

Why Most AI Trading Claims Fail on Inspection

CHEAT SHEET

A backtest that looks perfect is the strongest evidence something has gone wrong — flawless numbers are the fingerprint of overfitting, not genius.

CAST

Richard Feynman

physicist who warned against self-deception

John von Neumann

mathematician who illustrated overfitting's limits

Tarun Chordia

Emory researcher who quantified p-hacking at scale

Amit Goyal

Lausanne co-researcher on 2.1M strategy study

Campbell Harvey

documented p-hacking across asset-pricing literature

David Bailey

named trial omission a form of fraud in 2014

Norgate Data

provider of point-in-time index membership data

Overfitting

core failure mode memorizing noise not markets

P-Hacking

running tests until significance appears accidentally

THE STORY

1

Perfect backtest metrics signal catastrophic structural failure

300% returns, Sharpe above 3, drawdown under 5%, win rate above 75% — all those numbers together at those magnitudes are a warning, not validation. Every one of the failure modes ahead originates in normal, intelligent, well-intentioned behavior.

2

Overfitting memorizes historical noise not market structure

A strategy tuned too closely to past data stops describing how markets work and starts describing random noise in that specific dataset. Von Neumann's point: given enough adjustable parameters, any model fits any dataset perfectly — fitting is not understanding.

3

Data traps corrupt backtests before testing even begins

Testing a low-volatility strategy on today's S&P 500 constituents quietly removes all the high-volatility firms that blew up — the most dangerous outcomes are laundered out before you start. Point-in-time data like Norgate's reconstructs actual historical membership but costs money most retail platforms don't spend.

4

Multiple testing compounds false positives to near certainty

Five outcome measures across four conditions at 5% significance yields a 64% chance of at least one spurious result. Fifty independent tests at 5% significance produces a 92.3% probability of at least one false positive — the backtest reports the winner with the same straight face it always wears.

5

2.1 million strategies prove most discoveries are statistical accidents

Chordia, Goyal, and Saretto built the largest controlled p-hacking laboratory in finance history and found that under standard testing, most rejections of no outperformance were false positives. After applying multiple-hypothesis corrections, only a tiny fraction of strategies survived — none with theoretical grounding.

6

Omitting failed trials is functionally a form of fraud

David Bailey argued in 2014 that presenting a winning backtest without disclosing prior trial count permits recipients to draw statistically unsupported conclusions. Campbell Harvey showed published asset-pricing literature is largely a collection of selection artifacts from exactly this practice.

CONNECTIONS



REMEMBER

- ◆ A backtest needs **30 independent trades per free parameter** to carry any statistical weight — a standard widely known and routinely ignored.
- ◆ Testing 45 strategy parameter variations across five years of daily data produces a **greater than 50% probability of overfitting**.
- ◆ Running 50 independent tests at a standard 5% significance threshold gives a **92.3%** chance of generating at least one spuriously significant result.
- ◆ Just five outcome measures and four treatment conditions tested at 5% significance produces a **64%** chance of at least one false positive.
- ◆ Chordia, Goyal, and Saretto generated approximately **2.1 million trading strategies** — the largest controlled p-hacking laboratory ever assembled in finance.
- ◆ David Bailey wrote in 2014 that failing to report the number of trials behind a successful backtest is a **similar kind of fraud**.
- ◆ More than half of all backtests show signs of overfitting when enough parameter combinations are tested without statistical adjustment.