 5 forecast horizons · hourly cadence · 4 published research reports

Research-to-date platform counters sourced from platform_counters_2026-08-10 (owner-verified, supplied directly -- not derived from weekly_metrics_2026-08-03.json).

Snapshot principle. Every figure in this report is frozen at the Monday 16:00 UTC cutoff for the Aug 3-9 window. Any later data correction is handled as a note in a future issue, not by silently editing this one.

What we're testing next

- Next issue: does the confidence-bucket inversion persist? Does the 60-70 vs 50-60 non-ranking repeat?
- Monthly test: is the overconfidence gap stable across market regimes (trend vs flat)?

Related research

Stability Index Run 2 — marketmania.ai/research/reports/si-run-2.pdf
Consensus Watch #1 — marketmania.ai/research/reports/consensus-watch-2026-08-03.pdf
Weekly Model Watch #1 — marketmania.ai/research/reports/model-watch-2026-08-03.pdf

These publish together as one wave; links are stable.

CITE THIS REPORT

```
@misc{mm_weekly_calibration_2026w32,
  title  = {Weekly Calibration #1: confidence vs. direction-hit, Aug 3-9 2026},
  author = {{MarketMania Research}},
  year   = {2026}, month = {August}, day = {10},
  url    = {https://marketmania.ai/research},
  note   = {Methodology v1.1, hash e66c7e8c864a2233; source weekly_metrics_2026-08-03.json}
}
```