Weekly Model Watch #11

October 21, 2026 · MarketMania Research · Weekly series · window Oct 12-18, 2026 UTC

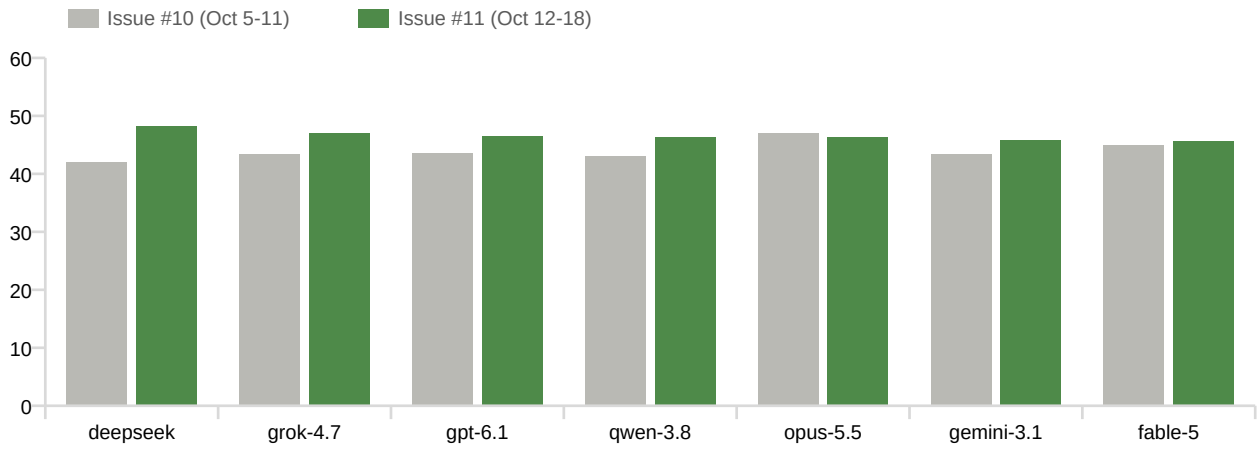
RESEARCH SNAPSHOT			
5,084 directional calls scored	**7** models (3 lines spliced)	of **9,198** mature forecasts	window **Oct 12-18, 2026 UTC**
hit = direction rule: tp1/tp2->hit, sl->miss, expiry->sign of gross pnl	FH gate **1h / 4h / 1d**	base field hit **46.6%**	cutoff **Mon Oct 19, 2026, 16:00 UTC**
market: BTC net **-1.98%** (prior -3.19%)	ann vol **27.4%** (was 27.2%)	TOP5 volume **$15.84B**, -1% w/w	pairwise corr **0.94** (was 0.85)

KEY FINDING · OBSERVATION (one weekly window)

Field accuracy went 43.8% -> 46.6% (+2.8pp) with 6 of 7 models improving, and no weekly title was awarded, for the eleventh issue running.

TL;DR

- **OBSERVATION -- No weekly title again.** deepseek-v4-pro leads descriptively at 48.2% [44.7%, 51.8%]; the gap to #2 (grok-4.7 (spliced), 47.0%) is 1.2pp. 6 of 7 models improved week over week; issue #10's leader claude-opus-5 is now 5 of 7 as claude-opus-5.5 (spliced).
- **No ticker cleared 51%.** XRP is the week's most readable coin (49.0%, n=1,008) and BTC the hardest at 44.0%; 0 of the 5 assets cleared 51%. Pair of the week: deepseek-v4-pro x XRP 52.4% (n=164); worst: claude-opus-5.5 (spliced) x BTC 39.5% (n=114).
- **2 of 6 qualifying turns found a caller.** BTC long +2.40%, 0 caller(s), first NOBODY; ETH short -3.33%, 7 caller(s), first deepseek-v4-pro; ETH long +3.24%, 0 caller(s), first NOBODY; ETH long +2.60%, 0 caller(s), first NOBODY; ETH short -2.56%, 13 caller(s), first gpt-6.1-sol (spliced); ETH long +3.52%, 0 caller(s), first NOBODY.
- **Self-agreement read as a penalty in every cell.** Cross-TF/FH lifts ran -6.2pp to -2.2pp (issue #10: -2.0pp to +6.8pp). On trend days, 4h TF-agreement hit 43.2% (n=227) against 36.9% (n=352) on flat days.



Directional hit-rate (%) by model, week over week, ranked by this week's rate. Short names: opus-5.5 = claude-opus-5.5, fable-5 = claude-fable-5, gemini-3.1 = gemini-3.1-pro, gpt-6.1 = gpt-6.1-sol, qwen-3.8 = qwen-3.8-max, deepseek = deepseek-v4-pro. Field base 43.8% -> 46.6%. The opus-5.5, gpt-6.1 and grok-4.7 columns compare the spliced line against issue #10's pure claude-opus-5 / gpt-5.6-sol / grok-4.6. "Was" values are the numbers published in issue #10; the leaderboard's W/w column is computed on unrounded rates and can differ by 0.1pp from the difference of the rounded columns.

Why it matters

MarketMania scores 7 model lines against the same market, hour after hour. 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. The claude-opus-5.5, gpt-6.1-sol and grok-4.7 rows are lineage splices in this issue -- the new builds took over at Oct 17, 2026 09:01 UTC, inside the window -- so read those three week-over-week rows as lines, not single model builds.

How to read this

Every block below is scored on the same pool: matured directional calls inside the 1h/4h/1d FH gate, hit by the methodology v1.1 direction rule. The weekly title is awarded only on a >=5pp gap AND non-overlapping 95% Wilson CIs with N>=10 per side -- a descriptive lead is not a title. Reversals use the 4h grid with a >=2.0% counter-move sustained 8h, BTC/ETH only in v1. Cross-TF/FH blocks ask whether a model agreeing with itself does any better, and are split trend vs flat wherever the week supplies both sides. Three model lines flipped at Oct 17, 2026 09:01 UTC, INSIDE this window: gpt-5.6-sol -> gpt-6.1-sol, claude-opus-5 -> claude-opus-5.5 and grok-4.6 -> grok-4.7. Every row of those three lines in this issue is a lineage splice (old-era rows Oct 12 00:00-Oct 17 09:01, new-era rows after), week A (the issue-#10 window) is entirely old-era, so each of the three week-over-week rows joins a spliced line to a single-build line. Engine ids are kept; the vendors' names are Claude Opus 5.5, GPT-6.1 Sol and Grok 4.7. claude-opus-5.5 (spliced) = claude-opus-5.5 (incl. opus-5-era, Oct 12 00:00-Oct 17 09:01 UTC); gpt-6.1-sol (spliced) = gpt-6.1-sol (incl. 5.6-era, Oct 12 00:00-Oct 17 09:01 UTC); grok-4.7 (spliced) = grok-4.7 (incl. 4.6-era, Oct 12 00:00-Oct 17 09:01 UTC).

Leaderboard of the week

Model	N	Cov.	Hit rate	95% CI	W/w pp
deepseek-v4-pro	767	59.0%	48.2%	44.7%-51.8%	+6.2
grok-4.7 (spliced)	695	53.0%	47.0%	43.4%-50.8%	+3.7
gpt-6.1-sol (spliced)	815	62.2%	46.5%	43.1%-49.9%	+3.0
qwen-3.8-max	766	57.6%	46.3%	42.8%-49.9%	+3.2
claude-opus-5.5 (spliced)	572	43.7%	46.3%	42.3%-50.4%	-0.7
gemini-3.1-pro	765	58.5%	45.8%	42.2%-49.3%	+2.4
claude-fable-5	704	52.9%	45.6%	41.9%-49.3%	+0.6

No weekly title is awarded this issue -- for the eleventh time. Title rule: **>=5pp gap AND non-overlapping 95% CI, N>=10** per side. deepseek-v4-pro's lead over grok-4.7 (spliced) is 1.2pp and their CIs overlap, so neither condition is met. The claude-opus-5.5, gpt-6.1-sol and grok-4.7 rows are lineage splices (the three lines flipped at Oct 17, 2026 09:01 UTC, inside this window; week A is entirely old-era) -- see How to read this.

Ticker of the week

Symbol	N	Hit rate	95% CI	Issue #10
XRP	1,008	49.0%	45.9%-52.1%	42.1%
SOL	1,072	47.4%	44.4%-50.4%	44.2%
ETH	1,049	47.4%	44.4%-50.4%	43.9%
BNB	990	44.8%	41.7%-47.9%	42.4%
BTC	965	44.0%	40.9%-47.2%	46.1%

XRP was the field's most readable ticker (49.0%) and BTC the hardest (44.0%); 0 of 5 assets cleared 51% this week (issue #10: 0 of 5). Issue #10's best ticker was BTC and its hardest XRP -- this issue's ordering is XRP > SOL > ETH > BNB > BTC.

Pair of the week

Top 6 of the 12 model x ticker pairs tracked this week (best hit-rate), then the 5 worst in their own table -- the pair list is split rather than paged so neither table breaks across a page. Pairs shown require N of 8 or more calls.

Model (top 6)	Symbol	N	Hit rate
deepseek-v4-pro	XRP	164	52.4%
claude-opus-5.5 (spliced)	SOL	123	52.0%
deepseek-v4-pro	ETH	165	51.5%
grok-4.7 (spliced)	XRP	144	50.7%
claude-opus-5.5 (spliced)	XRP	100	50.0%
qwen-3.8-max	XRP	165	48.5%

Model (5 worst)	Symbol	N	Hit rate
claude-opus-5.5 (spliced)	BNB	109	44.0%
claude-fable-5	BTC	143	43.4%
gemini-3.1-pro	BTC	144	43.1%
claude-fable-5	BNB	134	42.5%
claude-opus-5.5 (spliced)	BTC	114	39.5%

deepseek-v4-pro x XRP (52.4%, n=164) is the week's best pair and claude-opus-5.5 (spliced) x BTC (39.5%, n=114) the weakest. Pair history accumulates across issues -- read this as the eleventh data point, not a ranking.

Who saw the reversal first

Reversal definition: 4h grid; trend = sign of prior 24h; counter-move of 2.0% or more (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	Tue Oct 13, 00:00	Long	+2.40%	0	NOBODY
ETH	Mon Oct 12, 16:00	Short	-3.33%	7	deepseek-v4-pro
ETH	Tue Oct 13, 00:00	Long	+3.24%	0	NOBODY
ETH	Thu Oct 15, 08:00	Long	+2.60%	0	NOBODY
ETH	Fri Oct 16, 12:00	Short	-2.56%	13	gpt-6.1-sol (spliced)
ETH	Sun Oct 18, 20:00	Long	+3.52%	0	NOBODY

The BTC long turn drew NOBODY; the ETH short turn drew 7 caller(s), first deepseek-v4-pro on a 4h/1h call at Oct 12, 12:01 with confidence 58; the ETH long turn drew NOBODY; the ETH long turn drew NOBODY; the ETH short turn drew 13 caller(s), first gpt-6.1-sol (spliced) on a 1d/1d call at Oct 16, 00:01 with confidence 76; the ETH long turn drew NOBODY. Scope in v1 remains BTC/ETH only; the event count across eleven issues is still far too small to characterise turn-detection skill either way.

Cross-TF and cross-FH confirmation: does agreeing with yourself help?

Block	FH	vs.	Agree n	Agree hit	Dis n	Dis hit	Base hit	Lift
cross-TF	4h	TF 4h vs 1h	579	39.4%	47	66.0%	44.4% (873)	-5.1
cross-TF	1d	TF 1d vs 4h	141	49.6%	0 *	n/a	51.9% (160)	-2.2
cross-FH	4h	by FH 1h	525	38.3%	42	54.8%	44.4% (873)	-6.2
cross-FH	1d	by FH 4h	127	48.8%	0 *	n/a	51.9% (160)	-3.1

* = N<10 (insufficient); insufficient cells are never used to rank models. Mixed pairs (a model with only one side of the comparison present): 245 / 19 / 303 / 33. Pooled across all four cell families the lifts run -6.2pp to -2.2pp this week (issue #10: -2.0pp to +6.8pp).

Regime control. Methodology calls for a trend-vs-flat split on this table. This window has 3 trend days of 7: 10-12 -2.17% (trend), 10-13 +1.56% (trend), 10-14 -1.68% (trend), 10-15 +0.52%, 10-16 -0.77%, 10-17 -0.26%, 10-18 +0.83%. On trend days, 4h TF-agreement hit 43.2% (n=227) against 36.9% (n=352) on flat days.

Side-mix and the week's failure

Model	N	Long	Short	Sideways	L:S
claude-fable-5	1,330	21.6%	31.4%	47.1%	0.69
claude-opus-5.5 (spliced)	1,310	15.3%	28.4%	56.3%	0.54
deepseek-v4-pro	1,300	15.8%	43.1%	41.0%	0.37
gemini-3.1-pro	1,308	19.3%	39.2%	41.5%	0.49
gpt-6.1-sol (spliced)	1,310	12.4%	49.8%	37.8%	0.25
grok-4.7 (spliced)	1,310	11.8%	41.3%	46.9%	0.28
qwen-3.8-max	1,330	19.2%	38.4%	42.4%	0.50

L:S = each model's own long-share divided by its short-share; the field average of the per-model ratios is 0.45 this week. 'Sideways' was 44.7% of mature forecasts (4,114 of 9,198). Consensus skew: 605 symbol/FH/TF cells had 5 or more models on the same side and hit 45.6% (n=3,811) against the 46.6% field base.

Model	Symbol	Side	Conf	FH / TF	Slot (UTC)	Exit	Net PnL
gpt-6.1-sol (spliced)	ETH	Short	82	4h / 1h	Oct 15, 04:01	EXPIRY	-$0.74

Fail of the week. The week's single highest-confidence individual miss, exit reason expiry. Listed as a single card, never as a model ranking.

Calibration bridge

deepseek-v4-pro is both this week's hit-rate leader (48.2%) and the lowest-Brier line (Brier 0.2657, gap +11.6pp). Per-model ok-rate this window ran 98.50% to 100.00%. The full confidence-bucket analysis lives in Weekly Calibration #11.

Market check: bigger daily moves and higher BTC volatility

OBSERVATION — mean |1d move| 1.54% -> 2.05% (+33% rel), BTC realized vol 30.9% -> 34.9% (ann., hourly); field directional accuracy 43.8% -> 46.6% (+2.8pp), 6 of 7 models improved, sim win-rate up for 7 of 7, field sim PnL -$520.82 -> -$352.09 (adjacent calendar weeks Oct 5-11 vs Oct 12-18; descriptive, one pair of weeks, not a claim).
Robustness: raw price-sign accuracy 45.5% -> 44.1% — the opposite direction to the trade-based hit rule.

Measure	Oct 5-11 (week A)	Oct 12-18 (week B)	Change
Mean |1d move| (5 assets)	1.54%	2.05%	+33% rel
BTC realized vol (ann., hourly)	30.9%	34.9%	+3.9pp
Field directional accuracy	43.8%	46.6%	+2.8pp
Models improving hit-rate	--	6 of 7	--
Raw price-sign accuracy	45.5%	44.1%	-1.4pp
Field sim win-rate	34.5%	39.2%	+4.7pp
Field sim net PnL	-$520.82	-$352.09	--

Week A is exactly the issue-#10 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: the model cutover (2026-10-17T09:01:00+00:00, inside week B, day 6 of 7) splices three lines -- gpt-5.6-sol -> gpt-6.1-sol, claude-opus-5 -> claude-opus-5.5, grok-4.6 -> grok-4.7. Week A (2026-10-05..2026-10-11) is entirely before the cutover: its old-id rows are aggregated under the heads (it reproduces the published issue #10); week B mixes the old-id era with the new ids. The grok 4.5 -> 4.6 flip (2026-08-24T09:22:00+00:00) sits seven windows back (inside issue #4's window).

Practical implications

- Do not read this week's leaderboard as a skill ranking: no title was awarded and the top CIs overlap.
- Treat per-ticker readability as weekly weather. Eleven issues, and the best ticker changed hands in every one except issues #6 and #10 (SOL -> BNB -> BTC -> SOL -> BNB -> BNB -> ETH -> XRP -> BTC -> BTC -> XRP): XRP leads this week at 49.0%, after two issues of BTC.

- Self-agreement lifts run -6.2pp to -2.2pp with small disagree cells; carry the regime split into the monthly test, do not act on it.

Limitations

- Single week (Oct 12-18, 2026 UTC). Descriptive for this window only; no claim about next week or any model's underlying skill. This window ran 3 trend days of 7 (issue #10: 2 of 7); every comparison with issue #10 is a comparison across regimes as well as across weeks.
- Observations inside one window are not independent, and the 95% Wilson CIs shown throughout are descriptive, not inferential.
- The reversal detector remains v1 and BTC/ETH-only, on price series reconstructed from trade entry prices. 23 events across eleven issues cannot characterise detector performance either way.
- Cross-TF and cross-FH disagreement cells are small (n=47, 0, 42, 0); cells under N=10 are insufficient and never used to rank models. Week-over-week deltas are computed on unrounded rates and may differ by 0.1pp from the difference of the rounded columns.
- Model lines in this window: claude-fable-5, claude-opus-5.5 (spliced), deepseek-v4-pro, gemini-3.1-pro, gpt-6.1-sol (spliced), grok-4.7 (spliced), qwen-3.8-max. DeepSeek announced routing of V4 Pro to V4.1 Flash from Sep 14, 04:00 UTC; an API check on Oct 21 still returns deepseek-v4-pro as the served model for our requests, with the same system fingerprint as on Oct 13, so this issue keeps the label; it will change the day the served id changes. In this window the line's ok-rate is 100.0% (invalid 89 this window). Series density and lineage notes (wave 11): (1) rolling-1w forecasts emit daily since Aug 22, 2026 and rolling-1M daily since Aug 23, 2026, so inside this window (Oct 12-18) daily coverage is FULL, as it was in issues #4, #5, #6, #7, #8, #9 and #10: 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 483). (2) Model lineage is a splice. Three lines flipped at Oct 17, 2026 09:01 UTC, INSIDE this window (day 6 of 7): "claude-opus-5.5 (incl. opus-5-era, Oct 12 00:00-Oct 17 09:01 UTC)" (1,155 rows from the claude-opus-5 era, 315 from claude-opus-5.5); "gpt-6.1-sol (incl. 5.6-era, Oct 12 00:00-Oct 17 09:01 UTC)" (1,155 rows from the gpt-5.6-sol era, 315 from gpt-6.1-sol); "grok-4.7 (incl. 4.6-era, Oct 12 00:00-Oct 17 09:01 UTC)" (1,155 rows from the grok-4.6 era, 315 from grok-4.7). Week A (the issue-#10 window) carries no new-build rows, so each of the three week-over-week rows joins a spliced line to a single-build line, and a remapped comparison is not the same model build. (3) The two hourly slots before the flip (Oct 17, 07:01 and 08:01 UTC) hold 60 invalid rows (claude-opus-5 20, gpt-5.6-sol 20, grok-4.6 20), counted in the invalid counter as they are; in those two hours the other 4 lines (claude-fable-5, deepseek-v4-pro, gemini-3.1-pro, qwen-3.8-max) carry ok rows. (4) deepseek-v4-pro carries 1,440 rows in this window against 1,470 for every other line (Oct 15: 180 rows instead of 210). (5) The grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits seven windows back (inside the issue-#4 window), and 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 #10 after the audit that found the raw table mixes two exchanges; issue #10 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 19 16:00 UTC) -- the definition pinned in issue #3 and carried by issues #4, #5, #6, #7, #8, #9 and #10, which printed 66,861 at the Oct 12 16:00 cutoff. The pack reproduces that pin at the previous cutoff (control OK), so this issue's 71,945 is an additive step under one definition; counter deltas against issue #9 and earlier remain definitional.

Counters & lineage

Counter	Value	Counter	Value
Forecasts total	10,260	Mature (scored pool)	9,198
OK in gate	9,198	-- of them directional	5,084
Out of gate (1w / 1M)	490 / 483	-- of them sideways	4,114
Invalid	89	Pending (next issue)	0
Late closes	0	Uptime, all grids	266 / 266 slots
Source file	weekly_metrics_2026-10-12.json	Report cutoff	Mon Oct 19, 2026, 16:00 UTC
Generated at (pipeline)	2026-10-21 03:36 UTC	Methodology	v1.1 (2026-08-10) · hash e66c7e8c864a2233

THIS REPORT: 5,084 scored observations. MARKETMANIA RESEARCH TO DATE (as of Oct 19, 2026 cutoff): 71,945 directional forecasts resolved since Jul 11 · 7 models tracked (current line-up; earlier versions folded into their successors' lineage) · 5 assets · 5 horizons · hourly · 57 published reports.

Issue #11. Weekly Model Watch is a living series. Issue #10's open questions -- "does claude-opus-5 hold the top rank for a second issue, and does the title rule ever fire?" and "Does a qualifying reversal fire again after an issue with none, and does a second turn in the same pair ever draw a caller?" -- read this issue as: deepseek-v4-pro at the top (48.2%), not the line that led issue #10 (claude-opus-5), so the rank did not hold for a second issue; no weekly title was awarded, for the eleventh issue running, and 2 of 6 qualifying turns found a caller, on turns in BTC / ETH; five of the six were ETH turns, and the second of them (Tue Oct 13, 00:00) drew NOBODY while the fourth (Fri Oct 16, 12:00) drew 13 callers. Three rows are lineage splices this issue (claude-opus-5.5 (incl. opus-5-era, Oct 12 00:00-Oct 17 09:01 UTC); gpt-6.1-sol (incl. 5.6-era, Oct 12 00:00-Oct 17 09:01 UTC); grok-4.7 (incl. 4.6-era, Oct 12 00:00-Oct 17 09:01 UTC)) because the flip (Oct 17, 2026 09:01 UTC) fell inside the window; in issue #10 they were pure claude-opus-5, gpt-5.6-sol and grok-4.6. The grok 4.5 -> 4.6 flip (Aug 24, 2026 09:22 UTC) sits seven 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: does deepseek-v4-pro hold the top rank for a second issue, and does the title rule ever fire?
- Does a qualifying reversal fire outside ETH for a second issue running, and does a second turn in the same pair ever draw a caller?
- Monthly series: cross-TF/FH lifts with regime control across mixed weeks (trend vs. flat), block bootstrap.

Related research

Report	Direct PDF link
Consensus Watch #11	https://marketmania.ai/research/reports/consensus-watch-2026-10-12.pdf
Weekly Calibration #11	https://marketmania.ai/research/reports/weekly-calibration-2026-10-12.pdf
Weekly Model Watch #10	https://marketmania.ai/research/reports/model-watch-2026-10-05.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_modelwatch_2026w42,
  title  = {Weekly Model Watch #11: weekly leaderboard, ticker/pair reads and cross-confirmation, Oct 12-18 2026},
  author = {{MarketMania Research}},
  year   = {2026}, month = {October}, day = {21},
  url    = {https://marketmania.ai/research/reports/model-watch-2026-10-12.pdf},
  note   = {Methodology v1.1 (2026-08-10), hash e66c7e8c864a2233; window Oct 12-18, 2026 UTC; source weekly_metrics_2026-10-12.json}
}
```