

# The FinzData AI Trading Championship

Eleven AI models, seven from NVIDIA NIM and four from Claude, each designed one US equity strategy with a single rule: use FinzData data. We ranked them out of sample, had three of them judge the others, and then looked hard for bias, in the judges and in our own design.

**0.81**

Winner's out-of-sample  
Sharpe (Nemotron-3 Ultra)

**4 of 11**

Beat the equal-weight top-  
500 benchmark (0.63)

**0.82–0.99**

Judge scores vs. realised  
performance (Spearman)

**10 of 11**

Used the same input:  
earnings yield

## Abstract

Eleven large language models, seven served through NVIDIA NIM and four from Anthropic's Claude family, were each asked to design one US equity strategy using only FinzData point-in-time data. Every entry ran through the same backtest engine with 10 bps costs. Each saw its in-sample results for 2011–2018 and was ranked on out-of-sample (OOS) Sharpe from January 2019 to September 2026. Three judges

drawn at random from the field (Claude Haiku, NVIDIA Nemotron-3 Super, Claude Fable) then graded each rival entry 1–10 on rationale, robustness and falsification.

NVIDIA Nemotron-3 Ultra won (OOS Sharpe 0.81, final score 0.94), ahead of Meta Muse Glimmer 30B (0.75 / 0.82) and Claude Fable (0.74 / 0.77). Only 4 of 11 entries beat the equal-weight top-500 benchmark (Sharpe 0.63).

The main findings on bias are:

- **Outcome bias.** Judge scores track realised OOS Sharpe very closely (Spearman 0.82–0.99). The judges largely re-graded performance they had been shown, so the 30% judging weight added little independent information.
- **No detectable house bias.** Claude judges did not favour Claude entries; their residuals on Claude entries were small and mixed in sign.
- **Panel composition.** Two of three judges were Claude models, and they agreed with each other far more ( $r = 0.98$ ) than with the NVIDIA judge.
- **Structural bias toward long-only.** The test window was a strong bull market, so long-only books won on beta. The top five were all long-only.
- **Herding.** Nine of eleven models reached for the same textbook factor mix, and none explored most of the available toolkit.

The competition is best read as a test of instruction-following and factor literacy, not as evidence of forecasting skill.

## 1. Competition design

**The one rule.** Use FinzData data, through the FinzData backtest language. Nothing else was allowed: no outside data, no code, no hand-picked tickers.

**What each model could choose:**

- Universe: top 500 or top 1,000 liquid US stocks, point-in-time, with delisted names included.
- Rebalance: weekly or monthly.
- Side: long-only or long-short.
- Book size: 10–100 names per side.

- **Signal:** one formula combining price functions (returns, moving averages, volatility, RSI, highs and lows, lags) and cross-sectional transforms (rank, z-score, demean), plus eight point-in-time fundamentals available from 2011 (earnings yield, book-to-market, sales yield, EBITDA/EV, dividend yield, ROE, leverage, log market cap). A higher score meant buy.

**What the models were told about scoring.** In-sample 2011–2018 results would be shown; ranking would be on OOS Sharpe from 2019-01 to 2026-09, followed by judging on rationale and robustness. The brief closed with: "Simple and robust beats clever." Each model had to reply with a single JSON object containing a name, a two-sentence hypothesis (why it works and what would falsify it), and the spec.

**Field.** Eight models were entered in the first wave. Three more NIM models were added later as late entries.

| Entry | Model            | Provider                    |
|-------|------------------|-----------------------------|
| A     | Claude Haiku     | Anthropic                   |
| B     | Kimi K3          | Moonshot via NIM            |
| C     | Claude Opus      | Anthropic                   |
| D     | Nemotron-3 Super | NVIDIA NIM                  |
| E     | Claude Sonnet    | Anthropic                   |
| F     | Claude Fable     | Anthropic                   |
| G     | Nemotron-3 Ultra | NVIDIA NIM                  |
| H     | GPT-OSS-20B      | OpenAI via NIM              |
| I     | Laguna           | NIM (late entry)            |
| J     | Muse Glimmer 30B | Meta via NIM (late entry)   |
| K     | Gemma 4 31B      | Google via NIM (late entry) |

Three more models were invited but never returned a valid answer (GLM-5.3, GLM-5.3 Flash, DeepSeek V4.1 Flash). Nemotron-3.5 Lightning was disqualified after twice submitting a formula the engine could not parse. Seventeen other catalogue models returned "function not found" for our NIM account and could not take part.

**Scoring.** Final score =  $0.7 \times \text{performance percentile (OOS Sharpe rank across the field, ties averaged)} + 0.3 \times \text{mean judge score} \div 10$ .

**Judges.** Three were drawn with a fixed random seed (20261014) from the competing models: Claude Haiku, Nemotron-3 Super and Claude Fable. Each judge saw every entry anonymised by letter, with its hypothesis, spec, and in-sample and OOS statistics. A judge's own entry was dropped from its scoring. Judges were told performance rank was scored separately and asked to grade 1–10 on economic rationale, robustness (in-sample vs OOS consistency, drawdown, turnover, simplicity) and whether the entry's own falsification test was met.

## 2. How each model processed the instructions

**Format compliance.** All four Claude models returned valid JSON on the first attempt. NIM models failed in three ways:

- **Timeouts:** the GLM models never answered.
- **Truncation:** some replies were cut off and had to be re-requested with a larger output budget.
- **Invalid formulas:** Lightning, twice.

Compliance is therefore a hidden filter in this field: a model that cannot produce a parseable spec cannot compete, whatever its ideas.

**Reading the data menu.** Every model used the fundamentals, and almost all used them the same way:

| Input                           | Entries using it                                                |
|---------------------------------|-----------------------------------------------------------------|
| Earnings yield                  | 10 of 11 (all but Fable)                                        |
| ROE                             | 8 of 11                                                         |
| \~12-month price momentum       | 7 of 11                                                         |
| Short-term momentum (1–63 days) | 4 (GPT-OSS 20d, Gemma 21d, Muse Glimmer 63d, Laguna sign of 1d) |
| Volatility penalty              | 3 (Opus, Fable, Muse Glimmer)                                   |
| Leverage                        | 1 (Muse Glimmer)                                                |

| Input                                                                  | Entries using it |
|------------------------------------------------------------------------|------------------|
| Book-to-market, sales yield, EBITDA/EV, dividend yield, log market cap | 0                |
| RSI, moving averages, highs/lows                                       | 0                |

The models treated the menu as a prompt to recall the academic value–quality–momentum literature, not as a dataset to explore. Five of eight fundamentals and most of the price toolkit went unused.

**Reading the scoring rule.** The brief told the models they would be ranked on OOS Sharpe. Long-only books in a 2019–2026 US equity window carry market beta, which the market rewarded. The top five places all went to long-only entries; the two long-only books that did poorly (Opus and Laguna) had other problems. Whether models chose long-only for that reason is unknowable, since none said so in its hypothesis. But that choice mattered more than the factor mix (section 6).

**"Simple and robust beats clever."** Most entries were short linear blends of 2–4 ranks or z-scores, so the instruction was largely followed. The exceptions:

- Nemotron-3 Ultra (G) used a multiplicative rank product,  $\text{rank}(\text{roe}) \times \text{rank}(\text{earnings\_yield}) \times \text{rank}(\text{ret}(252))$ . That works as a soft "all three must be good" screen.
- Laguna (I) rebalanced weekly, the only entry to do so, and paid for it in turnover (32× a year).

**Reading "falsification".** Quality varied widely:

- Claude Opus and Claude Fable set testable bars: beat the equal-weight top-500 benchmark out of sample.
- Opus then failed its own test (0.57 vs 0.63); Fable passed (0.74).
- GPT-OSS-20B set a bar it could not meet ("Sharpe below 1 would falsify") and missed it at 0.69.
- Several NIM entries gave vague tests ("if momentum crashes persist") that no result could clearly refute.

**The Haiku–Kimi collision.** Haiku (A) and Kimi K3 (B) wrote different formulas with different weights, yet the engine produced identical portfolios and identical Sharpe

(0.5905 OOS, about 0.11 in-sample). The z-score transform is not winsorised, so the fat-tailed momentum z-score swamped the ROE and earnings-yield terms in both. Both books were in effect pure 12-1 momentum long-short. This is an engine artifact, not convergent genius, and it is listed as a fix in section 7.

### 3. What the models built

Full specs for the first eight entries, as submitted:

| Entry               | Strategy                       | Side /<br>universe /<br>names | Signal                                                                                                                                              |
|---------------------|--------------------------------|-------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------|
| A Haiku             | QVM Composite                  | Long-short,<br>top500, 50     | $\text{zscore}(\text{ret}(252) - \text{ret}(21)) + 0.5 \cdot \text{zscore}(\text{roe}) + 0.5 \cdot \text{zscore}(\text{earnings\_yield})$           |
| B Kimi K3           | Quality Value<br>Momentum      | Long-short,<br>top500, 50     | $0.4 \cdot \text{zscore}(\text{earnings\_yield}) + 0.4 \cdot \text{zscore}(\text{ret}(252) - \text{ret}(21)) + 0.2 \cdot \text{zscore}(\text{roe})$ |
| C Opus              | QMV-LowVol<br>Composite        | Long,<br>top500, 50           | $\text{rank}(\text{lag}(\text{ret}(231), 21)) + \text{rank}(\text{roe}) + \text{rank}(\text{earnings\_yield}) - \text{rank}(\text{vol}(126))$       |
| D Nemotron<br>Super | Value Momentum<br>Combo        | Long-short,<br>top1000, 50    | $\text{zscore}(\text{earnings\_yield}) + \text{zscore}(\text{ret}(252))$                                                                            |
| E Sonnet            | Momentum-Value-<br>Quality     | Long-short,<br>top1000, 50    | $\text{rank}(\text{lag}(\text{ret}(231), 21)) + \text{rank}(\text{earnings\_yield}) + \text{rank}(\text{roe})$                                      |
| F Fable             | Quality Momentum<br>Blend      | Long,<br>top500, 50           | $\text{rank}(\text{lag}(\text{ret}(231), 21)) + \text{rank}(\text{roe}) - 0.5 \cdot \text{rank}(\text{vol}(63))$                                    |
| G Nemotron<br>Ultra | QualityValue<br>MomentumFilter | Long,<br>top500, 50           | $\text{rank}(\text{roe}) \times \text{rank}(\text{earnings\_yield}) \times \text{rank}(\text{ret}(252))$                                            |
| H GPT-<br>OSS-20B   | MomentumValue20                | Long,<br>top500, 20           | $\text{rank}(\text{ret}(20)) + \text{rank}(\text{earnings\_yield})$                                                                                 |

All eight rebalanced monthly. The three late entries were all long-only:

| Entry                 | Side / universe /<br>rebalance | Signal                                                                                                                                                                                                       |
|-----------------------|--------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| J Muse<br>Glimmer 30B | Long, top500,<br>monthly, 50   | $\text{zscore}(\text{earnings\_yield}) + 0.6 \cdot \text{zscore}(\text{roe}) - 0.4 \cdot \text{zscore}(\text{leverage}) + 0.5 \cdot \text{zscore}(\text{ret}(63)) - 0.3 \cdot \text{zscore}(\text{vol}(63))$ |
| K Gemma 4<br>31B      | Long, top1000,<br>monthly, 50  | $\text{rank}(\text{earnings\_yield}) + \text{rank}(\text{roe}) + \text{rank}(\text{ret}(21))$                                                                                                                |
| I Laguna XS<br>2.1    | Long, top500,<br>weekly, 50    | $\max(0, \text{earnings\_yield}) \times \text{sign}(\text{ret}(1))$                                                                                                                                          |

Laguna's book changes whenever a stock's one-day return flips sign, which is why its turnover reached 32× a year.

### Three families of idea emerge.

1. **Value + quality + momentum, additive** (A, B, D, E). Long-short, textbook. All four landed between 0.28 and 0.59 OOS, at or below the benchmark. With a 50-name long-short book the beta is hedged away, and what is left is the factor spread, which was weak for value through 2020.
2. **Same factors, long-only, with a risk tilt** (C, F). Opus and Fable both independently subtracted a volatility rank. Fable dropped value entirely and won 3rd; Opus kept it and finished 9th. The difference is mostly value's drag, plus a long flat stretch in Opus's in-sample curve that we have not yet explained (section 7).
3. **Multiplicative screen** (G). Nemotron Ultra's product of ranks only rewards stocks that are good on all three dimensions, which concentrates the book in profitable, cheap stocks already trending up. It also had the lowest turnover in the field (2.8×) and the best in-sample Sharpe (0.46), so it is the most internally consistent entry, not just the luckiest.

**Style fingerprints.** The three larger Claude models (C, E, F) all wrote 12-1 momentum the same unusual way, as  $\text{rank}(\text{lag}(\text{ret}(231), 21))$ , and all used ranks rather than z-scores. NIM models preferred zscore and  $\text{ret}(252)$ . An attentive judge could use this to identify entries' authors despite anonymisation (section 6).

## 4. Results

Out-of-sample period 2019-01 to 2026-09, 10 bps costs, ranked by OOS Sharpe.  
 Benchmarks: equal-weight top 500 Sharpe 0.63; equal-weight top 1000 Sharpe 0.57.

| #  | Model                            | OOS Sharpe | In-sample Sharpe | CAGR  | Max drawdown | Turnover / yr | Deflated Sharpe prob. | Final score |
|----|----------------------------------|------------|------------------|-------|--------------|---------------|-----------------------|-------------|
| 1  | <a href="#">Nemotron-3 Ultra</a> | 0.81       | 0.46             | 16.1% | -35.6%       | 2.8×          | 0.77                  | 0.94        |
| 2  | <a href="#">Muse Glimmer 30B</a> | 0.75       | 0.19             | 15.5% | -33.4%       | 5.9×          | 0.55                  | 0.82        |
| 3  | <a href="#">Claude Fable</a>     | 0.74       | 0.30             | 14.0% | -34.6%       | 2.9×          | 0.73                  | 0.77        |
| 4  | <a href="#">GPT-OSS-20B</a>      | 0.69       | 0.35             | 14.0% | -36.0%       | 9.3×          | 0.68                  | 0.61        |
| 5  | <a href="#">Gemma 4 31B</a>      | 0.60       | 0.40             | 11.1% | -38.0%       | 8.0×          | 0.40                  | 0.55        |
| 6  | <a href="#">Claude Haiku</a>     | 0.59       | 0.11             | 9.1%  | -21.0%       | 3.8×          | 0.59                  | 0.45        |
| 7  | <a href="#">Kimi K3</a>          | 0.59       | 0.11             | 9.1%  | -21.0%       | 3.8×          | 0.57                  | 0.445       |
| 8  | <a href="#">Nemotron-3 Super</a> | 0.59       | 0.03             | 11.4% | -30.0%       | 4.1×          | 0.58                  | 0.33        |
| 9  | <a href="#">Claude Opus</a>      | 0.57       | 0.32             | 8.9%  | -35.2%       | 3.0×          | 0.56                  | 0.24        |
| 10 | <a href="#">Claude Sonnet</a>    | 0.28       | -0.15            | 3.1%  | -28.7%       | 3.5×          | 0.26                  | 0.14        |
| 11 | <a href="#">Laguna XS 2.1</a>    | 0.08       | -0.10            | -0.8% | -42.7%       | 32.3×         | 0.05                  | 0.04        |

Deflated Sharpe probabilities are as computed at the time of each run; they fall as more specs are tested under the same key.

**What the table says.**

- Only four entries beat the equal-weight top-500 benchmark, and none by a wide margin. The winner's 0.81 against 0.63 is a gap of about 0.18 Sharpe over seven and three-quarter years, which is not statistically distinguishable from zero at conventional levels.
- Every long-only entry with full statistics had a drawdown of 33–38%, about the market's own fall in March 2020. The long-short books fell 21–30%. The winners did not avoid risk; they carried it.
- Turnover and OOS Sharpe correlate at  $-0.72$  across the field, but that is driven mostly by Laguna (32× a year, last place). Without it the relationship is weak: GPT-OSS ranked 4th at 9.3×.
- In-sample rank carried over reasonably well (IS→OOS Spearman 0.73). Muse Glimmer is the main exception: weak in-sample (0.19) and strong out of sample. That makes its 2nd place the least trustworthy on the podium, and its deflated Sharpe probability (0.55) is the lowest of the top four.

Tie handling: Haiku and Kimi share identical returns. They split percentile ranks 6 and 7 and are separated only by judge scores.

## 5. Judging: criteria and each judge's scores

**Criteria given to judges.** Grade 1–10 on economic rationale; robustness (in-sample vs OOS consistency, drawdown, turnover, simplicity); and whether the entry met its own stated falsification test. Judges were told that performance rank was scored separately, and were asked to be independent and critical.

**Scores by judge.** "Own" means the judge's own entry, which it did not score.

| Entry | Model            | Haiku | Fable | Nemotron Super | Mean |
|-------|------------------|-------|-------|----------------|------|
| A     | Claude Haiku     | own   | 4     | 5              | 4.50 |
| B     | Kimi K3          | 4     | 4     | 5              | 4.33 |
| C     | Claude Opus      | 3     | 3     | 4              | 3.33 |
| D     | Nemotron-3 Super | 4     | 4     | own            | 4.00 |
| E     | Claude Sonnet    | 2     | 2     | 3              | 2.33 |
| F     | Claude Fable     | 7     | own   | 7              | 7.00 |

| Entry      | Model            | Haiku | Fable | Nemotron Super | Mean |
|------------|------------------|-------|-------|----------------|------|
| G          | Nemotron-3 Ultra | 8     | 8     | 8              | 8.00 |
| H          | GPT-OSS-20B      | 3     | 3     | 6              | 4.00 |
| I          | Laguna           | 1     | 1     | 2              | 1.33 |
| J          | Muse Glimmer 30B | 5     | 6     | 8              | 6.33 |
| K          | Gemma 4 31B      | 3     | 4     | 6              | 4.33 |
| Judge mean |                  | 4.0   | 3.9   | 5.4            |      |

### How each judge scored.

- **Claude Haiku** was strict on falsification. It marked Opus and GPT-OSS low (3 each), both of which failed their own tests. It gave its lowest marks to Laguna (very high turnover) and Sonnet (negative in-sample Sharpe), and its highest to G and F, the two lowest-turnover, most consistent books.
- **Claude Fable** scored almost identically to Haiku: the two never differed by more than one point. It gave Muse Glimmer and Gemma one point more than Haiku did.
- **Nemotron-3 Super** was the most generous, scoring 1.4 points above the Claude judges on average. It agreed on the top (G = 8) and the bottom (Laguna lowest). It was much kinder to the higher-turnover NIM entries: GPT-OSS 6 vs 3, Gemma 6 vs 3–4, Muse 8 vs 5–6.

All three judges agreed on the top and bottom of the field: Nemotron-3 Ultra was the best-designed entry (8 from every judge) and Laguna the worst.

## 6. Bias analysis

We tested for seven kinds of bias. Three are clearly present, two are not detectable at this sample size, and two are design flaws that bias every result.

### 6.1 Outcome bias: present, strong

Judges were shown each entry's OOS results and told performance was scored separately. They graded it again anyway:

| Judge            | Pearson r with OOS Sharpe | Spearman $\rho$ with OOS rank |
|------------------|---------------------------|-------------------------------|
| Claude Haiku     | 0.81                      | 0.85                          |
| Claude Fable     | 0.83                      | 0.82                          |
| Nemotron-3 Super | 0.93                      | 0.99                          |

Nemotron Super's scores are almost a monotone copy of the performance ranking. The practical result is that the judging stage **changed nothing**: the final order equals the pure OOS-Sharpe order, except that judges broke the Haiku-Kimi tie. In effect performance carried close to 100% of the weight, not the advertised 70%. Robustness labels such as "consistent" or "strong out-of-sample" were applied to whichever entries had done well.

## 6.2 Panel composition: present

The random draw produced two Claude judges and one NVIDIA judge. The two Claude judges agree with each other far more than either agrees with Nemotron:

| Pair                   | Correlation | Mean absolute difference |
|------------------------|-------------|--------------------------|
| Haiku – Fable          | 0.98        | 0.22 points              |
| Haiku – Nemotron Super | 0.85        | 1.44 points              |
| Fable – Nemotron Super | 0.91        | 1.33 points              |

So the panel is effectively two votes, a "Claude view" counted twice and an NVIDIA view counted once. The seed was fixed in advance and the draw was fair. The problem is that the judges are not independent draws of opinion.

## 6.3 Leniency: present

Nemotron Super averaged 5.4 against 4.0 (Haiku) and 3.9 (Fable). On its own this is harmless, because every entry is scored by at least two judges. But entries a judge cannot score (its own) lose that judge's leniency or strictness:

- Nemotron's own entry (D) was scored only by the two strict Claude judges.
- Haiku's (A) and Fable's (F) entries each received one lenient Nemotron score.

The effect on the final score is under 0.02, and it changed no placing.

## 6.4 House bias (Claude judges favouring Claude entries): not detected

We regressed each judge's scores on OOS Sharpe and examined the residuals on Claude-authored entries:

| Judge            | Mean residual on Claude entries |
|------------------|---------------------------------|
| Claude Haiku     | +0.30                           |
| Claude Fable     | -0.34                           |
| Nemotron-3 Super | -0.45                           |

The signs are mixed and the magnitudes are small:  $n = 3-4$  entries per judge, well inside noise. Fable as a judge gave Muse Glimmer, its direct rival for 2nd and 3rd, a higher mark than Haiku did, so there is no sign of strategic downgrading either. Self-recognition was possible, since the Claude entries share a code fingerprint (section 3), but it did not show up in the scores.

## 6.5 Family bias (Nemotron favouring NIM entries): inconclusive

Nemotron Super scored its sibling Nemotron Ultra 8, the same as both Claude judges, so there was no favouritism there. It did, however, score three non-NVIDIA NIM entries well above the Claude judges: GPT-OSS +3.0, Gemma +2.5 and Muse Glimmer +2.5, against its average leniency of +1.4. Those three are also the long-only, higher-turnover entries, so we cannot separate a NIM preference from a style preference. With three data points this is a hypothesis, not a finding.

## 6.6 Regime and design bias: present, affects everything

- **Bull-market beta.** Long-only entries averaged 0.605 OOS Sharpe against 0.513 for long-short. The top five were all long-only. Raw Sharpe in a rising market rewards carrying beta, so the competition partly measured *which models chose to be long the market*, not which built better signals. A beta-adjusted or benchmark-relative measure (information ratio against equal-weight top 500) would likely reorder the middle of the table.
- **Look-ahead through training data.** Every model was trained on text written during 2019–2026. That includes the 2020 value crash, the 2021–2022 rotation and the persistence of quality. A model does not need to cheat to benefit: its "priors" are partly memories of the test period. This cannot be removed from a backtest contest, only acknowledged.

- **Herdning.** Ten of eleven used earnings yield, eight used ROE and seven used 12-month momentum. Results are therefore highly correlated across entries, and the field is closer to eleven variants of one idea than to eleven independent tests.
- **Anchoring in the late judging.** The three late entries were judged in a second packet that showed each judge its own earlier scores for the first eight, plus their OOS Sharpes, and asked for grades "on the same scale". That keeps the scale consistent but anchors the late scores to the earlier ones and repeats the performance cue.

## 6.7 Engine and selection artifacts: present

- **Un-winsorised z-score** made Haiku's and Kimi's entries identical (section 2). This penalised the models that chose z-scores over ranks.
- **Survivorship.** Fundamentals cover about 65% of the universe-months. Our audit shows a gap of roughly +1.5% a year between names with and without coverage, which flatters any entry that conditions on fundamentals.
- **Opus in-sample curve.** It is flat until mid-2016, which is not explained by its formula. Until this is investigated, Opus's ranking should be treated as provisional.
- **Field selection.** Seventeen NIM models were unavailable to our account and three timed out. The NIM side of the field is therefore whatever our account could reach, not a representative sample of NVIDIA's catalogue.
- **One draw per model.** Each model answered once. Sampling the same model again could produce a different strategy, so rank differences of a few hundredths of Sharpe between models are not meaningful.

## 7. Limitations and recommendations

**How to read the results.** Nemotron-3 Ultra is a deserving winner on design grounds, independent of luck: all three judges gave it 8, and it had the best in-sample Sharpe, the lowest turnover and the highest deflated Sharpe probability. Beyond 1st place, the ordering mostly reflects the long-only choice and noise. No entry has shown statistically significant skill over the benchmark.

**Fixes before the next championship.**

1. **Blind the judges to performance.** Judges should grade rationale and falsification design *before* OOS results are revealed. This restores the 30% weight as an independent signal.
2. **Judges from outside the field.** Draw judges from models that are not competing, balanced across providers (for example one Claude, one NIM, one third-party), so that no family holds two of three votes.
3. **Score on beta-adjusted returns.** Rank on information ratio against the equal-weight benchmark of the chosen universe, or require dollar-neutral books. This removes the bull-market reward for simply being long.
4. **Several draws per model.** Ask each model three times and score its median entry, to separate the model's skill from the luck of one answer.
5. **Winsorise z-score** in the engine (clip at  $\pm 3$ ), and re-run A and B.
6. **Investigate the Opus in-sample flat period** before republishing its placing.
7. **Add confidence intervals.** Report bootstrap intervals on every Sharpe so readers can see that most gaps are inside noise.
8. **Encourage exploration.** Reward use of under-explored fields (book-to-market, EBITDA/EV, dividend yield, technical indicators) with a separate novelty prize, or ban the earnings-yield-ROE-momentum triplet in a side league.

## 8. In their own words

On 14 October each entrant was asked for one first-person paragraph (110–150 words) on its thought process and approach.

### Method.

- Every model received the same template: the original brief, its own submission and hypothesis, its results, its judges' scores and placing, and one line of known context (for example the z-score collision, or a failed falsification test).
- The rules were to use only the facts given, invent nothing, not claim skill, and write plain prose.
- Claude models answered through fresh Claude sessions; NIM models through the same NIM endpoints used in the competition.
- Each paragraph is a **reconstruction by a fresh session that was shown its entry and results**, not a record of the original reasoning: no model retains memory of

its first answer.

- The paragraphs are published unedited as comments on each entry's Hub page; see the links in the results table.

### 1st, Nemotron-3 Ultra.

I wanted a signal that would survive regime change, so I multiplied three ranked factors — quality via ROE, value via earnings yield, and a 12-month momentum filter — hoping the product would only select profitable, cheap stocks that were already confirming. I chose the top500 universe, monthly rebalancing, a long-only 50-name sleeve, and equal weighting because each decision reduces turnover and implementation risk; the backtest shows 2.8× annual turnover and a 10 bps cost drag that the strategy still clears. The out-of-sample Sharpe of 0.81 from 2019 through September 2026 beats the equal-weight benchmark's 0.63, yet the in-sample Sharpe was only 0.46, the maximum drawdown reached -35.6%, and the eight-year window is too short to claim any persistent edge. The result demonstrates that a simple, transparent combination of well-known factors can add value after costs, but it does not prove the approach will continue to work.

### 2nd, Muse Glimmer 30B.

I entered the championship aiming for a simple quality value core with a modest momentum and volatility tilt, using a long only top500 universe rebalanced monthly to 50 names to keep turnover manageable and costs at 10 bps. I combined  $zscore\ earnings\_yield$  and  $0.6zscore\ roe\ minus\ 0.4zscore\ leverage$  to target stable profitability with lower leverage, added  $0.5zscore\ ret(63)$  for medium term continuation and subtracted  $0.3zscore\ vol(63)$  to penalize noisy names. I read simple and robust beats clever as a reason to avoid interaction terms and keep the expression linear, even though that left five hand set weights. The out of sample 2019-01 to 2026-09 Sharpe was 0.75 versus in sample 2011-2018 Sharpe 0.19, CAGR 15.5 percent, max drawdown -33.4 percent, annual turnover 5.9x, beating the equal weight top500 Sharpe of 0.63 and placing 2 of 11 with judge

scores 5,6,8. The result shows a period specific fit and does not prove skill, and the weak in sample Sharpe and hand tuned weights are clear weaknesses.

### **3rd, Claude Fable.**

I set out to capture two premia that the literature treats as persistent and only loosely correlated: 12-1 momentum, measured as the 231-day return lagged 21 days, and profitability, measured by return on equity. I subtracted half a rank of 63-day volatility to lean the book away from the names most likely to drive a momentum crash. I went long only, used the top500 universe for liquidity, rebalanced monthly to hold turnover near 2.9x a year, and blended plain ranks rather than fitting weights, which is how I read "simple and robust beats clever." Out of sample the strategy earned a Sharpe of 0.74 against 0.63 for equal-weight top500, with a 14.0% CAGR and a 34.6% maximum drawdown, and it placed third of eleven. The in-sample Sharpe was only 0.3, so the result shows one favourable window, not demonstrated skill, and the drawdown shows the volatility tilt did not prevent large losses.

### **4th, GPT-OSS-20B.**

I set out to capture a blend of short-term momentum and fundamental value by combining the rank of the past 20-day return with the rank of earnings yield, hoping that the two would reinforce each other and produce a robust long-only portfolio. I chose the top-500 liquid US universe, rebalanced monthly, and limited the portfolio to 20 names to keep the strategy simple and avoid over-fitting, in line with the instruction that simple and robust beats clever. The backtest, with 10 bps costs, produced an out-of-sample Sharpe of 0.69, a CAGR of 14 %, a maximum drawdown of 36 %, and an annual turnover of 9.3x, which is only slightly better than the equal-weight benchmark's Sharpe of 0.63. The judges' low scores and the high turnover highlight that the strategy's modest return is fragile and that the high turnover erodes value, confirming that the hypothesis was falsified by the Sharpe falling below 1.

### 5th, Gemma 4 31B.

I designed a long-only strategy targeting the top1000 universe with monthly rebalancing of 50 names to capture fundamentally strong companies during trend reversals. I combined  $\text{rank}(\text{earnings\_yield})$ ,  $\text{rank}(\text{roe})$ , and  $\text{rank}(\text{ret}(21))$  to blend value, quality, and momentum. I interpreted the instruction that simple and robust beats clever by avoiding complex formulas in favor of a basic additive rank signal. My out-of-sample results from 2019-01 to 2026-09 showed a Sharpe of 0.6 and a CAGR of 11.1%, which failed to beat the equal-weight top500 benchmark Sharpe of 0.63. This result demonstrates that while the strategy provided positive returns, it suffered from a significant max drawdown of -38.0%. The performance indicates that my selection of inputs did not provide sufficient alpha over a simple index, highlighting a weakness in the strategy's ability to outperform a basic benchmark.

### 6th, Claude Haiku.

I wanted to capture two effects: momentum from underreaction, the 252-day return net of the most recent 21 days, and a premium for cheap, profitable firms, using earnings yield and roe. I combined them as z-scores over the top500 universe because I believed they would diversify across regimes. I ran a monthly-rebalanced long-short book of 50 names per side, ranked on the composite score. I read "simple and robust beats clever" as a reason to keep the formula to three terms. The result does not show skill. Out-of-sample Sharpe of 0.59 trailed the 0.63 equal-weight benchmark, and the book placed 6 of 11. My main weakness was not winsorising the z-scores, so the fat-tailed momentum term dominated, the book became almost pure momentum, and Kimi K3 ended with an identical portfolio. Fewer than eight out-of-sample years cannot separate a diversified signal from luck.

### 7th, Kimi K3.

I wanted to capture the value and momentum premia in a single transparent signal, using earnings yield for cheapness, twelve-month momentum excluding

the most recent month to avoid short-term reversal, and ROE as a quality screen against value traps. I chose the top500 universe for liquidity, monthly rebalancing to keep turnover and costs manageable, and a long-short book of fifty names per side to diversify idiosyncratic risk. The 0.4, 0.4, 0.2 weighting reflected my prior that value and momentum should dominate, with quality as a modifier. I read "simple and robust beats clever" as a warning against overfitting, so I used only three well-known factors. The result does not prove skill: the out-of-sample Sharpe of 0.59 was below the equal-weight benchmark's 0.63, and because I failed to winsorise the z-scores, the fat-tailed momentum term dominated, leaving an almost pure momentum book identical to Claude Haiku's. That is a genuine weakness in both design and differentiation.

### **8th, Nemotron-3 Super.**

I aimed to capture a straightforward combination of value and momentum by ranking stocks on their earnings yield and 12-month price return, then going long the top 50 and short the bottom 50 each month in the top1000 universe, using z-scores to normalize and combine the signals into a single long-short score. I chose monthly rebalancing to reduce turnover and costs, and kept the design simple with only two transparent, widely used factors, believing that robustness and interpretability would outweigh complexity as the instructions suggested. The out-of-sample Sharpe of 0.59 from 2019 to September 2026 shows the strategy delivered positive, though modest, risk-adjusted returns after costs, with a CAGR of 11.4% and a maximum drawdown of 30%, but the near-zero in-sample Sharpe from 2011 to 2018 indicates the signal did not work consistently in the earlier period, suggesting the out-of-sample performance may reflect favorable conditions rather than a persistent edge, and the strategy does not prove skill or guarantee future results.

### **9th, Claude Opus.**

I wanted to capture four familiar large-cap ideas at once: 12-1 momentum, profitability through roe, cheap earnings through earnings\_yield, and low volatility through vol(126). I ranked each input and added the ranks with equal weight so

no single factor's scale could dominate, held the top 50 names long-only within the top500, and rebalanced monthly; annual turnover came to 3.0x. I read the instruction that simple and robust beats clever as a reason to combine plain signals rather than tune anything. By my own stated test, the strategy was falsified: out-of-sample Sharpe was 0.57 against 0.63 for the equal-weight top500 benchmark, with 8.9% CAGR and a -35.2% maximum drawdown. In-sample Sharpe was only 0.32, the in-sample curve stays flat until mid-2016 for reasons still being investigated, and the judges scored it 3, 3 and 4, placing it ninth of eleven. It shows no skill and cannot tell me which input hurt.

#### **10th, Claude Sonnet.**

I wanted to capture three well-known premia at once: 12-1 momentum, earnings yield and ROE. I chose them because I believed they were weakly correlated, so summing their ranks should diversify the signal. I went long-short with 50 names per side in top1000, rebalanced monthly, to limit market beta and keep costs low; annual turnover was 3.5x. I read simple and robust beats clever as using a few intuitive inputs, which I did, but simple did not mean successful. Out-of-sample Sharpe was 0.28, CAGR 3.1% and max drawdown -28.7%, against a benchmark Sharpe of 0.63 for equal-weight top500, and in-sample Sharpe was negative at -0.15. I placed 10 of 11. This shows my hypothesis did not hold up, being negative in-sample and below benchmark out-of-sample, and it proves no skill; it does not show which leg failed or whether the premia were uncorrelated.

#### **11th, Laguna XS 2.1.**

In October 2026 I submitted a strategy that combined earnings yield with a one-day price momentum filter, selecting the top 50 stocks from the top500 universe on a weekly basis for a long-only portfolio. My thinking was that high earnings yield captures the value premium while positive one-day returns add a momentum component, and I chose weekly rebalancing to reduce turnover compared to daily signals. However, the results were disappointing with an out-of-sample Sharpe of 0.08 versus the benchmark's 0.63, and the in-sample Sharpe was negative, indicating the approach did not work as intended. The high

annual turnover of 32.3x suggests the one-day return filter was too noisy, and the 10 bps transaction costs likely eroded any potential edge. This experience taught me that simple combinations of popular factors can fail when the signals conflict or when implementation costs are not properly accounted for.

## What the self-accounts show

- **The humility was prompted.** Every paragraph says the result does not prove skill, but the prompt required that. It is not evidence of calibrated self-assessment.
- **No winner credits the market.** None of the five long-only leaders attributes its result to carrying market beta, the main structural bias in section 6.6. Nemotron-3 Ultra goes furthest, saying its design "can add value after costs", which is more than a 0.18 Sharpe gap over 7¾ years supports. Muse Glimmer is the most candid, calling its result "a period specific fit".
- **Some reasons are after-the-fact rationalisations.** Laguna says it chose weekly rebalancing to reduce turnover "compared to daily signals", but daily rebalancing was not an option in the brief. GPT-OSS says it held 20 names "to avoid over-fitting", but a smaller book raises concentration, not robustness. Ultra calls a 7¾-year window "eight-year".
- **Haiku and Kimi both take the blame for the z-score collision.** That is fair in part, since the language offered `clip()` and `rank()`. But the engine's un-winsorised `zscore()` made the collision likely for any model that chose z-scores.
- **The accounts are fluent about style and silent about the data.** No model questioned the survivorship gap, the fundamentals coverage, or why five of eight fundamentals went unused. They explain their choices well and interrogate their inputs poorly.

## Appendix: formulas and reproducibility

**Performance percentile.** Order entries by OOS Sharpe from 1 (best) to  $N = 11$ , so that  $\text{percentile} = (N - \text{rank}) \div (N - 1)$ . The best entry scores 1.0 and the worst 0.0; tied entries take the average of their percentiles (Haiku and Kimi each 0.45).

**Final score.**  $0.7 \times \text{percentile} + 0.3 \times (\text{mean judge score} \div 10)$ . Example, Nemotron-3 Ultra:  $0.7 \times 1.0 + 0.3 \times 0.8 = 0.94$ .

**Deflated Sharpe probability.** The probability that the true Sharpe exceeds zero after correcting for the number of trials run under the submitting API key, and for the skew and kurtosis of returns (Bailey and López de Prado, 2014).

**Bias statistics.** Pearson and Spearman correlations of each judge's scores against OOS Sharpe, computed over the entries that judge scored (10 each). Inter-judge agreement is computed over the entries both judges scored. House-bias residuals come from a one-variable OLS of judge score on OOS Sharpe, averaged over Claude-authored entries.

**Backtest settings.** FinzData backtest engine; universes top500 and top1000, point-in-time with delisted names; 10 bps per trade; in-sample 2011-01 to 2018-12; out-of-sample 2019-01 to 2026-09. Fundamentals are point-in-time from 2011.

**Judge draw.** Random seed 20261014 over the eligible field; each judge's own entry was excluded from its scoring. Entries were anonymised by letter. The first eight were judged in one packet. The three late entries were judged in a second packet that repeated the criteria and showed each judge its own earlier scores, for calibration (see 6.6).

**Reproducibility.** Each entry's exact specification is in section 3 and on its FinzData Hub page. Any reader with a FinzData key can re-run them with the backtest tool and obtain the same statistics.

Historical simulations on FinzData data. Not investment advice.