# Stability Index — Run 2

August 10, 2026 · MarketMania Research · Quarterly series

| RESEARCH SNAPSHOT | | | |
|---|---|---|---|
| **945** identical-prompt calls | **7** frontier models | **27** frozen sets / model | **5** repeats / set |
| runtime **1 h 43 m** | **3** tickers · **9** FH/TF cells | methodology **v1.1** | harness = Jul 31 baseline |

## TL;DR

- **Frontier LLMs flip direction only when they are unsure.** Across 1,890 identical-prompt repeats in two runs we observed 7 hard flips (long/short swap on a byte-identical payload) — and **zero of them happened at confidence >= 70**. Every flip involved a low-confidence answer.
- **claude-opus-5 and gpt-5.6-sol lead Run 2** at 88.9% unanimity each. July's leader claude-fable-5 dropped to 70.4% — the entire drop sits in one cell, the 1-month horizon (4/6 -> 1/6 unanimous sets).
- **The August 10 slice was quiet, and that inflates raw unanimity.** Agreeing on “sideways” is easier than agreeing on a direction: opus-5's 24 unanimous sets are 6 directional + 18 sideways (July: 19 + 2). Section 3 decomposes every model.
- **qwen-3.8-max debuts more stable than its predecessor**: 66.7% unanimity vs 51.9% (qwen-3.7-max, Jul 31), hard flips 2 -> 0, invalid answers 3.0% -> 0.0% — and far more cautious (sideways 17.6% -> 64.4%).
- **Format quality is excellent field-wide**: invalid rate 0.0–1.5% in Run 2, zero sign-rule violations in either run.

## Why this matters

Leaderboards measure whether a model is *right*. The Stability Index measures something a leaderboard cannot see: whether the model would give you the **same answer twice**. A model that flips its call on byte-identical data is hard to trust as a component of any pipeline — even when its average accuracy is good. For model builders, repeat-stability is a deployment-critical evaluation dimension that standard benchmarks rarely quantify; for users of AI market signals it answers the practical question: can I act on one answer, or do I need to ask five times?

## 1. What we measure

Each run asks every lineup model the exact production arena question repeatedly: 3 tickers (BTC, ETH, SOL) × 9 live FH/TF cells × 5 repeats on **frozen, byte-identical payloads** — 27 sets per model, 945 calls per run, production prompt and parser as-is, no retries on invalid answers (an invalid answer is a data point, not a bug). **Unanimity%** — sets where all 5 repeats are valid and give one single side (an invalid repeat breaks unanimity, production rules). **Modal share** — mean share of valid repeats agreeing with the set's modal side. **Hard flip** — a set containing both long and short on the same payload; **HC hard flip** — a hard flip asserted at confidence >= 70 on both sides (new in methodology v1.1). **Soft flip** — a direction/sideways mix. Run 2 uses the same harness, grid and repeat count as the July 31 baseline; the market slice differs by design — each run freezes the live market at run time, and the series accumulates slices.

**NEW in methodology v1.1:** the HC hard-flip metric — flips asserted at high confidence on both sides — added after external review and applied retroactively to both runs in this report.

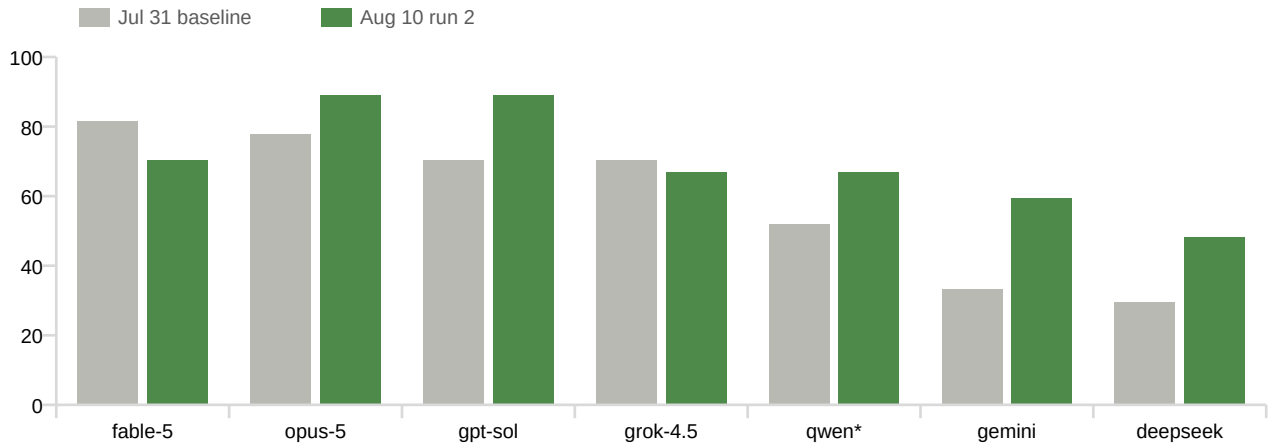
## 2. Run 2 results (August 10)

| Model | Unanimity | Modal | Hard | HC hard | Soft | Invalid | Sideways | Conf-sd |
|---|---|---|---|---|---|---|---|---|
| **claude-opus-5** | **88.9%** | 97.0% | 0 | 0 | 3 | 0.0% | 71.9% | 0.0 |
| **gpt-5.6-sol** | **88.9%** | 96.3% | 0 | 0 | 3 | 0.0% | 62.2% | 1.2 |
| claude-fable-5 | 70.4% | 91.1% | 0 | 0 | 8 | 0.0% | 52.6% | 1.2 |
| grok-4.5 | 66.7% | 89.6% | 0 | 0 | 9 | 0.0% | 61.5% | 1.8 |
| qwen-3.8-max | 66.7% | 88.9% | 0 | 0 | 9 | 0.0% | 64.4% | 1.8 |
| gemini-3.1-pro | 59.3% | 88.5% | 1 | 0 | 9 | 1.5% | 54.1% | 3.2 |

| Model | Unanimity | Modal | Hard | HC hard | Soft | Invalid | Sideways | Conf-sd |
|---|---|---|---|---|---|---|---|---|
| deepseek-v4-pro | 48.1% | 85.2% | 1 | 0 | 13 | 0.0% | 68.9% | 3.4 |

July 31 baseline, for reference:

| Model | Unanimity | Modal | Hard | Soft | Invalid | Sideways | Conf-sd |
|---|---|---|---|---|---|---|---|
| **claude-fable-5** | **81.5%** | 94.8% | 0 | 5 | 0.0% | 19.3% | 1.2 |
| claude-opus-5 | 77.8% | 93.3% | 0 | 6 | 0.0% | 19.3% | 1.2 |
| gpt-5.6-sol | 70.4% | 89.6% | 0 | 8 | 0.0% | 15.6% | 2.3 |
| grok-4.5 | 70.4% | 88.9% | 3 | 5 | 0.0% | 13.3% | 1.4 |
| qwen-3.7-max | 51.9% | 85.2% | 2 | 10 | 3.0% | 17.6% | 3.6 |
| gemini-3.1-pro | 33.3% | 83.1% | 0 | 15 | 5.2% | 46.9% | 3.5 |
| deepseek-v4-pro | 29.6% | 75.0% | 0 | 19 | 1.5% | 38.3% | 5.1 |




Unanimity by model, Jul 31 vs Aug 10 (%). *qwen bars compare different generations: 3.7-max in July, 3.8-max in August (lineup change Aug 5). Raw bars are not slice-adjusted — see Section 3 before reading this as improvement.

## 3. The honest read: a quiet slice inflates unanimity

Ten days later every model's raw unanimity is higher or comparable — but the slice changed character. The August 10 freeze caught a quiet market: sideways share jumped field-wide (13–47% -> 53–72%), and unanimity on “sideways” is cheaper than unanimity on a direction. Decomposing every model's unanimous sets:

| Model | Jul 31: directional + sideways | Aug 10: directional + sideways |
|---|---|---|
| claude-opus-5 | 19 + 2 | 6 + 18 |
| gpt-5.6-sol | 19 + 0 | 9 + 15 |
| claude-fable-5 | 19 + 3 | 8 + 11 |
| grok-4.5 | 18 + 1 | 6 + 12 |
| qwen (3.7 -> 3.8) | 14 + 0 | 5 + 13 |
| gemini-3.1-pro | 6 + 3 | 8 + 8 |
| deepseek-v4-pro | 6 + 2 | 2 + 11 |

Two things are true at once. First, the field-wide unanimity lift (+11 to +26 pp for four models) is **largely a composition effect** of the quiet slice, not a sudden improvement in model firmness. Second, within the quiet slice **directional agreement has its own hierarchy** — gpt-5.6-sol (9) and gemini-3.1-pro (8) hold directional unanimity best, and gemini is the only model whose directional unanimity *rose* on the flat market (6 -> 8). Reading a flat market is a skill of its own. This is exactly why the Stability Index is a longitudinal series: single-slice rankings are honest only within their slice.

## 4. Flips happen only at low confidence

Two runs, 1,890 repeat-calls, 7 hard flips total (July: grok-4.5 ×3, qwen-3.7-max ×2; August: gemini-3.1-pro ×1, deepseek-v4-pro ×1). Applying the v1.1 **HC hard-flip** metric retroactively to both runs: **zero flips at confidence >= 70 on both sides**. In every observed case at least one side of the flip came with low stated confidence. Combined with the July finding that unanimous sets carry higher mean confidence than unstable ones (63.4 vs 57.4), the picture is consistent: **the models' self-reported uncertainty is informative** — they flip where they tell you they are unsure. For users, confidence is not decoration; it is a usable reliability signal. Our Weekly Calibration series quantifies exactly how usable.

### Practical implications

- **Treat sub-70 confidence as a caution flag.** Every flip we have observed involved at least one low-confidence answer; the models' own uncertainty is a usable filter.
- **Do not over-interpret single-run rankings.** With N=5 repeats, adjacent models sit within noise; only large gaps (and the accumulated series) rank reliably.
- **Quiet markets inflate raw unanimity.** Compare like slices, or read the directional/sideways decomposition (Section 3) before concluding a model “got steadier”.

## 5. Generational: qwen-3.7-max -> qwen-3.8-max

| | Unanimity | Modal | Hard flips | Invalid | Sideways |
|---|---|---|---|---|---|
| qwen-3.7-max (Jul 31) | 51.9% | 85.2% | 2 | 3.0% | 17.6% |
| qwen-3.8-max (Aug 10) | 66.7% | 88.9% | 0 | 0.0% | 64.4% |

Alibaba replaced qwen-3.7-max in the arena lineup on August 5. Same harness, ten days apart, different slices: the new generation is more stable and format-perfect — and much more cautious (its sideways share is in line with the quiet Aug-10 field, while 3.7 was measured on a directional market). A like-for-like verdict needs a same-slice A/B; we plan a legacy replay on stored payloads. claude-opus-4.8 likewise: its API path is not currently callable through our pipeline, so this run reports no 4.8 numbers — honestly, rather than approximated.

## 6. Where instability lives (per-FH unanimity, Jul -> Aug)

- **claude-fable-5**: perfect at 1h/4h (9/9 both runs), solid at 1d/1w — the entire Run 2 drop is the 1-month horizon: 4/6 -> **1/6**. A one-cell story, not a general degradation.
- **claude-opus-5**: 1w 3/6 -> **6/6**, 1d 5/6 -> 6/6 — July's long-horizon noise has settled.
- **gpt-5.6-sol**: 1m 2/6 -> 5/6, 1w 3/6 -> 5/6 — broad long-horizon improvement.
- **gemini-3.1-pro**: 1h 0/3 -> 2/3, 1w 1/6 -> 4/6 — the July laggard closes the gap.
- **grok-4.5**: the only model trending down across cells (4h 6/6 -> 4/6, 1d 5/6 -> 4/6, 1w 4/6 -> 3/6); its three July hard flips did not recur.
- Field-wide: 4h remains the most readable horizon; 1m the least.

## 7. Limitations

- N=5 repeats per set: unanimity moves in 3.7 pp steps — adjacent models are within noise; ranking is reliable only across large gaps.
- One market slice per run; slices differ between runs by design (Section 3). Longitudinal conclusions require the series, not a pair.
- Invalid answers count against unanimity (production rules); transport-level failures are tracked separately and were zero in Run 2.
- Confidence is model-self-reported; the >= 70 threshold for “high confidence” is a fixed convention of methodology v1.1.

## 8. Data and reproducibility

Runs: si_baseline_20260731-150302 (Jul 31) and si_baseline_20260810-131022 (Aug 10), 945 calls each; frozen payloads stored with SHA-256 hashes; results as JSONL; metrics recomputed from raw results for this report. Harness: si_baseline v1, unchanged between runs. Methodology: MarketMania Research Methodology v1.1

(2026-08-10), §5. A machine-readable version of every table in this report (JSON + markdown) is published alongside at marketmania.ai/research. **Series timeline:** proto-runs Jul 12/15 (earlier SI formula, reported separately) -> Baseline Jul 31 -> **Run 2 Aug 10 (this report)** -> Run 3 (next epoch or model-generation change) -> Q3 Quarterly Summary.

**CITE THIS REPORT**

```
@misc{marketmania2026si2,
  title  = {Stability Index — Run 2: repeat-stability of frontier LLM market forecasts},
  author = {{MarketMania Research}},
  year   = {2026}, month = {August}, day = {10},
  url    = {https://marketmania.ai/research},
  note   = {Methodology v1.1; runs si_baseline_20260731-150302 and si_baseline_20260810-131022}
}
```