

AI Futures Championship

12 AI models designed futures strategies across 48 markets.
One beat the trend benchmark.

Winner: Claude Haiku · out-of-sample Sharpe 0.75 vs 0.65 for 12-month trend

Why futures, and why this test matters

Stocks

Our first championship asked 11 AI models to pick US stocks; Nemotron-3 Ultra won with a Sharpe of 0.81. This one moves to the harder macro problem

Long/short

Futures go short as easily as long, across every asset class: the question becomes which way rates, oil, gold, grains and currencies move

Trend

The benchmark is 12-month trend following: one line of code, decades of evidence, and the core of the managed-futures industry

2019-26

The test window holds every kind of market: the 2020 crash, the 2022 inflation shock and bond sell-off, and the 2026 oil shock

Breadth and depth: 48 markets, 7,041 futures contracts

Every listed contract since 2007, rolled and back-adjusted: 3.7 million daily contract records feed the test

Sector	Mkts	Markets traded	Contracts
Equity indexes	9	S&P 500, Nasdaq-100, Russell 2000, Dow Jones, Euro Stoxx 50, DAX, FTSE 100, CAC 40, Nikkei 225	1,054
Government bonds	8	US 2-year, US 5-year, US 10-year, US 30-year, Schatz, Bobl, Bund, Buxl	624
Currencies	8	euro, yen, Aussie dollar, pound, Canadian dollar, Mexican peso, NZ dollar, Swiss franc	1,300
Energy	6	WTI crude, Brent crude, gasoil, natural gas, gasoline, heating oil	1,373
Metals	5	gold, silver, copper, platinum, palladium	1,208
Agriculture	12	corn, soybeans, wheat, soy oil, soy meal, KC wheat, sugar, coffee, cocoa, cotton, cattle, hogs	1,482
Total	48	4,873 trading sessions, January 2008 to October 2026	7,041

Breadth of the field and depth of the test

12

AI models from 9 labs: Anthropic, DeepSeek, Google, Meta, MiniMax, NVIDIA, OpenAI, Poolside, Z.ai. Frontier and small open models side by side

36

independent strategies, three per model; the median draw counts, so one lucky idea cannot carry a model

233k

daily carry readings from front versus next contract, beside 227k market-days of returns, all computed point in time

2,000

bootstrap resamples per strategy, a deflated Sharpe for 36 trials, and three blind judges from three labs

How the championship worked

48

futures markets in six sectors: equity indexes, rates, currencies, energy, metals and agriculture, every contract back-adjusted and rolled

12

AI models, three independent draws each (36 strategies). Each wrote a hypothesis, a falsification test and one signal in a fixed formula language

2009

to 2018 was the design window. Models saw only that period. 2019 to September 2026 was held out and scored once

10%

volatility target, next-session execution, 3 bps per unit of turnover. Same engine and same risk for every entry

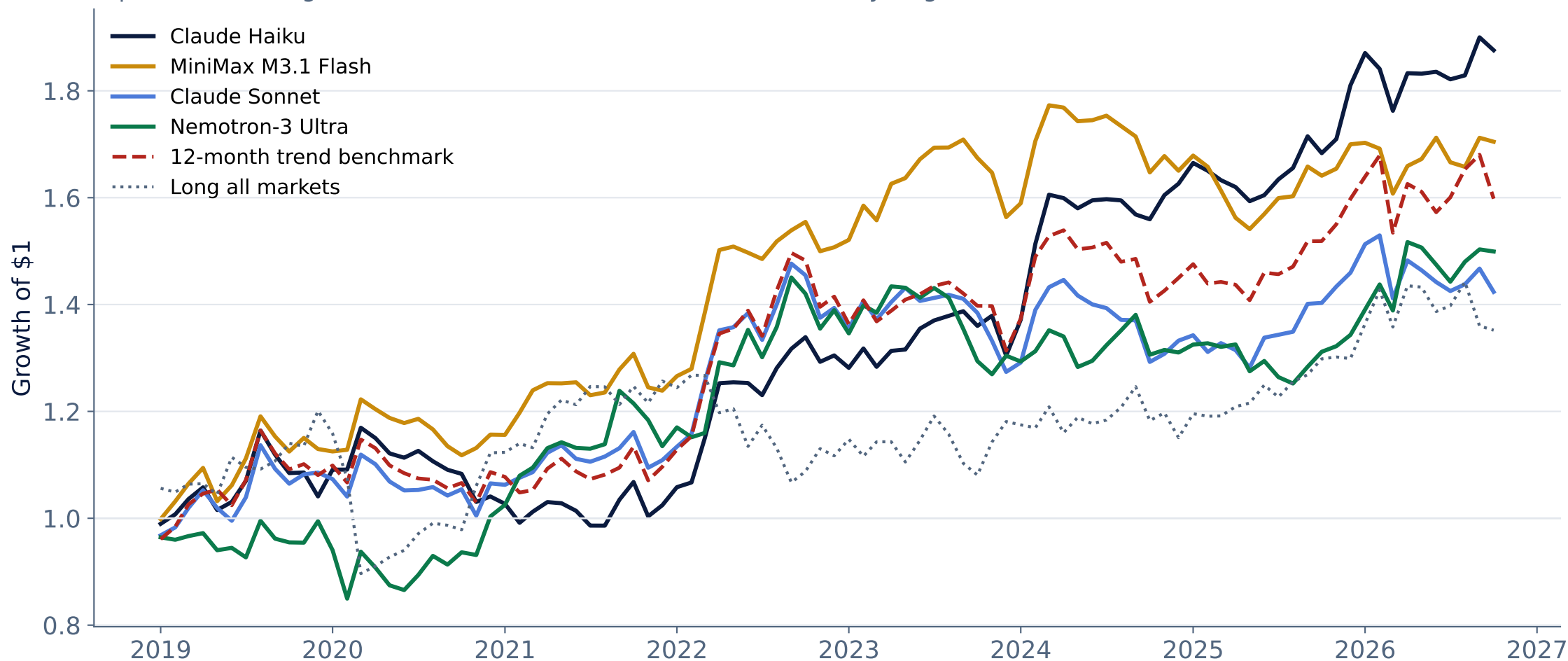
Final standings

Counted draw = each model's median out-of-sample draw. Final score = 70% performance rank + 30% blind judges.

#	Model	OOS Sharpe	IS Sharpe	OOS CAGR	Max DD	Judges	Final
1	Claude Haiku	0.75	0.36	+8.4%	-19%	7.3	0.92
2	MiniMax M3.1 Flash	0.64	0.68	+7.0%	-18%	8.0	0.88
3	Claude Sonnet	0.48	0.69	+4.9%	-18%	8.0	0.75
4	Nemotron-3 Ultra	0.48	0.05	+5.3%	-16%	5.3	0.73
5	Muse Glimmer 30B	0.36	0.47	+3.5%	-23%	7.3	0.67
6	Claude Fable	0.35	0.93	+3.1%	-19%	7.7	0.61
7	Gemma 4 31B	0.30	0.68	+2.8%	-23%	6.3	0.51
8	GLM-5.3 Flash	0.06	0.57	+0.0%	-31%	8.3	0.44
9	Laguna XS 2.1	0.06	0.07	+0.1%	-18%	5.3	0.41
10	GPT-OSS-20B	0.05	-0.42	+0.0%	-28%	4.3	0.26
11	Nemotron-3.5 Lightning	-0.04	-0.17	-1.1%	-34%	3.3	0.16
12	DeepSeek V4.1 Flash	-0.31	0.44	-3.9%	-42%	5.3	0.16
	Benchmark: 12-month trend	0.65	0.71	+6.7%	-14%		
	Benchmark: long all markets	0.39	0.44	+3.9%	-36%		

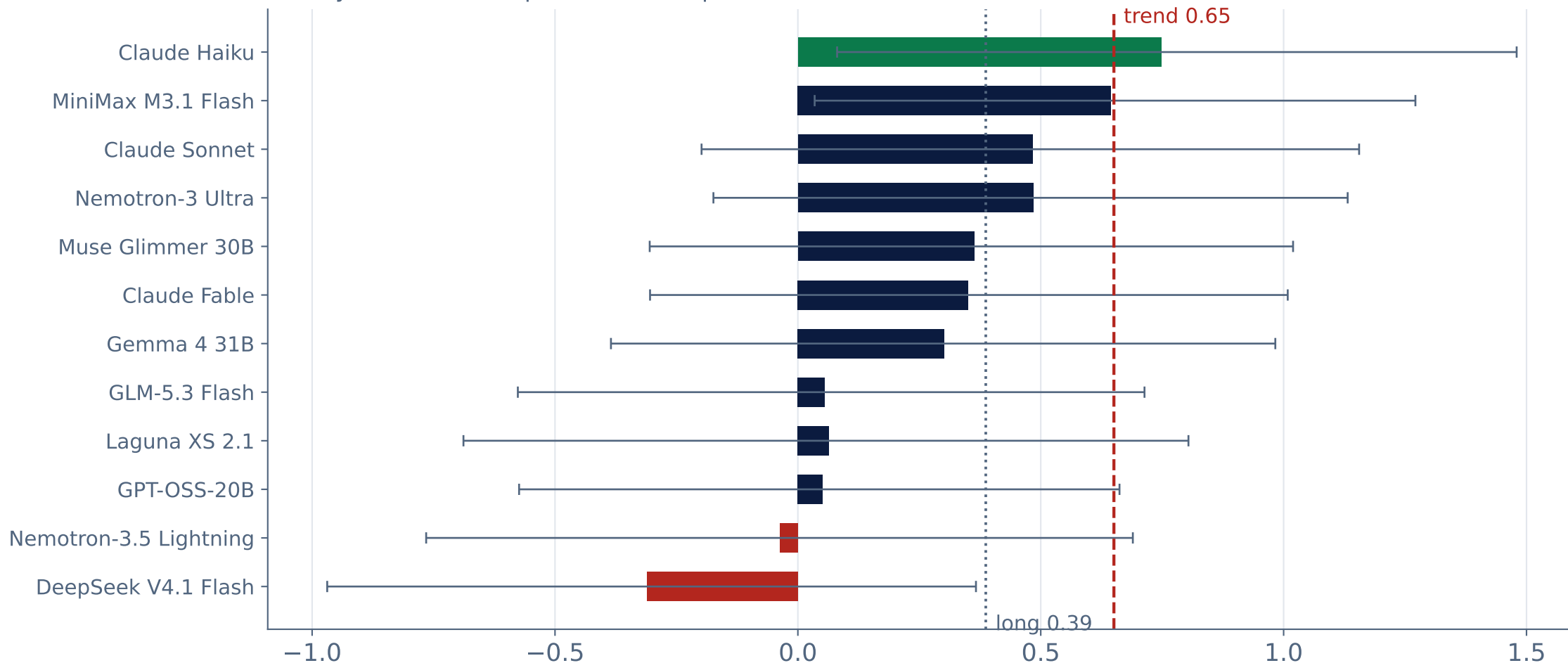
Out of sample: growth of \$1, January 2019 to September 2026

Top four models against the benchmarks, after costs, 10% volatility target



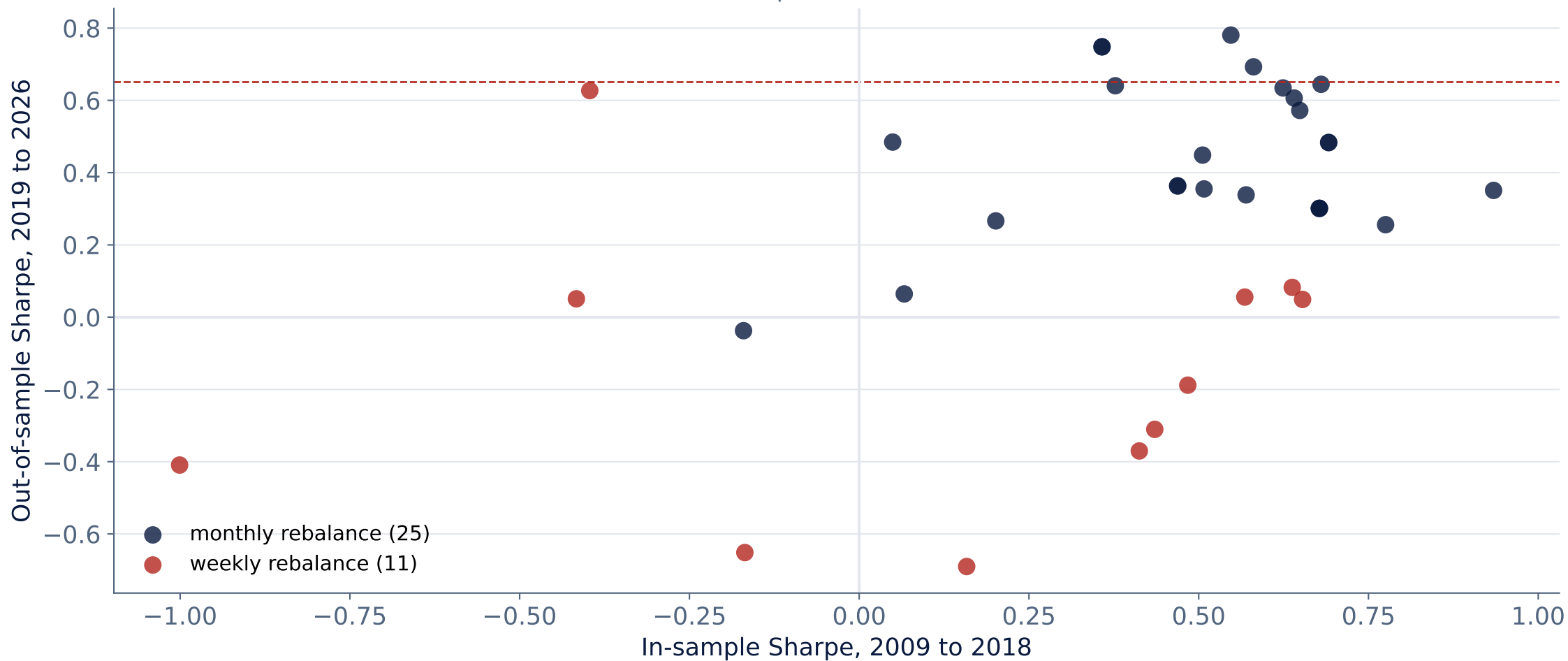
Out-of-sample Sharpe with 95% bootstrap intervals

20-day block bootstrap, 2,000 resamples. Green = above the trend benchmark



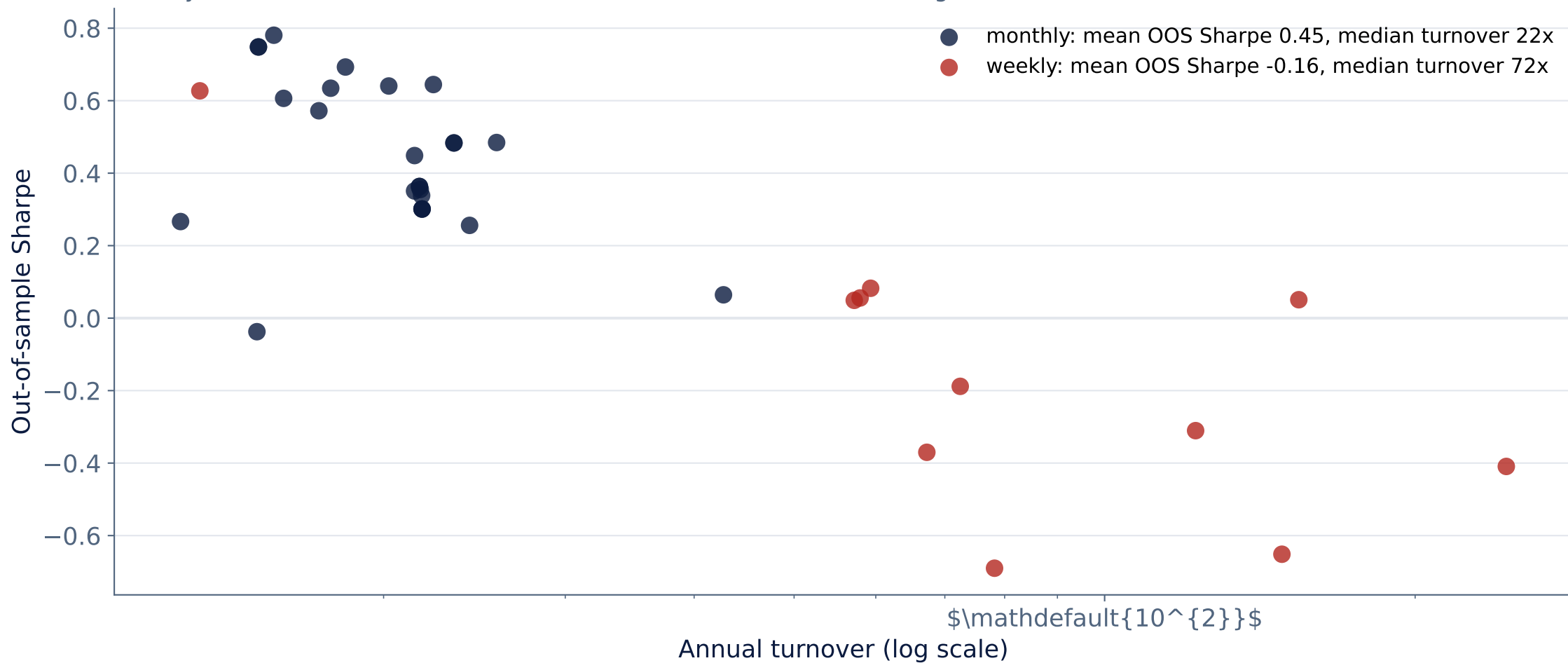
The design window barely predicted the test: rank correlation 0.26

All 36 draws. Dashed line = trend benchmark out of sample



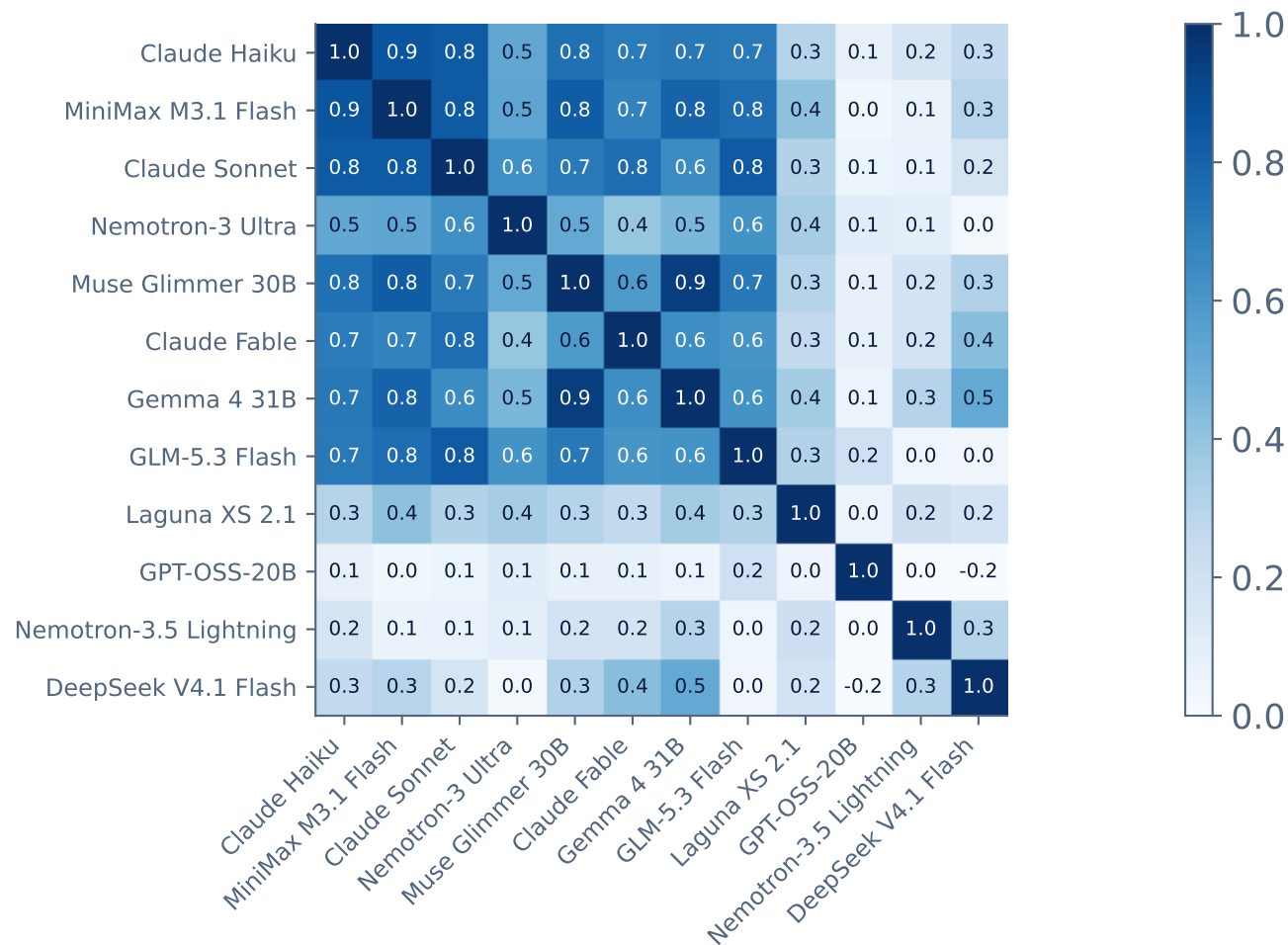
Trading more paid less: rank correlation -0.72

Weekly rebalancers traded about three times as much and lost the edge to costs and noise



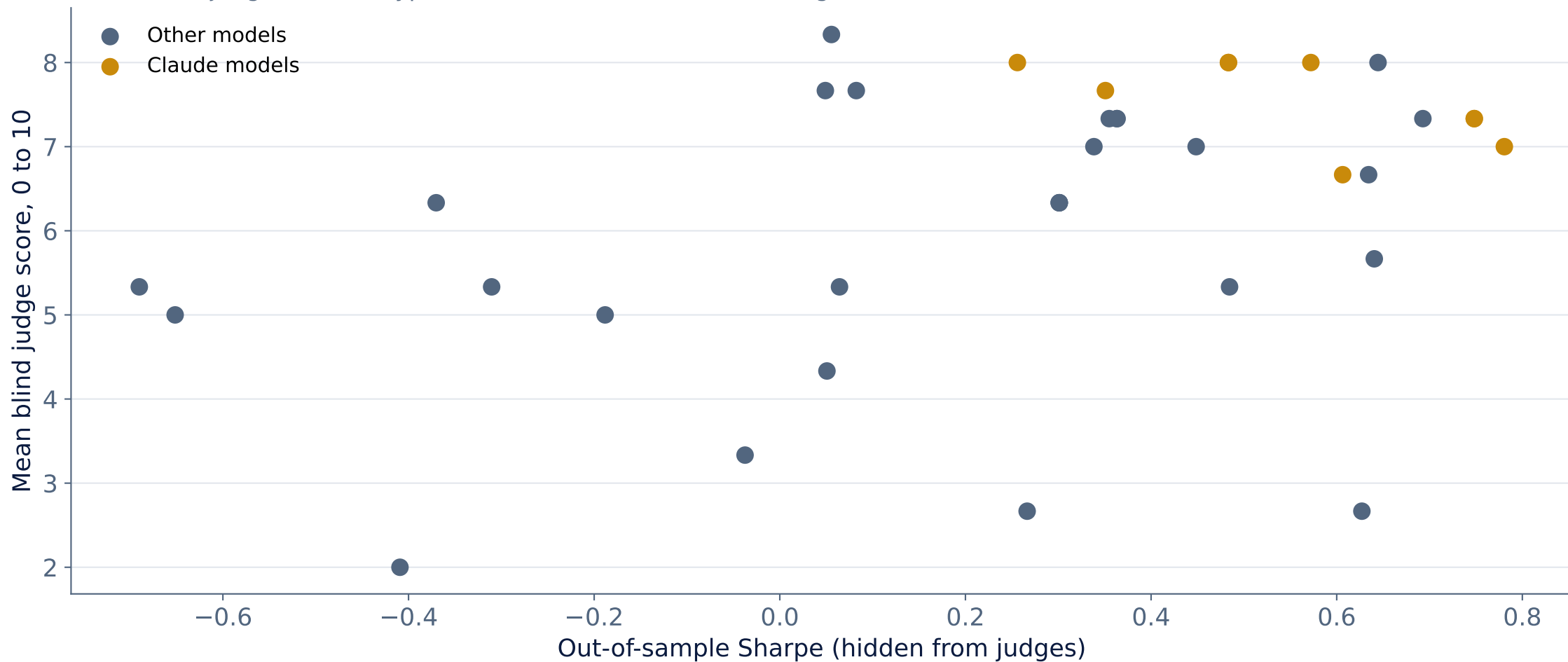
Herding: 36 of 36 strategies used carry, most added trend

Correlation of daily out-of-sample returns between counted draws



Judges tracked the in-sample record (0.75) more than the test (0.40)

Three blind judges scored hypothesis, falsification test and design. Rank correlations shown



The winner: Claude Haiku

0.75

out-of-sample Sharpe, the only entry above the 0.65 trend benchmark.
Information ratio against trend +0.23

Signal

$\text{zscore}(\text{ret}(252)/\text{vol}(252)) + 0.5 * \text{zscore}(\text{carry}())$: twelve-month trend scaled by its own volatility, plus half-weight carry, monthly, all six sectors

0.08

lower end of the 95% bootstrap interval, above zero. Deflated Sharpe probability 0.41 after 36 trials, so not yet proof of skill

0.04

average correlation of GPT-OSS-20B, the most original strategy in the field (novelty mention)

What we learned, and Round 2 changes

Anchor

Seven draws were close variants of the example signal in the brief.
Round 2 briefs give no example signal

Cost

Weekly entries averaged -0.16 out of sample, monthly 0.45. Turnover now shows in the brief and score

Judges

Claude entries scored +1.7 points higher with all three judges. Round 2 removes same-family judges and strips style cues

Repeats

3 signals were submitted twice. Duplicate draws will be redrawn so every model gets three distinct ideas