

overfit_meanrev.py

Strategy stress test report

Assets	SPY	Data period	1993-01-29 to 2026-10-02
Development period	2000-01-01 to 2012-12-31	Costs and fills	5 bps per side; filled at the next close
Generated	2026-10-09 01:05 UTC	Engine	Quantgrounds 0.2.0
Code fingerprint	067e 4b9a 8316 ff63	Replay code	1/067e4b9a8316ff63/C/SPY/2000-01-01/2012-12-31/5/10

Verdict

The edge largely disappears outside the period the strategy was developed on.

Diagnostic score

10 / 100

Serious issues

Weighted mean of six sub-scores. Capped at 10 by the edge over buy & hold (-0.87 Sharpe). Not a forecast.

Sub-score	Weight	Score	Based on
Edge over buy & hold	25%	0	out-of-sample Sharpe 0.18 vs 1.05 held
Risk-adjusted return	15%	18	out-of-sample net Sharpe 0.18
Parameter robustness	15%	52	99% of Sharpe kept (median), 4% at worst $\pm 10\%$
Cost resilience	15%	82	break-even at 5.6x your cost
Generalization	15%	70	100% of held-out assets positive, Sharpe ratio 0.41
Drawdown at equal risk	15%	21	68.0% at equal volatility vs 55.2% held

Stress tests

Stress test	Result	Status
Out-of-sample periods	Sharpe 0.95 in-sample, 0.18 out-of-sample	Fail
Walk-forward	Sharpe 0.38 over 28 one-year tests; 18 of 28 positive	Pass
Parameter sensitivity	Worst $\pm 10\%$ change keeps 4% of Sharpe (rsi_days)	Fail
Transaction costs	Break-even at 27.9 bps per side (5.6x baseline)	Pass
Execution delay	Sharpe 0.46 to 0.34 with a one-day delay	Pass
Unseen assets	Positive on 10 of 10 held-out assets; median Sharpe 0.19	Warning
Market regimes	Weakest: bear / crash, Sharpe -0.23 over 577 days	Pass
Look-ahead bias	Prevented by design: on_bar() sees past data only	Pass
Resampled histories	Sharpe 90% range 0.18 to 0.71	Pass
Random timing	Beats 14% of 1,000 random timings with the same exposure (after the development period)	Warning
Market exposure	Out-of-sample Sharpe 0.18 vs 1.05 held; full history: beta 0.07, alpha 1.4%/yr (t 1.9)	Fail
Research selection	Variants tried not given; selection risk unknown	Not run
Code review	1 note; deterministic	Warning
Drawdowns and concentration	Max drawdown -23.3%; longest underwater 10.2 yrs	Warning

Key metrics

Metric	Strategy (net of costs)	Buy & hold SPY
CAGR	2.2%	10.8%
Sharpe ratio (90% range)	0.46 (0.18 to 0.71)	0.65
Maximum drawdown	-23.3%	-55.2%
Total trades / yearly turnover	156 / 9.3x	1 / 0.0x

1. Summary and findings

Over 1993-01-29 to 2026-10-02, the strategy returned 2.2% a year net of costs, with a Sharpe ratio of 0.46 (90% resampling range 0.18 to 0.71) and a worst drawdown of -23.3%. Buy & hold of the same assets: Sharpe 0.65, CAGR 10.8%, worst drawdown -55.2%. The most important issue: performance collapses outside the development period.

Findings (7), most important first

F1 **High** Performance collapses outside the development period · line 12

Evidence. In-sample (2000-01-01 to 2012-12-31): Sharpe 0.95, CAGR 4.4%. Out-of-sample: Sharpe 0.18 (before the development period: Sharpe 0.55, CAGR 2.6%; after the development period: Sharpe -0.00, CAGR -0.1%).

Likely cause. Performance is concentrated in the period you developed on, which is the signature of fitting to that period. See the tuned-looking values at line 12.

Change to try. Re-derive the rules with fewer, rounder parameters. Then judge them only on data you did not look at, such as a walk-forward split, before trusting the in-sample numbers.

F2 **High** Parameter cliff: 'rsi_days' · line 11

Evidence. Baseline rsi_days=4 gives Sharpe 0.46. A $\pm 10\%$ nudge to rsi_days=5 gives Sharpe 0.02 (4% retained). All nudges: rsi_days=3: Sharpe 0.49; rsi_days=5: Sharpe 0.02.

Likely cause. PARAMS['rsi_days'] (line 11) sits on a narrow peak. Small changes in market behavior act like a parameter nudge.

Change to try. Pick 'rsi_days' from the middle of a plateau where neighbors perform similarly, or average the signal over several values of 'rsi_days' instead of betting on one.

F3 **High** Buy & hold of the same assets did better

Evidence. Out-of-sample Sharpe 0.18 vs 1.05 held (same assets, same days, same costs). The score is capped at 10.

Likely cause. holding the assets earned more per unit of risk than trading them

Change to try. Unless the strategy has another purpose it delivers (lower drawdowns, less time invested), holding is simpler and cheaper.

F4 **Medium** Logic doesn't transfer to assets you didn't pick

Evidence. Median Sharpe on declared assets 0.46 vs 0.19 on 10 held-out assets; 100% of held-out assets were profitable.

Likely cause. The rule may be exploiting something specific to the declared asset and period rather than a general effect.

Change to try. State the mechanism you expect to work and why it should be specific to this asset. If it shouldn't be, simplify until it works across similar assets.

F5 **Medium** Timing is no better than random

Evidence. Over the years after the development period, rotating the strategy's own positions against the market 1,000 times (same exposure, holding periods and costs; only the timing changes) gives a median Sharpe of 0.23; 95% of rotations stay below 0.65. The strategy's Sharpe of -0.04 beats 14% of them.

Likely cause. when the strategy is invested adds little; its returns come from how much it is invested

Change to try. A fixed allocation with the same average exposure would do about as well and is simpler.

F6 **Medium** Very long time underwater

Evidence. Longest stretch below a previous equity peak: 10.2 yrs.

Likely cause. Returns arrive in bursts separated by long flat or losing periods.

Change to try. Decide in advance how long you would tolerate being underwater, and compare that with this figure.

F7 Medium Unusually specific parameter values · line 12

Evidence. Line 12: "entry_rsi": 10.2,

Likely cause. Values like entry_rsi=10.2, exit_rsi=80.7, trend_days=51 are typical of optimizer output rather than a reasoned choice. Specific values often sit on a narrow performance peak.

Change to try. Prefer round values chosen from reasoning, and confirm the parameter-fragility section shows a plateau around them.

What held up

- + Cost resilient: break-even at 27.9 bps per side (5.6x baseline).
- + Walk-forward holds up: Sharpe 0.38 when parameters are re-chosen from past data only (18 of 28 test years positive).

2. Stress tests in detail

2.1 Out-of-sample periods

One continuous backtest, split by date into the development period you declared (the data you tuned on) and everything before and after it.

Period	Days	CAGR	Sharpe	Volatility	Max drawdown
Before development period	1,749	2.6%	0.55	5.0%	-7.0%
Development period (2000-01-01 to 2012-12-31)	3,269	4.4%	0.95	4.6%	-4.2%
After development period	3,459	-0.1%	-0.00	5.1%	-23.3%
All out-of-sample	5,208	0.8%	0.18	5.1%	-23.3%



Figure 1. Growth of \$1, log scale, net of 5 bps per side. Shaded: development period.

2.2 Parameter sensitivity

Each numeric parameter in PARAMS is moved by -20%, -10%, +10% and +20% while the others stay fixed (integers are rounded).
Baseline net Sharpe: 0.46.

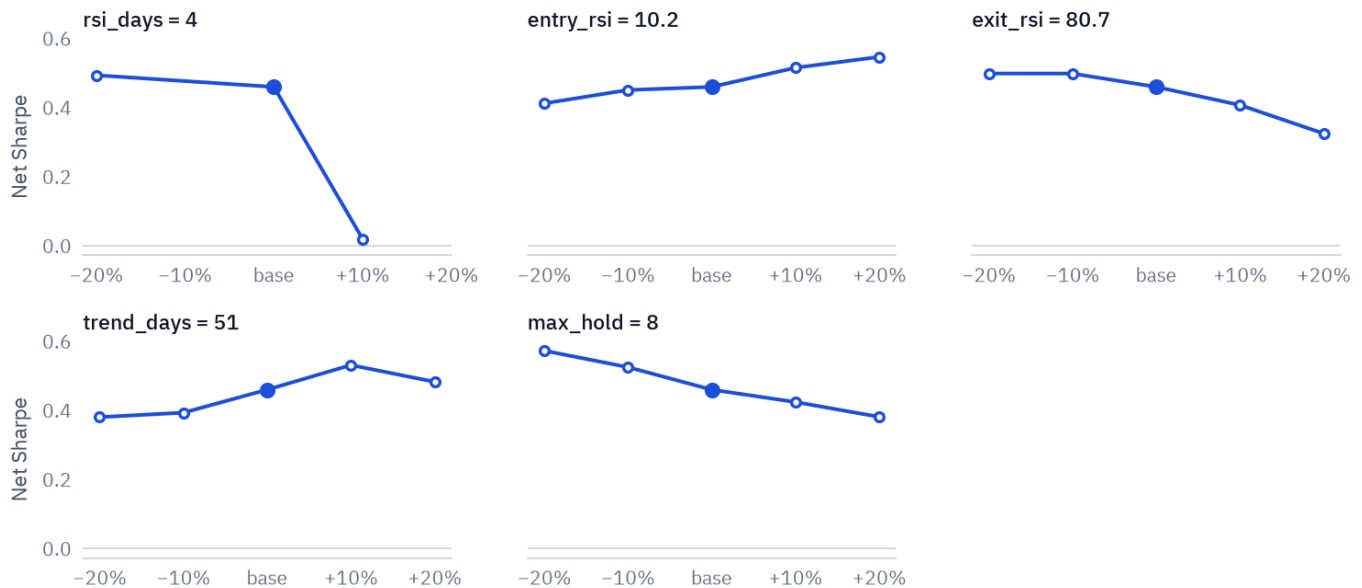


Figure 2. Net Sharpe as each parameter is moved. Filled point: the declared value.

Parameter	Change	Value	Net Sharpe	Net CAGR	Sharpe kept
rsi_days	-20%	3	0.49	3.1%	107%
rsi_days	+10%	5	0.02	-0.0%	4%
entry_rsi	-20%	8.16	0.41	1.9%	90%
entry_rsi	-10%	9.18	0.45	2.1%	98%
entry_rsi	+10%	11.22	0.51	2.5%	112%
entry_rsi	+20%	12.24	0.55	2.7%	119%
exit_rsi	-20%	64.56	0.50	2.2%	108%
exit_rsi	-10%	72.63	0.50	2.3%	108%
exit_rsi	+10%	88.77	0.41	2.0%	89%
exit_rsi	+20%	96.84	0.32	1.6%	71%
trend_days	-20%	41	0.38	1.7%	83%
trend_days	-10%	46	0.39	1.8%	85%
trend_days	+10%	56	0.53	2.6%	116%
trend_days	+20%	61	0.48	2.4%	105%
max_hold	-20%	6	0.57	2.5%	125%
max_hold	-10%	7	0.52	2.4%	114%
max_hold	+10%	9	0.42	2.1%	92%
max_hold	+20%	10	0.38	1.9%	83%

2.3 Transaction costs

The same positions re-costed at 0x to 5x the baseline cost per side, charged on traded notional. Break-even is the per-side cost at which net CAGR reaches zero.

Cost level	bps per side	Net CAGR	Net Sharpe
0x	0	2.6%	0.55
1x	5	2.2%	0.46
2x	10	1.7%	0.36
3x	15	1.2%	0.27
5x	25	0.3%	0.08

Break-even: 27.9 bps per side (5.6x baseline).

Turnover: 9.3x capital per year.

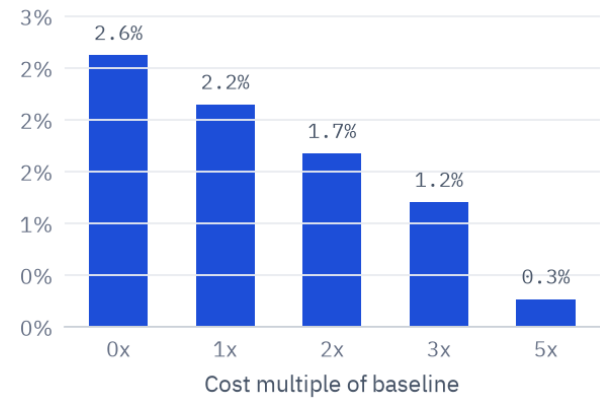


Figure 3. Net CAGR by cost level.

2.4 Execution delay

The strategy's own decisions, each filled one trading day later than in the baseline (two closes after the signal instead of one). Everything else is unchanged.

	Baseline	One-day delay
Net Sharpe	0.46	0.34
Net CAGR	2.2%	1.6%
Maximum drawdown	-23.3%	-19.5%

2.5 Unseen assets

The same code and parameters run on each asset alone, over that asset's own history, including assets you did not declare (other symbols in the data set with at least three years of history).

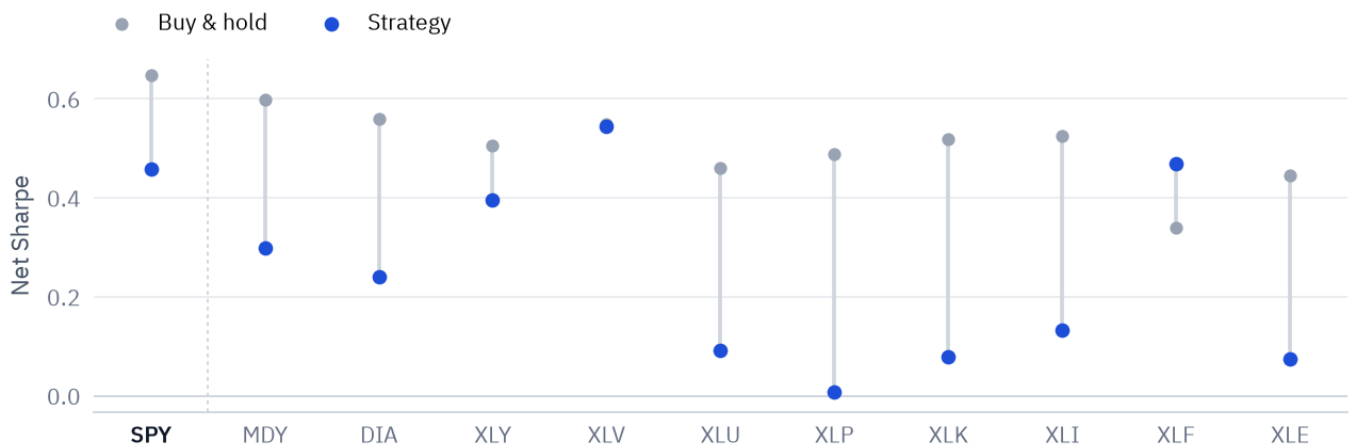


Figure 4. Net Sharpe per asset, strategy vs buy & hold. Bold: declared assets.

Asset	Role	Period	Strategy Sharpe	Buy & hold Sharpe	Strategy CAGR
SPY	declared	1993–2026	0.46	0.65	2.2%
MDY	held-out	1995–2026	0.30	0.60	1.7%
DIA	held-out	1998–2026	0.24	0.56	1.1%

Asset	Role	Period	Strategy Sharpe	Buy & hold Sharpe	Strategy CAGR
XLY	held-out	1998–2026	0.40	0.51	2.5%
XLV	held-out	1998–2026	0.55	0.55	2.8%
XLU	held-out	1998–2026	0.09	0.46	0.4%
XLP	held-out	1998–2026	0.01	0.49	-0.0%
XLK	held-out	1998–2026	0.08	0.52	0.3%
XLI	held-out	1998–2026	0.13	0.53	0.6%
XLF	held-out	1998–2026	0.47	0.34	2.8%
XLE	held-out	1998–2026	0.07	0.44	0.3%

2.6 Market regimes

Each day is labeled from trailing SPY data (first matching rule wins): Bear / crash: SPY trailing 126-day return below -10%; High volatility: not bear, and SPY trailing 21-day realized volatility above 25% annualized; Bull: not the above, and SPY trailing 126-day return above +10%; Sideways: everything else (SPY trailing 126-day return between -10% and +10%, volatility normal). Returns are arithmetic annualized means over non-contiguous days.

Regime	Days	Strategy return	Strategy Sharpe	SPY return
Bull	2,811	4.8%	1.07	29.8%
Sideways	4,417	0.8%	0.16	5.1%
High volatility	546	3.8%	0.87	69.0%
Bear / crash	577	-0.9%	-0.23	-74.6%

2.7 Look-ahead bias

The strategy has no `prepare()` step. `on_bar()` only receives data up to the current bar, so it cannot read future prices by construction. Static checks still scan the code for future-reading patterns (section 4).

2.8 Resampled histories

The strategy's daily net returns were resampled 1,000 times in 20-day blocks (circular block bootstrap), which keeps short-range patterns and never treats overlapping days as independent. Ranges assume the future resembles this history; they are not forecasts.

Statistic	Actual	90% range
Sharpe ratio	0.46	0.18 to 0.71
CAGR	2.2%	0.8% to 3.3%
Maximum drawdown	-23.3%	-23.8% to -9.1%
Share of histories with negative CAGR		0.3%

2.9 Random timing

Over the years after the development period, the strategy's own positions were rotated against the market by 1,000 different offsets. Each rotation keeps the exposure, holding periods and trading costs and only changes when the strategy is invested (a Monte Carlo permutation test, after Masters and Aronson). Good timing should beat most rotations.

Statistic	Value
Strategy Sharpe	-0.04
Random timings: median Sharpe	0.23
Random timings: 5th to 95th percentile	-0.15 to 0.65
Share of random timings beaten	14%

Statistic	Value
Average exposure	10%

2.10 Market exposure

How much of the return is the market's? Daily net returns are regressed on SPY (Jensen's alpha), with Newey-West standard errors because overlapping positions make daily errors correlated. The edge over buy & hold, which caps the score, is shown alongside.

Statistic	Value
Sharpe edge over buy & hold	-0.87 (out-of-sample Sharpe 0.18 vs 1.05 held)
Beta to SPY	0.07
Alpha per year	1.4% (t = 1.9)
Share of daily variation explained by SPY	7%

2.11 Walk-forward

Each year, the best of the 19 parameter sets that were run (the declared values and each $\pm 10\%$ / $\pm 20\%$ change) is chosen by its Sharpe over the previous 5 years, then traded for the next year. Every choice uses only past data. This re-tunes within a small neighborhood of your parameters, so it can't see a wider search you did before uploading.

Statistic	Value
Walk-forward Sharpe (stitched test years)	0.38
Full backtest Sharpe	0.46
Test years with a positive Sharpe	18 of 28
Median / worst test-year Sharpe	0.22 / -1.12
Years the declared values were chosen	0%

2.12 Research selection

Not run: the number of variants you tried before this one wasn't given, so the risk that this result is the luck of picking the best of many backtests is unknown. Give it ("Variants tried" under Your setup in the demo, --trials on the command line) to run this test.

3. Performance and risk

Metric	Strategy (net)	Buy & hold SPY
Total return	104.8%	3075.5%
CAGR	2.2%	10.8%
Sharpe ratio (rf = 0)	0.46	0.65
Sortino ratio	0.66	0.92
Annualized volatility	4.9%	18.5%
Maximum drawdown	-23.3%	-55.2%
CAGR before costs	2.6%	
Trades	156	
Win rate per trade	71%	
Average holding period (days)	5.7	
Turnover (x capital per year)	9.3	
Time in market	10%	100%



Figure 5. Drawdown from the previous equity peak. Shaded: development period.

Largest drawdowns

Depth	Peak	Trough	Recovered	Decline	Recovery	Regime at the time
-23.3%	2016-06-23	2023-03-01	not yet	1682 d	n/a	Sideways
-7.0%	1998-07-27	1998-11-13	1999-03-04	78 d	74 d	High volatility
-7.0%	1999-07-28	2000-04-14	2001-06-05	182 d	286 d	Sideways
-5.9%	1997-03-10	1998-01-27	1998-04-30	223 d	65 d	Bull
-5.3%	1996-02-22	1996-12-16	1997-03-05	207 d	54 d	Bull

The five deepest of 80 drawdowns, in trading days. Longest time below a previous peak: 10.2 yrs.

Return concentration

Across 156 trades, the best trade produced 7% of net trade P&L, the best five 29%, and the best 10% (16 trades) 73%. Shares above 100% mean the remaining trades lost money in aggregate.

4. Code review

Line	Severity	Note	Code
12	Medium	Unusually specific parameter values	"entry_rsi": 10.2,

Guaranteed by the engine

- on_bar() only receives prices up to and including the current bar, as read-only arrays.
- Fills: decision at close t, filled at close t+1; a decision never earns the return of the bar it was made on.
- Costs of 5 bps per side are charged on every change in position.
- Gross exposure is capped at 1x (0 decisions were scaled down).
- Determinism: a second identical run produced identical positions.
- Execution: 30 runs in a separate process with a 600 s time limit, a 4096 MB memory limit, file writes blocked, and network access blocked (isolated network namespace).

5. Methodology and limitations

Data

Adjusted daily closes from the cached data set (see data/SOURCES.md). This run: 1993-01-29 to 2026-10-02.

Execution model

Event-driven, one bar at a time. The strategy returns target weights; the engine recomputes holdings, drift, turnover, costs and P&L itself, so strategy code never touches the accounting.

Statistics

Daily simple returns net of costs; 252 trading days per year; Sharpe and Sortino with a zero risk-free rate. Trades open when an asset's weight leaves zero and close when it returns to zero or changes sign. The only probability figure is the deflated Sharpe ratio, when the number of variants tried is given.

Scores

Random-shard mode (the default) gives the Quantgrounds score; "Your setup" mode gives a diagnostic score for that setup only. Both are a weighted mean of six sub-scores, each 0 to 100 with linear interpolation between anchors, judged against buy & hold of the same assets over the same days. Edge over buy & hold (25%): out-of-sample Sharpe minus buy & hold's, $-0.5 \rightarrow 0$, $0 \rightarrow 40$, $+0.25 \rightarrow 65$, $+0.5 \rightarrow 85$, $+1.0 \rightarrow 100$. Risk-adjusted return (15%): net Sharpe $0 \rightarrow 0$, $0.5 \rightarrow 50$, $1.0 \rightarrow 80$, $1.5 \rightarrow 100$. Parameter robustness (15%): $100 \times (\text{half the median Sharpe kept under nudges} + \text{half the worst kept at } \pm 10\%)$. Cost resilience (15%): $\text{break-even} \div \text{your cost } 1x \rightarrow 0$, $2x \rightarrow 40$, $3x \rightarrow 60$, $5x \rightarrow 80$, $10x \rightarrow 100$. Generalization (15%): $100 \times (\text{half the share of held-out assets with positive Sharpe} + \text{half the held-out} \div \text{declared median Sharpe, capped at } 1)$. Drawdown at equal risk (15%): $\text{worst drawdown after scaling to buy \& hold's volatility} \div \text{buy \& hold's worst drawdown}$, $0.3 \rightarrow 100$, $0.7 \rightarrow 70$, $1.0 \rightarrow 40$, $1.5 \rightarrow 0$. Missing sub-scores are dropped and weights re-normalized. Edge cap: the score can't exceed $50 + 100 \times \text{the Sharpe edge}$ (never below 10). Confirmed look-ahead caps it at 20. Bands: 90+ Exceptional, 80–89 Strong, 65–79 Promising, 45–64 Mixed results, 25–44 Weak, below 25 Serious issues. Scores are shown as whole numbers with halves rounded up.

Stress-test statuses

Each test's status is the most severe finding attached to it: critical or high $\rightarrow$ Fail, medium $\rightarrow$ Warning, low $\rightarrow$ Note, none $\rightarrow$ Pass. If look-ahead is confirmed, every performance-based test is marked Not valid.

Random timing

The strategy's positions are rotated against the market by 1,000 offsets of at least 63 days, keeping exposure, holding periods and costs. Reported: the share of rotations whose Sharpe is below the strategy's. Uses the years after the development period when one is declared. With no skill the share is uniform on 0-100%, so: high finding below 5%, medium below 50%, low below 95%. Not run when exposure never changes.

Market exposure

OLS of daily net returns on the benchmark (SPY for your setup, the whole US market for random shards) with Newey-West standard errors: beta, annualized alpha and its t-statistic. Medium finding when alpha is at most -1% a year with $t \leq -2$; low when beta is at least 0.3 and $|t| < 1$. The edge over buy & hold findings are shown under this test.

Walk-forward

Each year, the best of the parameter sets already run (declared values and each $\pm 10\%/\pm 20\%$ change) by Sharpe over the previous five years is traded for the next year; the test years are stitched together. High finding if the full Sharpe is above 0.3 and the walk-forward Sharpe is at or below zero; medium if it is below half, or if fewer than half the test years are positive.

Research selection

Deflated Sharpe ratio (Bailey and López de Prado, 2014) for the number of variants the user says they tried: the chance the true Sharpe is above zero after picking the best of that many. High finding below 50%, medium below 95%. Runs only when the number is given; it does not change the score.

Random shards

The default mode. Shards are random (asset, window) pairs: one asset from the data set with at least three years of history, a window of 2 to 8 years starting after the strategy's first trade on that asset. Each shard is scored with the six sub-score formulas above, using slices of full-history backtests (no development period; generalization compares the other assets over the same dates). The Quantgrounds score applies the usual weights to each sub-score's average over all shards. The $\pm$ figure is 1.96 standard deviations of that score across 300 resamples of the shards; the seed reproduces the run exactly. The shard test fails below 50% of shards with a positive Sharpe and warns below 70%.

Limitations

Daily bars and US equities only; close prices only, so fills are at the next close; a flat cost per trade with no market impact, borrow fees or taxes; the demo data set is small, ends in 2018–2019 and has survivorship bias; regimes use simple SPY rules; static code checks are heuristics (the truncation test is the proof). Passing every test does not show that a strategy has an edge.

Verify this report

The code fingerprint is the SHA-256 hash of the strategy source (line endings normalized, trailing spaces and blank edge lines ignored), so it changes if the code changes. The replay code holds its first 16 characters, every setting that affects the score (the random-shard seed, or the assets, development period and cost) and the score itself. Runs are deterministic, so at quantgrounds.com/verify anyone can paste the code and the replay code: the fingerprint is checked in the browser, then the strategy is re-run with those settings. The same score means the report is genuine for this code; an edited score or a different strategy does not reproduce. No key or account is involved.

Fingerprint: 067e4b9a8316ff63f1e2c51a626837ed372aca7066742ead5f5975a99277bf6

Replay code: 1/067e4b9a8316ff63/C/SPY/2000-01-01/2012-12-31/5/10

Educational diagnostics only; not investment advice. Results are simulated on historical data and do not predict future returns. The score summarizes this report's tests; it is not a measure of trading skill or future profitability.

Every number in this report is read from the run's results.json file.