THANAAYA TECHNOLOGIES

RESEARCH NOTE

Counting Honestly

An AI model banked 488 crypto trading signals. Counted properly, none held up. Eight signals named in advance then met three months of new data.

ABOUT THANAAYA TECHNOLOGIES

Thanaaya Technologies is a software company licensed by the Dubai Department of Economy and Tourism, building since 2021. We design, build and operate production systems across financial software, distributed ledger, gaming, commerce, media and quantitative research — and we run two of them as our own products rather than only delivering them for clients.

We take a small number of engagements at a time, and the engineer who designed a system is the one who gets called when it misbehaves. This paper comes out of our own research programme, which exists to find out cheaply whether an idea is real. We publish the negative results as well as the positive ones: an undocumented failure gets rediscovered, and a threshold nobody wrote down cannot be checked.

2021	24	80+	9
BUILDING SINCE	PRODUCTION DOMAINS	SERVICES IN PRODUCTION	INDUSTRIES SERVED

Subject: AI-assisted signal search on Binance USDT-margined perpetual futures, trial counting and a pre-registered forward test
Programme: Thanaaya internal quantitative research
Status: search closed; no capital committed
Published: November 2026 · **Licence:** Dubai Department of Economy and Tourism 964356 · **Unified licence:** BL1484

PART ONE

In plain English

What we asked, what we built, how we caught our own mistake, and what a three-month forward test could and could not tell us. No mathematics in this section.

1. The question

If you let an AI model search crypto futures for a trading edge, how do you know whether anything it finds is real?

Our own research programme set out to answer that on Binance's USDT-margined perpetual futures. A model proposed trading signals in bulk. A strict statistical test banked the survivors. When the search stopped, 488 signals were in the bank and every one had passed.

That is the moment a search is most dangerous. Search long enough and some ideas will look excellent on past data by luck alone. The only defence is to count honestly how many ideas were tried, and then ask for a correspondingly better result. We had a test built for exactly that. The question this paper answers is whether we had used it honestly.

2. The mistake we found in our own count

We had not. The search ran in rounds, and the tally of ideas tried restarted at the start of every round. Each round was therefore judged against an easier bar than the whole search deserved. Later rounds were also steered by what earlier rounds had rejected, so the ideas were not independent attempts.

~1,930
genuinely different ideas tried, counted across the whole search

0.95 → 0.77
the best signal's score once re-priced at the honest count, against a floor of 0.80

0 of 81
strictest-gate signals that still cleared the floor

We re-priced all 488 at the honest count. The best fell from 0.95 to 0.77 on the scale, below the 0.80 floor we had set before the search began, and none of the 81 that had passed the strictest gate survived. We stopped the search.

Finding this was the first thing the programme got right. The error was in our own bookkeeping, nobody had asked about it, and it would have been invisible in any single result. It was found because we went looking for it, before any money was involved.

3. Two more checks before anything was believed

A paper-trading soak. Four of the top signals ran against live order-book prices for 51 hours. Every one of the 114 fills was priced from the real order book rather than an assumed fee. The four books together finished down 0.016%, which is what signals that no longer qualify should do. The useful result was the cost: real order-book pricing came in **80% below** a flat 0.075% assumption. That makes every later cost estimate in the programme better, whatever the signals did.

A forward test, named in advance. Before looking at any fresh data we wrote down eight signals, the best from each of eight families of ideas, and the pass mark: a t-statistic of at least 1.645 over about three months of new hourly data. Picking the winners out of 488 afterwards would have repeated the original mistake. The test was judged once, on a fixed date.

4. What came back

0 of 8	8 of 8	1.43 vs 1.645
signals that passed on three months of new data	earned a positive return on the new data	the best t-statistic against the bar it needed

None passed. But the shape of the failure matters, and it is not the shape of noise. All eight earned a positive return on data that did not exist when they were chosen, and seven did so at an annualised Sharpe ratio above 1.3. The best reached a t-statistic of 1.43 and the weakest was essentially flat (0.22).

The bar is demanding for a window this short. Three months of hourly data needs an annualised Sharpe near **3.3** to reach a t-statistic of 1.645. The best signal's realised 2.9 was a strong result that still fell short. That is a statement about the size of the sample as much as about the signal.

> WHAT WE CAN AND CANNOT SAY
>
> **Cannot say:** that any of the eight is a proven edge. The pre-registered test was not passed, and the result stands as a failure to pass. **Can say:** that eight signals picked blind came back all positive, which is what a faint real effect would look like, and also what a few correlated bets on one data field would look like. The bank is not 488 independent ideas: 420 of the 488 read the exchange's funding rate. The forward test cannot tell those two readings apart, and a longer window is what would.

5. What the programme leaves behind

- **A counting rule that does not reset.** Results are now judged against the all-time tally of ideas tried. A later round does not get an easier bar.
- **Real costs.** Order-book pricing replaced the flat fee assumption and cut the estimated cost of the trades tested by 80%.

- **A forward-test protocol.** Signals named in advance, a pass mark written before the data exists, one verdict on a fixed date. The same discipline carried into the next study.
- **A concentration warning.** 420 of the 488 banked signals read the same field. A search that finds one idea many ways has found one idea. New work starts from a different data family.
- **A clean record.** No capital was committed. The programme stays on research and paper trading, and the result is written down where it will be found.

PART TWO

The technical record

Data, search, counting, the soak, the forward test and the arithmetic of its power. Figures are as measured at the time of writing.

1. Data and search

Venue and contracts	Binance USDT-margined perpetual futures, about 490 coins, hourly bars over 18 months. The tradable set is re-chosen at each step using only what was known at the time.
Search	An AI model proposed signal formulas in bulk. A formula was banked only if it cleared a deflated Sharpe test (Bailey and López de Prado, 2014) at a floor fixed before the search began.
Output	488 banked signals.
Concentration	420 of the 488 read the exchange funding rate. Of the eight forward-tested signals, all eight did.

2. Counting the trials

The deflated Sharpe ratio adjusts a result for how many independent ideas were tried. Its honesty depends entirely on that count. We found two faults. First, the count was **reset** at the start of each search round, so a late-round result was deflated for only that round's trials. Second, later rounds were conditioned on earlier rejections, so the trials were not independent draws. Clustering ideas by the correlation of their return histories and counting across the whole life of the search gave about **1,930** distinct ideas, against the much smaller per-round counts that had been used (Harvey, Liu and Zhu, 2016, make the general point that thresholds must rise with the number of tests).

Re-pricing the bank at the honest count.

QUANTITY	AT BANKING	AT HONEST COUNT
Best deflated score in the bank	0.952	0.766
Floor	0.80	0.80
Strictest-gate signals above the floor	**81**	**0**

3. The paper-trading soak

Four of the top signals, 51.1 hours, paper only. The books held $400,000 of paper capital in total, and the point of the run was instrumentation. Every fill was priced from real order-book depth, and none fell back to the flat assumption.

MEASURE	VALUE
Fills	114
Closed trades	53
Aggregate result	−0.016%
Cost paid vs a flat 0.075% assumption	80% lower

The aggregate drifted from positive to zero during the final day, and one book travelled from clearly positive to negative without any change in market regime. With 53 closed trades across four books, none of the individual results means anything, and none was read as a survivor.

4. The forward test

Eight signals were fixed before the data existed: the best by deflated score at the honest count from each of the eight highest-ranking families. The pass mark was a t-statistic of at least 1.645 and an annualised Sharpe above 0.5 on at least 2,160 hourly bars of new data, judged once. Signals are shown by letter, in registration order.

SIGNAL	OUT-OF-SAMPLE SHARPE AT BANKING	DEFLATED SCORE AT BANKING	DEFLATED SCORE, HONEST COUNT	FORWARD SHARPE	FORWARD T
A	2.57	0.952	0.766	1.94	0.96
B	**2.66**	**0.948**	**0.756**	**2.88**	**1.43**
C	3.48	0.938	0.726	2.15	1.07
D	3.12	0.929	0.704	1.35	0.67
E	3.09	0.922	0.686	1.41	0.70
F	2.08	0.914	0.666	0.45	0.22
G	3.40	0.907	0.651	2.57	1.28
H	3.53	0.906	0.648	2.21	1.10

Each signal’s row is the first evaluation at or after the 2,160-bar mark (2,161 to 2,179 bars). The forward Sharpe is lower than the banking Sharpe for seven of eight, as selection from a large search predicts. The relation t ≈ Sharpe × √(years) holds: 2.88 × √(2,161 / 8,760) = 1.43.

5. The arithmetic of power

The bar was set before the data existed and is not criticised here. It does, however, determine what the test could detect.

QUESTION	ANSWER
Annualised Sharpe needed for t = 1.645 on 2,160 hourly bars (0.247 years)	≈ 3.3
Best realised forward Sharpe	2.88
Expected t if the true Sharpe were 1.5, over 0.247 years	≈ 0.75
Time for the expected t to reach 1.645 at a true Sharpe of 1.5	≈ 1.2 years
Time for an 80% chance of clearing 1.645 at a true Sharpe of 1.5	≈ 2.8 years

A signal with a genuine Sharpe near 1.5 would usually fail this test. The three-month window could confirm only an unusually strong effect or an unusually lucky one. The test's purpose, to stop an unsupported signal from going near capital, was met. Its power to confirm a modest real one was low by construction.

6. Across the whole bank

Banked signals evaluated on new data	488
Cleared t ≥ 1.645	10
Expected by chance alone at 5%	≈ 24
Of the ten, named in advance	0
Positive return on new data	326
Average t-statistic across the bank	+0.28

Fewer than the chance expectation cleared the bar, and none had been named in advance, so the ten carry no weight. That two-thirds of the bank made a positive return is a weak signal too: with 420 signals sharing one input, the bank behaves like a handful of bets rather than 488.

7. Limits

- Eight signals from a bank dominated by one input are not eight independent tests. Their agreement in sign is weaker evidence than it looks.
- About three months is one market regime. Funding-driven signals can behave very differently in other regimes.
- The deflated-score floor is a convention. The conclusion at the honest count (no signal clears it) does not depend on the exact floor, since the best sits at 0.77 against 0.80, but it would change at a lower one.

- The verdict is one reading on a fixed date. Later readings move (signal B’s t-statistic is lower days after the verdict), which is why only the registered reading is reported as the result.

8. What stays closed

The historical search is closed, and the banked signals are not promoted. The forward record can be extended and re-read under a new registration written in advance. Nothing here is a trading recommendation.

References

Bailey, D. H. and López de Prado, M. (2014). The Deflated Sharpe Ratio: Correcting for Selection Bias, Backtest Overfitting, and Non-Normality. Journal of Portfolio Management 40(5), 94–107. https://doi.org/10.3905/jpm.2014.40.5.094

Harvey, C. R., Liu, Y. and Zhu, H. (2016). … and the Cross-Section of Expected Returns. Review of Financial Studies 29(1), 5–68. https://doi.org/10.1093/rfs/hhv059

Scope. Thanaaya Technologies is a Dubai Department of Economy and Tourism licensed company (licence 964356). This paper describes our own research programme, carried out by the company itself; no client is involved. No strategy or signal design is published, and nothing here constitutes investment advice or an offer of any service relating to trading. Enquiries: info@thanaaya.com · thanaaya.com