Weekly Model Watch #1

August 10, 2026 · MarketMania Research · Weekly series

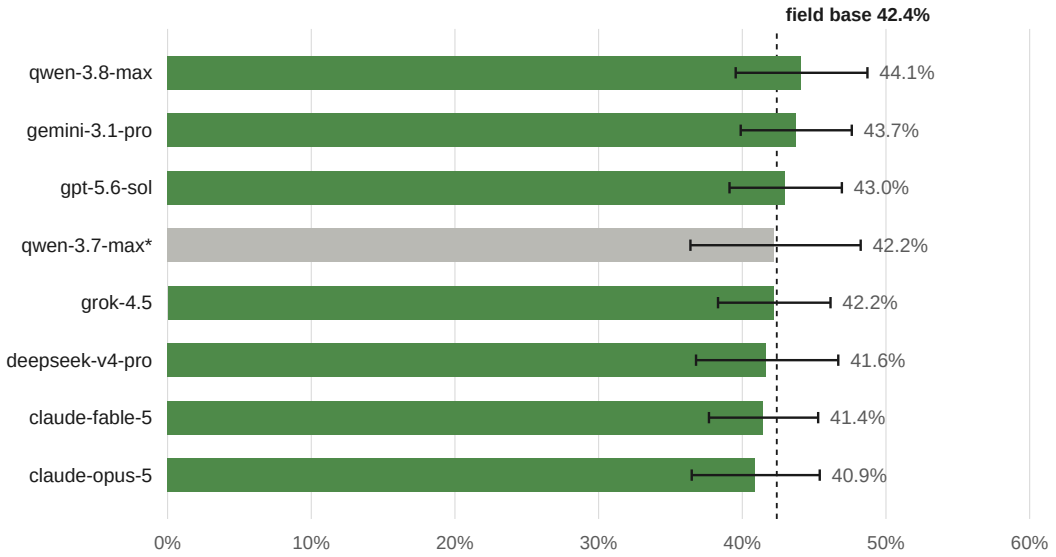
RESEARCH SNAPSHOT			
window **Aug 3-9, 2026 UTC**	**8** models (7 current + qwen-3.7-max legacy to Aug 5)	**4,042** directional calls scored of **9,121** mature	FH gate **1h / 4h / 1d**
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	base field hit **42.4%**	cutoff **Mon Aug 10, 2026, 16:00 UTC**	methodology **v1.1 (2026-08-10)**, hash **e66c7e8c864a2233**

KEY FINDING · OBSERVATION (one weekly window)

The week's only trend reversal went uncalled: zero of 8 models caught BTC's +2.1% Monday turn.

TL;DR

- OBSERVATION -- **No weekly title awarded.** The top four models sit within ~1.9pp of hit-rate (qwen-3.8-max 44.1%, gemini-3.1-pro 43.7%, gpt-5.6-sol 43.0%, qwen-3.7-max 42.2%) with overlapping 95% CIs -- the title rule (>=5pp gap AND non-overlapping 95% CI) was not met. qwen-3.8-max leads descriptively.
- **Ticker of the week: SOL; hardest: ETH.** SOL was the field's most readable coin (48.2%, n=852) and ETH the hardest (34.8%, n=724) -- the widest gap in the table, and the one split where 95% CIs land clean of each other (44.9-51.6% vs. 31.4-38.4%); every adjacent pair in the ranking overlaps its neighbor. Pair of the week: qwen-3.8-max x SOL 52.9% (n=104) -- the same model also owns the worst pair (x ETH, 28.4%, n=88).
- **Self-agreement added nothing this week.** When a model's two timeframes (or neighboring horizons) agreed on direction, hit-rates ran 1.9-4.5pp BELOW the weekly base; the rare intra-model disagreements did far better (71.4% / 70.6%) but on just 7 and 17 calls -- descriptive only, not a strategy.
- **A one-regime week.** All 7 days were flat (max |BTC daily move| 0.98%, methodology flat/trend threshold 1.5%); the field skewed long over short (average of each model's own long:short ratio, ~2.2:1) in a sideways-heavy tape (55.7% of mature calls); its single most confident miss was a conf-85 BTC long (qwen-3.7-max, Aug 4).



Weekly leaderboard: directional hit-rate by model, ranked descending, with 95% Wilson CI whiskers. Dashed line = field base (42.4%). Grey bar = legacy model (qwen-3.7-max, replaced Aug 5). No weekly title is awarded this issue -- see Block 1 for the full table and rule.

Why it matters

MarketMania scores 8 models against the same market, hour after hour. This first issue of Weekly Model Watch asks four practical questions about that week of calls: Who led the week, which tickers were readable, who saw the turn first, and does a model agreeing with itself across timeframes/horizons make its call stronger? Each question gets its own block below, with its own table, its own N, and its own honesty caveats -- this is the fullest report in the weekly series (8 analysis blocks); later issues compress as the format settles.

Leaderboard of the week

The full 8-model field, ranked by this week's directional hit-rate.

Model	N	Coverage	Hit rate	95% CI
qwen-3.8-max	447	53.9%	44.1%	39.5%-48.7%
gemini-3.1-pro	629	48.3%	43.7%	39.9%-47.6%
gpt-5.6-sol	612	46.9%	43.0%	39.1%-46.9%
qwen-3.7-max *legacy, to Aug 5*	263	56.0%	42.2%	36.4%-48.2%
grok-4.5	607	46.7%	42.2%	38.3%-46.1%
deepseek-v4-pro	377	28.9%	41.6%	36.8%-46.7%
claude-fable-5	642	49.2%	41.4%	37.7%-45.3%
claude-opus-5	465	35.6%	40.9%	36.5%-45.4%

No weekly title is awarded this issue. Title rule: **>=5pp gap AND non-overlapping 95% CI, N>=10** (N>=10 per side). The top four sit within ~1.9pp of hit-rate (44.1% down to 42.2%) and every adjacent pair of 95% CIs overlaps, so neither condition is met -- qwen-3.8-max leads descriptively at 44.1% (n=447), not by title. Movement vs. prior week: first issue -- starts next week (issue #2). Hit-rate rank and Brier (calibration) rank diverge sharply for several models this week (e.g. claude-fable-5 ranks 7th of 8 by hit-rate but 2nd by Brier) -- see the calibration bridge in Counters & lineage, and the full picture in Weekly Calibration #1.

Ticker of the week

This week's 5-ticker field, ranked by directional hit-rate.

Symbol	N	Hit rate	95% CI
SOL	852	48.2%	44.9%-51.6%
XRP	800	46.5%	43.1%-50.0%
BNB	782	43.6%	40.2%-47.1%
BTC	884	38.4%	35.2%-41.6%
ETH	724	34.8%	31.4%-38.4%

SOL was the field's most readable ticker this week (48.2%, n=852) and ETH the hardest (34.8%, n=724). Their 95% CIs are well separated (44.9-51.6% vs. 31.4-38.4%) -- the widest gap in the table -- while every adjacent pair in the ranked list (SOL-XRP, XRP-BNB, BNB-BTC, BTC-ETH) overlaps its neighbor. Read this as a one-week readability signal, not a durable per-asset skill claim.

Pair of the week

Top 6 of the 12 model x ticker pairs tracked this week (best hit-rate), and the 5 worst. Pairs shown require N>=8 calls.

Top 6 pairs

Model	Symbol	N	Hit rate
qwen-3.8-max	SOL	104	52.9%
qwen-3.8-max	XRP	96	52.1%
deepseek-v4-pro	SOL	80	50.0%
gpt-5.6-sol	XRP	119	49.6%
claude-opus-5	SOL	110	49.1%
gemini-3.1-pro	XRP	141	48.9%

5 worst pairs

Model	Symbol	N	Hit rate
gpt-5.6-sol	ETH	113	34.5%
gemini-3.1-pro	ETH	108	34.3%
deepseek-v4-pro	BTC	81	33.3%
claude-opus-5	ETH	73	32.9%
qwen-3.8-max	ETH	88	28.4%

qwen-3.8-max owns both extremes this week: its best individual pairing, x SOL (52.9%, n=104), is also the week's best pair overall, and its worst pairing, x ETH (28.4%, n=88), is also the week's worst. Pair history accumulates across issues -- read this week's top/bottom-6 as a first data point, not a ranking.

Who saw the reversal first

Reversal definition: 4h grid; trend = sign of prior 24h; counter-move >=2.0% (BTC/ETH) sustained 8h. A model 'calls' the reversal if it has a matured directional hit call in the new direction, FH 4h or 1d, in the slot window [T-12h, T+FH].

Symbol	Time (UTC)	New side	Move (8h)	Callers	First caller
BTC	Mon Aug 3, 08:00	Long	+2.10%	0	**NOBODY**

This week's only qualifying reversal was BTC's turn to long at 08:00 UTC on Monday, Aug 3 -- an 8-hour move of +2.10%. Zero of the 8 active models had a matured hit call in the new (long) direction inside the [T-12h, T+FH] window; first caller: **NOBODY**. This is a single event -- the reversal detector's first live week -- so it is reported descriptively, not as a claim about any model's turn-detection skill. Scope in v1 is BTC/ETH only; expansion to more assets is planned after 2-3 issues.

Cross-TF confirmation: does agreeing with yourself help?

Within one forecast horizon (FH), does a model's call from a longer input timeframe (TF) agree with its call from a shorter one -- and if so, does the call do any better?

FH	vs.	Agree n	Agree hit	Dis n	Dis hit	Mixed	Base hit	Lift pp
4h	TF 4h vs 1h	441	39.7% [35.2%-44.3%]	7	*71.4% [35.9%-91.8%] **	208	41.7% (n=665)	-2.0
1d	TF 1d vs 4h	53	34.0% [22.7%-47.4%]	0	*n/a (0)*	25	38.5% (n=78)	-4.5

Agree hit / Dis hit = hit-rate with the 95% Wilson CI in brackets; Agree n / Dis n give the call counts. * = N<10 (insufficient, greyed); # = N<20 (thin, greyed); "n/a (0)" = no disagreeing calls this week. Insufficient and thin cells are never used to rank models.

Regime control (degenerate this issue). Methodology calls for a trend-vs-flat split on this table (by_regime). All 7 days this week were classified flat: 08-03 +0.09% · 08-04 +0.93% · 08-05 +0.98% · 08-06 -0.57% · 08-07 +0.88% · 08-08 +0.16% · 08-09 +0.09% (max |move| 0.98%). With zero trend days, the split is degenerate this issue: the flat bucket shown above IS the whole sample, and the trend bucket is n=0 throughout. We report the numbers as-is rather than fabricate a split; this becomes informative in a mixed (trend + flat) week.

Cross-FH confirmation: does a shorter horizon confirm a longer one?

Does a model's call at one forecast horizon (FH) agree with its call at a shorter FH on the same timeframe -- and if so, does the call do any better?

FH	vs.	Agree n	Agree hit	Dis n	Dis hit	Mixed	Base hit	Lift pp
4h	by FH 1h	410	39.8% [35.1%-44.6%]	17	*70.6% [46.9%-86.7%] #*	238	41.7% (n=665)	-1.9
1d	by FH 4h	44	36.4% [23.8%-51.1%]	1	*0.0% [0.0%-79.4%] **	33	38.5% (n=78)	-2.1

The 1d cells are thin across both this block's fh_1d_confirmed_by_4h (44 agree / 1 disagree) and Block 5's fh_1d_tf_1d_vs_4h (53 agree / 0 disagree) -- 44-53 agree, 0-1 disagree. Disagree cells with N<10 (both 1d disagree cells here and in Block 5, plus Block 5's 4h TF-pair disagree cell, n=7) are marked insufficient and are never used to rank models. The 4h FH-confirmation disagree cell below (n=17) clears the N>=10 floor but is still thin -- read descriptively only.

Side-mix and the week's failure

How each model split its calls this week across long / short / sideways, and the week's single costliest miss.

Model	N	Long	Short	Sideways	L:S
claude-fable-5	1,305	37.9%	11.3%	50.8%	3.34
claude-opus-5	1,305	28.1%	7.5%	64.4%	3.74
deepseek-v4-pro	1,305	16.9%	12.0%	71.1%	1.40
gemini-3.1-pro	1,302	31.7%	16.6%	51.7%	1.91
gpt-5.6-sol	1,305	30.5%	16.4%	53.1%	1.86
grok-4.5	1,300	26.9%	19.9%	53.3%	1.35
qwen-3.7-max *legacy, to Aug 5*	470	29.6%	26.4%	44.0%	1.12
qwen-3.8-max	829	39.3%	14.6%	46.1%	2.69

L:S = each model's own long-share divided by its short-share. Averaging that ratio across the 8 models (not pooling raw counts) gives the field's ~2.2:1 long-over-short skew quoted in the TL;DR; the field also called 'sideways' more than half the time overall (55.7% of mature forecasts, 5,079 of 9,121).

Consensus skew. 471 symbol/FH/TF cells this week had 5 or more of the (up to 7) active models on the same side. Those skewed cells' mature calls hit 41.7% (n=2,901) -- below the week's 42.4% field base rate, so leaning with a 5-of-7+ skew did not outperform the field this week either. (A stricter 6-of-6+ unanimity threshold is Consensus Watch's own metric -- see that report for the full herding analysis.)

Model	Symbol	Side	Conf	FH	Slot (UTC)	Exit	Net PnL
qwen-3.7-max	BTC	Long	85	1h	Aug 4, 19:01	Expiry	**-$0.28**

Fail of the week. The week's single highest-confidence individual miss: qwen-3.7-max, BTC, long, confidence 85, FH 1h, slot Aug 4 19:01 UTC, exit by expiry, net -$0.28. Cross-reference: that same slot was a 7-of-7 unanimous long across the active field that day (see Consensus Watch #1) -- the whole field, not just this one model, was wrong-footed.

Counters and the calibration bridge

Gate totals, uptime and data-quality for this window, plus a bridge to this week's calibration read.

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
Late closes	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)	2026-08-11 14:24 UTC
Methodology	v1.1 (2026-08-10), hash e66c7e8c864a2233
Report cutoff	2026-08-10 16:00 UTC (frozen)

Data quality. Per-model ok-rate (share of a model's forecasts that passed gate validation) was at or near 100% for every model this week: gemini-3.1-pro 99.85%, qwen-3.8-max 99.88%, and 100.0% for the other six (claude-fable-5, claude-opus-5, deepseek-v4-pro, gpt-5.6-sol, grok-4.5, qwen-3.7-max).

Calibration bridge. qwen-3.8-max was also this week's best-calibrated model by Brier score (0.2713, lowest of the 8-model field) -- the same model leading this issue's hit-rate leaderboard (Block 1), though for most other models Brier rank and hit-rate rank point in different directions (e.g. claude-fable-5 ranks 7th of 8 by hit-rate but 2nd by Brier). Full confidence-bucket and by-horizon calibration analysis lives in Weekly Calibration #1.

Practical implications

- Do not read this week's leaderboard as a skill ranking. The top four models overlap in both point estimate (within ~1.9pp) and 95% CI, and the title rule was not met -- qwen-3.8-max's lead is descriptive, for this window only.
- Treat SOL as this week's most-readable ticker and ETH as the hardest with real caution: it is one week, and the field also missed its only reversal outright (0 of 8 models), which points to a general weak spot at turns rather than an ETH-specific one.
- Do not use within-model cross-TF/cross-FH agreement as a confidence booster yet. In every tested cell, agreement underperformed the base rate this week (by 1.9-4.5pp); the small number of disagreements did better but on too few calls (7, 17) to act on.

Limitations

- Single week (Aug 3-9, 2026 UTC). Every finding in this issue is descriptive for this window only; no claim is made about next week or about any model's underlying skill.
- This week sat inside a single market regime: all 7 days were classified flat (max |BTC daily move| 0.98%). The trend-vs-flat regime control called for by methodology (Block 5) could not be exercised this issue -- there were no trend days to compare against.

- Observations inside one window are not independent: the same models watch overlapping symbol/FH/TF cells hour after hour, so agree/disagree calls made close in time are correlated, not i.i.d. draws.
- 95% Wilson CIs shown throughout are descriptive, not inferential, for the reason above -- read them as a range, not a formal coverage guarantee.
- The reversal detector (Block 4) is v1 and BTC/ETH-only. Its underlying price series is reconstructed from trade entry prices (median per symbol-slot), not an independent tick feed, per this week's metrics notes. One event this week is not enough to characterize detector performance either way.
- Cross-TF and cross-FH disagreement cells are tiny this week (n=0, 1, 7, 17) -- too small to rank models on; cells below N=10 are marked insufficient and are never used to rank models.
- This is issue #1 of Weekly Model Watch -- there is no prior issue to compare against, so no week-over-week deltas are reported (Block 1's 'movement vs prior week' begins with issue #2).

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

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 reversal blind spot repeat, and does cross-TF agreement stay flat-to-negative?
- **Monthly test:** cross-TF/FH lifts with regime control across mixed weeks (trend vs. flat), block bootstrap.

Related research

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

These three weekly reports -- Consensus Watch, Weekly Calibration and Weekly Model Watch -- publish together as one wave each week; links are stable.

CITE THIS REPORT

```
@misc{mm_model_watch_2026w32,
  title  = {Weekly Model Watch #1: weekly leaderboard, ticker/pair reads and cross-confirmation, Aug 3-9 2026},
  author = {{MarketMania Research}},
  year   = {2026}, month = {August}, day = {10},
  url    = {https://marketmania.ai/research},
  note   = {Methodology v1.1 (2026-08-10), hash e66c7e8c864a2233; window Aug 3-9, 2026 UTC; source weekly_metrics_2026-08-03.json}
}
```