OrderFlowAi

SCIENTIFIC RESEARCH REPORT · PUBLIC EDITION

The Search for the Edge

How to prove a statistical advantage in financial markets — and why almost nobody tries

A report on 471 working sessions, 54 self-retracted findings, and the construction of an evidence apparatus designed to make a result **certifiable** — before that result even exists. From the first alpha build to **CORTEX-ONE**.

471	54	350	0
WORKING SESSIONS	SELF-RETRACTED FINDINGS	SELF-TESTING TOOLS	UNSUBSTANTIATED EDGE CLAIMS

OrderFlowAi — OrderFlowAi Research
orderflowai.io/research

As of September 23, 2026 · Session 471
Public edition — sharing welcome

TABLE OF CONTENTS

What this report contains

How to read this. This report does **not** claim that an edge has been found. It documents how systematically it is being searched for — and why that systematic approach is the actual asset. Every number in it is measured; every refuted number is marked as refuted. If you read only one page, read page 3. p = probability of chance (lower is better, threshold 0.05) · n = sample size.

Shortcuts: Investors · pp. 3, 24–33, 36–43 | Engineers · pp. 13–22 | Auditors · pp. 15–19, 34–35 | Curious readers · pp. 4–14

Where the numbers come from

SOURCE	WHAT IS DRAWN FROM IT
Project archive 925 dated commits	Chronicle, milestones, phase assignment — every entry dated.
Retraction register 54 entries, machine-checked	Every statement marked false, with its cause and its cost.
Analysis tools 350 self-testing programs, 4,829 unit checks	All statistical figures — p-values, confidence intervals, decay curves.
Trading logs **598 trades / 57 days (June–September)**	**All performance figures. Simulation account, fully recorded.**

SUMMARY FOR THE IMPATIENT

We have not found the edge. We have built something rarer: the machine that could prove one.

In the trading software industry, hundreds of vendors claim an "advantage." Almost none can substantiate it — because doing so requires an apparatus almost nobody builds: pre-registration, placebo controls, null hypotheses, a register of one's own errors. **That apparatus has existed in full at OrderFlowAi since July 2026.** It is the product of twelve months of work and — this is the unusual part — of 54 findings we retracted ourselves rather than sold.

THE PRODUCT

Four software editions run in production, with paying customers and an official NinjaTrader listing. **Revenue does not depend on the edge** — it depends on visualization, order book analysis and journaling. That decoupling is deliberate.

THE RESEARCH

CORTEX — a self-learning neural engine with 49 features and five prediction heads, with a fresh learned state since September 23. Last repeatedly verified finding: **no demonstrable advantage**, globally and within every individual market regime. That is not a setback. That is a measurement.

THE ASSET

The evidence apparatus. **350 analysis tools** with built-in self-tests (4,829 individual checks), a pre-registered nine-stage test chain and an openly maintained retraction register. Whoever finds an edge with this can have it **certified**.

The five core statements

1. **Here, honesty is not a virtue but a method.**
   Fifty-four times we have retracted one of our own findings — including results already reported as successes. Every retraction is recorded with its cause and its cost in a machine-verified register. A system that finds its own errors is the precondition for any later certification.

2. **Negative results are expensively acquired knowledge.**
   That the regime label carries no directional information, that the microprice leads by only 90 milliseconds, that the size evidence beats its own placebo by 0.2 percentage points — each of these took weeks to establish and saves any successor exactly those weeks.

3. **We have stopped tuning thresholds.**
   The most expensive mistake in this industry is adjusting parameters until the backtest looks good. Since July 2026 the rule is fixed *before* the calculation — pre-registered, versioned, dated in the commit history.

4. **We follow leads to the end — even when the end is negative.**
   The July lead was left undecided; its endpoint was retired. In September three pictures fell against their null line; the candidate flush absorption hit 65.6 % but lost $5.17 per trade and is discarded by rule. Tick-exact replays show: the entries carry no edge — not even reversed. What stands: a shadow filter with a strong but exploratory signal (`p = 0.0001`) and a measurement basis checked against the exchange (97 %).
   So: honestly negative, not abandoned.

5. **Development speed grows with every model update.**
   The project is built by one founder and one AI agent operating in "autonomous co-founder" mode. What took days in 2025 takes hours today — and the next model generation will move that boundary again. Details in Part VI.

> **The investor sentence, in one line:** You are not investing in the claim of an edge — you are investing in the only laboratory in this industry capable of *refuting* one, and therefore the only one capable of proving one.

FUNDAMENTALS

What an "edge" actually is

An edge is not a feeling, not a hit rate, and not a good month. An edge is a **measurable, repeatable deviation from chance** that survives a serious attempt to refute it.

The four conditions

To speak of an edge, **all four** must hold:

- **A** **A benchmark exists.**
  You must know what *chance* would deliver in exactly this situation. That number is almost never 50 %.
- **B** **The deviation is large enough for the sample.**
  A 70 % hit rate over 7 trades is meaningless. The same rate over 400 trades is a finding.
- **C** **It holds outside the data used to find it.**
  Anything else is memorization.
- **D** **It survives a serious refutation attempt.**
  Placebo control, reversed evaluation order, shifted time axis.

Why the null line is almost never 50 %

An example from our own retraction register: we measured a hit rate for one prediction head and compared it against 50 %. Result: "32 points below chance" — alarm.

In fact the correct null line ranged between **2 % and 20 %** depending on the day, because a large share of outcomes were neither "up" nor "down" but sideways. Measured against the *correct* line, the head sat **almost exactly on chance**.

Retracted · S392

The most expensive sentence in this industry

"*The hit rate is 75 %.*"
This statement is worthless unless it also states: **75 % of what**, **over how many independent observations**, **against which null line**, and **measured in which time window**. Without those four, it is marketing.

The difference between describing and predicting

The most important and most frequently overlooked distinction. A model can describe a market movement perfectly and still be worthless — because it only describes the movement once it has already happened.

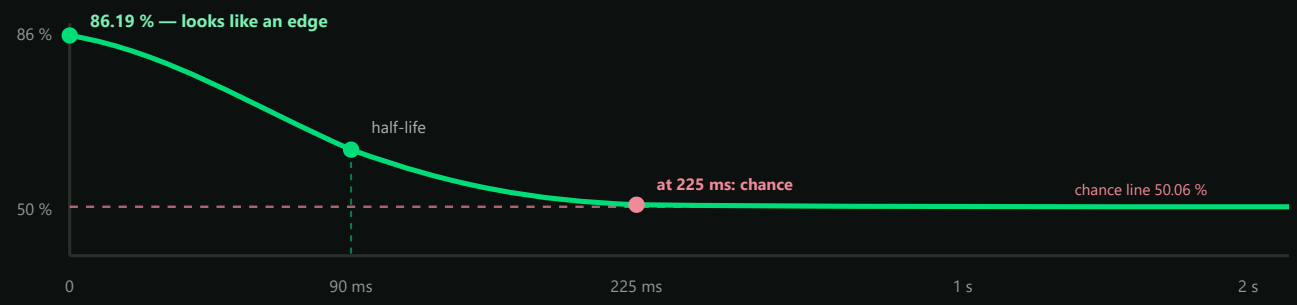


Figure 1: The decay curve of the microprice signal, measured across 936,000 de-duplicated individual trades (July 27, 2026). The 86 % hit rate is real — but it describes what is happening now and predicts nothing beyond 90 milliseconds. Without this curve it would have become an "edge."

The rule that followed: A metric without a decay curve is incomplete — exactly as a hit rate without a null hypothesis is. Both have been mandatory columns in every analysis tool since July 2026.

LESSONS FROM ERROR

The five traps — and how we fell into every one

This list is not theory. Every trap cost us time; every one is recorded with its date and its cost in the retraction register. We show them because an auditor will ask about exactly this.

Trap 1 — Guessing the null line instead of computing it

fell twice

You compare a result against 50 % because that intuitively sounds like "chance." But the actual chance line depends on the distribution of possible outcomes — in our case it ranged between 2 % and 20 %. **Consequences:** one false alarm ("32 points below chance") and, worse, a progress indicator that reported learning for a year where in reality the share of sideways outcomes was falling.

Trap 2 — A coverage rate without a control

cost: one evening + one concept

A rule matched in **75.4 %** of cases. This was reported as "the first tool to clear its pre-registered hurdle." The placebo control — the same rule applied to a data point **50,000 trades away**, guaranteed to be unrelated — matched **75.2 %**. Information content: **0.21 percentage points.** The rule was not measuring the market; it was measuring its own construction.

Trap 3 — Mistaking log lines for observations

fell three times

One event produces ten log lines. Count lines instead of events and a sample appears ten times larger than it is — invalidating every significance test. In our case: **311 lines = 13 market moments**. After correction the p-value rose from 0.005 to 0.119; a "finding" became "no verdict." The same trap in another disguise: duplicated trade records, because two program instances were writing to the same file (**44.5 % exact duplicates**).

Trap 4 — Comparing two numbers with different scopes

fell five times

Two individually correct measurements that use different time windows, accounts or sampling grids. The result looks like a contradiction in the market and is a defect in the display. Most prominent case: a delta metric reading **+13,225** in one window and **−16,816** in the other — no sign error, but two different session definitions.

Trap 5 — Verifying against your own calculation

most recently: July 29, 2026

You check a conversion by comparing its output against a number produced by that same conversion. The error stays invisible. In our case: a two-hour time zone shift that invalidated an entire analysis. It was found not by inspecting the result but by a **domain** question: "Does the reconstructed price path of a full loss ever touch its stop price?" Answer at the time: only in 35 % of cases. After the fix: 100 %.

What these five traps have in common: none is a programming error. All five are *measurement* errors — the code ran correctly and measured the wrong quantity. This is why software quality assurance is not enough; a second, independent verification path is required. Part III describes exactly that.

How each was found

Trap 1 by recomputing the baseline from raw data · 2 by a placebo arm · 3 by grouping lines into events · 4 by asking which window a number belongs to · 5 by a domain question, not by inspecting the output

What each now costs

Every one has an automated checker attached. A repeat of any of the five turns a build red rather than producing a report.

SYSTEMATICS

The ladder of evidence — six rungs from "interesting" to "certifiable"

Every claim about an advantage can be placed on this ladder. The industry typically sells at rungs 1 and 2. Certification requires rung 6.

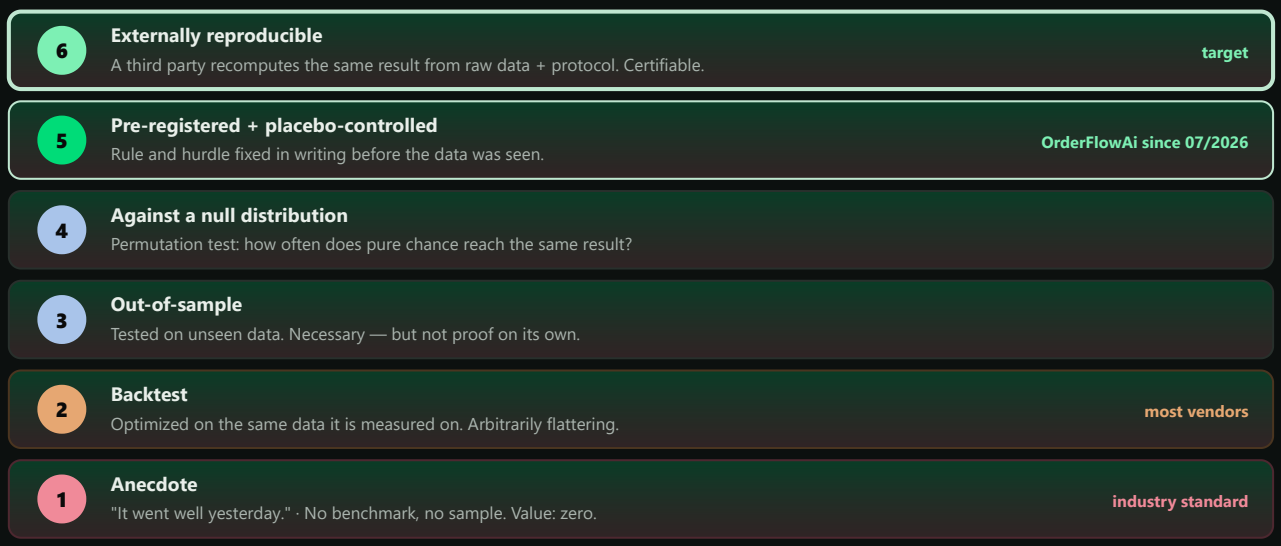


Figure 2: The ladder of evidence. The decisive jump is from rung 4 to rung 5 — it costs nothing but discipline and is nonetheless the rarest step in the industry.

Where we stand

The **apparatus** stands on rung 5 and is prepared for rung 6: raw data is archived, every tool tests itself, pre-registration is versioned and machine-enforced. The **result** stands at "no advantage demonstrable" — which is a clean rung-5 outcome, only a negative one.

Why that is worth more than it sounds

An apparatus that produces a *negative* result at rung 5 is proven to work. An apparatus that only ever produces positive results is suspect. The first is a measuring instrument; the second is a sales brochure.

PHASE 0 · AUGUST 2025 – FEBRUARY 2026

The first lines: a tool, not an oracle

On **August 1, 2025** the first internal alpha goes live. It predicts nothing. It displays something — and specifically something most trading programs cannot render at all: the **order book**, the list of all open buying and selling intentions, live and at full depth.

What was built in this phase

AUG 1, 2025 · V0.01.1
Internal alpha
First rendering of market depth as a heat map. Pure display tool.

AUTUMN 2025
Integration with NinjaTrader 8
The trading platform supplies the data; our own software computes and displays. This separation — **third-party execution, own analysis** — remains the architecture to this day.

JAN 1, 2026 · V0.01.6
Spoofing detection
First *interpretation* rather than plain display: orders placed only for show, which vanish before execution, are flagged.

FEB 1, 2026 · V0.01.7
Session filters & export
From here on, analyzable recordings exist — the foundation for everything that follows.

MAR 3, 2026
Four-phase reversal scanner
The first attempt to classify turning points. Rule-based, still without learning.

Why Level 2 data is the starting point

A conventional price chart shows *what* has happened. The order book shows *who is currently willing to trade* — and at what price. It is the only data source available to retail participants that makes intent visible before execution.

That is precisely why it is also the only source in which an advantage *could* plausibly be hidden. Anyone analyzing only candlestick charts is analyzing publicly available history.

The architectural decision that made everything possible

The software was built from the outset as an **observer**, not as a trading platform. It reads the data stream, computes in parallel, and records everything.

The side effect only became apparent a year later: because **every** session was archived in full, analyses could be run in 2026 on data from 2025 that nobody had planned for. **Without that archive there would be no research report.**

Position after Phase 0

1	0	~6 mo
PRODUCT	PREDICTIONS	ARCHIVE

What distinguishes this data source from a price chart

LAYER	WHAT IT SHOWS	WHY AN ADVANTAGE COULD HIDE THERE
Candlestick chart standard, universally available	Completed price movement in fixed time slices.	**Practically not.** Public, delayed, and analyzed simultaneously by millions of participants.
Time & sales widespread	Every trade with price, size and timestamp.	**Limited.** Shows execution, not intent — and who was the aggressor must be inferred.
Order book (Level 2) basis of this project	**All open buying and selling intentions at full depth, continuously updated.**	**Yes, potentially. The only layer accessible to retail participants where intent becomes visible before execution — including feigned intent.**

In hindsight the most important decision of this phase: recording everything even though there was no use case for it at the time. Data archives cannot be created retroactively — and the value of a market dataset grows with every day you have failed to record it.

PHASE 1 · MARCH 2026

From displaying to predicting — and the first temptation

In March 2026 the display tool becomes a learning system. In four weeks the building blocks that still carry the project today come into being — and at the same time the error that will later cost a quarter of a year to correct.

MAR 11, 2026
Overhaul of the AI/ML engine
First serious model architecture. Goal: estimate the next price direction from the order book state.

MAR 11, 2026 · V0.02.3
Multi-timeframe trend engine
Several time horizons at once — more context for the same model.

MAR 13, 2026
Market speed as an input
How many trades per second? A simple but effective contextual value — the very sensor whose silent failure will go undetected for months in July 2026 (page 13).

MAR 17, 2026
Institutional edition & GEX engine
Options-derived positioning data as an additional information layer.

MAR 26, 2026
"Prometheus" neural engine
The first in-house neural engine gets a name. Two days later it is renamed **CORTEX** — the name that stays.

The first temptation: 70–85 %

This phase produces numbers that later live on as recollection: hit rates of 70 to 85 % over individual trading days. They were not invented — they simply were not what people took them to be.

The July 2026 review found: the evidence consisted of **4 to 9 trades per day**, measured in a supervised window around the New York open. The phrase "70–85 %" appears repeatedly in the archive as a **target**, not as a measurement. One line reads verbatim: `10 trades 80 % WR +$25` — four out of five trades won, yield: 25 dollars.

How a recollection becomes a metric

Nobody lied. It happens gradually:
1. A good day is noted.
2. The note is quoted.
3. The quote loses the sample size.
4. "On one day, 7 trades" becomes "the hit rate."

Countermeasure, in force since July 2026: every rate is written only together with sample size, null line and time window. Tools that fail to output these are treated as defective.

The real progress of this phase was not the model but the infrastructure: from here on a continuous path existed from the trading platform through the analysis into an analyzable log. That chain is the precondition for any result later being *traceable* at all.

PHASE 2 · MARCH 28 – APRIL 17, 2026

CORTEX is born

On March 28, 2026 the neural engine receives its final name. On April 17 it ships to customers as **CORTEX v0.3.0**. Between those two dates lie three weeks in which a model becomes a product.

What CORTEX is technically

A recurrent neural network (LSTM) that learns from a sequence of market snapshots. Each snapshot consists of **41 features** — order book imbalances, absorption events, trading speed, distance to the volume point of control, options pressure and others.

The model runs **locally on the user's machine**, continues learning from that user's own market sessions, and sends no data anywhere. This is unusual — and it is one reason the research is cleanly possible at all: every instance is an independent experiment.

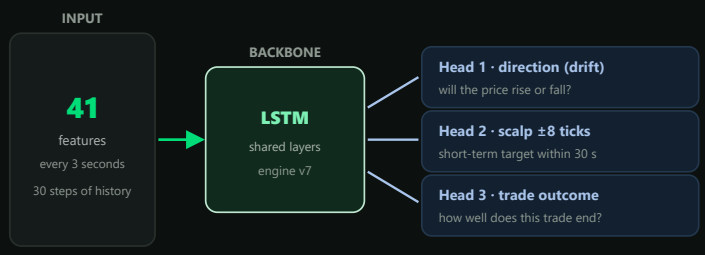


Figure 3: CORTEX architecture. One shared backbone, three specialized outputs. The third head was the first to be completely refuted, in July 2026 (page 17).

The number that costs everything

```
NUM_FEATURES = 41
ENGINE_VERSION = 7
```

Change the count or meaning of the features and the learned model no longer fits — **it must be deleted entirely**. The project term for this: a *wipe*.

Between March and July 2026 this happened repeatedly. Each time the model was "zero sessions old" again. **That is why CORTEX never improved** — not because it could not learn, but because it was never allowed to for long enough.

The lesson drawn

A learning system needs stationarity. You cannot test and learn at the same time: every change to the input resets the learned state, and every learning phase forbids changes.

Formulated on July 20, 2026, after four months in which both were attempted simultaneously.

What the 41 features consist of

Book structure
Bid/ask imbalance, depth distribution, walls, feigned orders

Flow
Trading speed, cumulative delta, absorption events, large single prints

Location
Distance to volume point of control, to the open, to prior-day levels, time of day

Context
Options-derived positioning pressure, market regime, session segment

Each feature is sampled every three seconds; the model sees the most recent **30 steps**, roughly a minute and a half of history. The selection was made on domain grounds — which of them actually carried information was only measured in July 2026 (result: the state vector as a whole is overfitted, see page 17).

April 17, 2026 — CORTEX launch v0.3.0. Pro, Trial and Starter editions ship with the neural engine. At that point there is **no** measurement substantiating an advantage — there is a working engine and the reasoned hope that one will emerge. The difference between those two only becomes painfully clear three months later.

PHASE 3 · APRIL – JUNE 2026

The euphoria — and the first numbers that disagreed

For two months everything runs at once: product development, marketing, customer acquisition, strategy work. It is the most productive and simultaneously least scientific phase of the project.

What succeeded

MAY 5, 2026
"NY Open Daily" — 11 winning days, 1 losing day
A publicly documented streak. Day 12 becomes the first losing day and ends the series at −$700.

MAY 7, 2026
CORTEX-OEM v0.5.0 — autonomous order engine
The AI takes over entry, exit and position management. 26 trades in a simulation account, +$600, 69 % hit rate.

MAY 19, 2026
Live 75 % hit rate
Another very good day. Another small sample.

JUN 27, 2026
Official NinjaTrader listing
Admission to the vendor ecosystem — external confirmation of product maturity.

What became visible at the same time

MAY 15, 2026
11 trades, 3 wins, 8 losses
27.3 % hit rate, −$747.82. A single day — but the first pointing in the same direction as the eventual aggregate.

MAY 30, 2026
Strategy evaluation: NO-GO
Our own trade analysis refuses to clear the new version.

JUN 29, 2026
A safety mechanism locks out the best day
Two early losses trigger a day-lock; the subsequent 22-point uptrend is missed entirely. **A protection that costs more than it protects.**

The pattern behind all three cases

Good days were documented and shared; bad days were analyzed and *fixed*. That sounds sensible — but it means the aggregate never gets computed. It was finally computed on **July 29, 2026**, on request, across all **422 trades from 23 days**.

The aggregate that arrived two months late

METRIC	MEASURED	MEANING
Overall hit rate	42.7 %	Break-even would be 49.0 % (reward-to-risk 1.04)
Result per trade	−$23.53	across 422 trades
Cumulative	−$9,930.88	simulation account, 23 trading days
Trades reaching the first target	100 % win	n = 100 · average +$320
Trades not reaching it	25 % win	n = 322 · average −$130
Losers dead within 30 seconds	40 %	**It is not a direction problem. It is a timing problem.**

This table is the turning point of the entire project. It says: if a trade survives the first half-minute and reaches its first target, it *always* wins. That shifts the research question from "which way is the market going?" to **"is this moment a good entry?"** — a question that is considerably smaller, more concrete and more provable.

PHASE 4 · JULY 6 – 20, 2026

The break: fourteen days in which nearly every hope fell

In July 2026 testing is **adequately powered** for the first time — with enough data, with correct null lines, with cross-validated procedures. The result is unambiguous and unpleasant. It is also the moment a software project becomes a research project.

JUL 6, 2026 · SESSION 337

CORTEX v5 — verdict: NO EDGE

The first complete evaluation of the neural engine against a correct benchmark. No demonstrable advantage.

JUL 7, 2026 · SESSION 340

Offline learnability study: no learnable advantage — and underpowered

The second part is the notable one: the study **itself determined** that its data volume was not yet sufficient for a firm verdict. A tool that reports the limits of its own conclusiveness is the beginning of serious measurement.

JUL 15, 2026 · SESSION 358

First adequately powered verdict: NO EDGE (global)

This time at full statistical power. Metric: Matthews correlation `MCC ≈ 0.02` — effectively chance.

JUL 15, 2026 · SESSION 359

No advantage in *any* market regime

The rescue hypothesis was: the advantage exists but only in certain market phases, so it averages out in the pooled sample. Tested across **26,940 observations**: range `p = 0.29` · uptrend `p = 0.30` · downtrend `p = 0.80`. **Falsified.**

JUL 16, 2026 · SESSION 362

Tick-exact test: NOT LEARNABLE

The last technical excuse — "the data resolution is too coarse" — falls. Even at exact resolution, the target is not learnable from these features.

JUL 20, 2026 · SESSION 368

The regime label carries no directional information

15 of 16 tested cells fall below baseline, and in roughly **19 %** of cases the label even points the wrong way. This also ends threshold tuning: **you cannot tune an uninformative signal into shape.**

JUL 20, 2026 · SESSION 369

Strategy shift — product and research are decoupled

The central commercial consequence. From now on the product sells what it demonstrably does: visualization, order book analysis, journaling, heat map. The edge investigation continues as a **separate research branch** — without sales pressure, without deadlines, without temptation.

What fell in these 14 days

- The neural engine's global advantage
- The regime-conditional advantage
- Learnability from the 41-feature state
- The regime label's directional information
- The hope that finer data would solve it
- Four months of threshold calibration

What arose in these 14 days

- A measuring apparatus that can say **no**
- The separation of product and research
- The principle: no verdict before the counter-check
- The retraction register as a mandatory instance
- The insight that learning systems need stationarity
- A research question small enough to prove

Why this section appears in an investor report: because a team that refutes its own core hypothesis six times in fourteen days and then restructures the business model displays exactly the behaviour one must require of a research company. The alternative — keep searching until some number fits — would have taken a year longer and collapsed at the first external review.

PHASE 5 · JULY 21, 2026

CORTEX-ONE — one brain instead of many opinions

The break is followed by reconstruction. The diagnosis was not "the model is too weak" but: **there were too many divergent truths inside the system.**

The problem: distributed truth

Until July 2026 several program components each computed the *same* signal independently — the main window, the heat map window, the trading strategy. Because each used slightly different time windows or caches, three displays could show three different directions for the same second.

For a display that is a cosmetic flaw. For a **measurement** it is fatal: you no longer know which number the model actually saw.

The solution: a signal bus

One producer generates the signal; all others consume it identically. Plus two fixed, clearly separated time horizons instead of arbitrarily many variants:

SCALP
±8 ticks within 30 seconds — short-term target

SWING
larger move, longer horizon

The frozen foundation

On July 21, 2026 version `v0.4.80` is **pinned as the learning baseline** — with a marker in the version history that cannot be moved.

From this point on: **no changes to the feature vector.** No new feature, no new engine version, no wiping of the learned state. The model is finally allowed to age over weeks.

41 features · engine v7 · frozen

The pre-registered gating protocol

The same day, the **CORTEX gating protocol** is written. It fixes in writing, *before* any result exists:
· when a signal counts as viable
· which statistical hurdle applies (lower bound of the confidence interval, not the point estimate)
· over how many sessions it must hold
· **three abort criteria** at which the search is discontinued

And explicitly what is **forbidden**: changing thresholds after seeing a result.

Figure 4: CORTEX-ONE. The rebuild did not produce a better prediction — it produced **traceability**. Without it, no certification is conceivable.

Why this is the actual precondition for a certificate: an auditor will ask "what value did the model see at this instant?" Before CORTEX-ONE there were three answers to that question. Since CORTEX-ONE there is one — and it is readable in the log.

PHASE 6 · JULY 28 – 30, 2026

From mirror to instrument

The most recent and so far most important diagnosis: **CORTEX was a mirror.** It was fed from the same raw quantities as the trading strategy — and therefore inevitably confirmed whatever the strategy already thought. On July 28 the two agreed on **33 out of 33** trades.

The mirror finding

Two systems using the same inputs are not two opinions. Their agreement proves nothing.

Compounding this: the model's **confidence** was *anti-predictive*. Winning trades averaged 53.2 % model confidence, losers 57.4 %. At confidence of at least 70 %: **0 % win rate** (n = 4).

A better model would not have solved this. It required a second, *independent* measurement channel.

The answer: a second channel

Two quantities were chosen that work **without** the error-prone attribution "was that a buyer or a seller?" and therefore *bypass* the broken path rather than repairing it:

OFI — order flow imbalance: measures the *flow* in the book rather than its state.
Microprice — the queue-weighted fair price.

And the objective was deliberately narrowed: **not a directional model, but an entry gate.** Not "where is the market going" but "is now a good moment."

Three findings in three days

1. **The size-evidence path carries no information (Jul 28–29)**
   A rule for detecting aggressive market participants matched in 75.4 % of cases. Against its own placebo: 75.2 %. Distribution analysis further showed that 87.7 % of all trades involve a single contract — and that on large trades (25 contracts and up) the rule's hit rate is
   0.0 %
   . It was blind to precisely what it was meant to find. retracted

2. **Our instrument was recording twice (Jul 29)**
   Two program instances were writing to the same file.
   44.5 %
   of all lines were exact duplicates. Every trade-weighted statistic before that was double-counted. Found only because each line now carries an instance identifier. recording repaired

3. **A dead sensor blocked 1,228 of 1,228 decisions (Jul 29)**
   The market-speed value was queried in a window where the corresponding variable does not exist. Instead of reporting an error, the code filled the gap with the fallback value `"CALM"` — quiet market. For months the strategy saw an apparently dead market and therefore did not trade.
   This became a governing rule in the code:
   sending nothing is better than sending a fallback value. A missing value looks like a defect — a disguised fallback looks like a measurement. fixed

The common thread of these three days: none of the findings concerned the market — all three concerned the **instrument**. That is not coincidence; it is the normal ordering in any empirical science. Before making claims about the world, you must prove the apparatus works. The industry almost always skips this step.

OVERVIEW

Twelve months at a glance

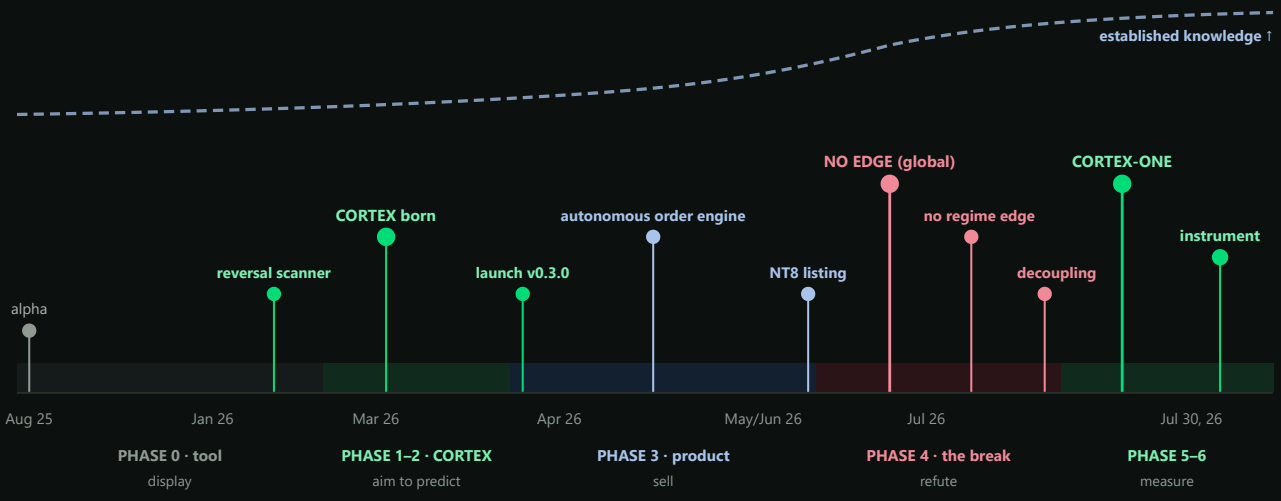


Figure 5: Timeline. The dashed line shows when *established* knowledge grew fastest — precisely when the results were at their most negative.

894
LOGGED WORK ENTRIES

~207,500
LINES OF CODE (JAVASCRIPT + C#)

808
KNOWLEDGE ENTRIES IN THE PROJECT BASE

4
PRODUCT EDITIONS IN PRODUCTION

What each phase contributed

PHASE	GUIDING QUESTION	LASTING CONTRIBUTION
0 · tool	What happens in the order book?	The data archive. Without the decision to record everything from day one there would be no dataset today — and datasets cannot be created retroactively.
1–2 · CORTEX	Can it be predicted?	The learning engine and its local execution on the user's machine. Every installation is an independent experiment, with no data leaving the premises.
3 · product	Will anyone buy this?	Four editions in production, paying customers, an official ecosystem listing — the commercial base that carries the research.
4 · the break	Is any of this true?	Six falsifications in fourteen days and the decoupling of product and research. **Measured in established knowledge, the most valuable phase.**
5–6 · instrument	**How do you prove it?**	**The complete evidence apparatus: pre-registration, placebo control, retraction register, self-testing tools, certificate generation.**

An observation worth pausing on: from August 2025 to June 2026 the *product* grew fast and *established knowledge* grew slowly. From July 2026 that reversed. Both phases were necessary — but only the second produces something that can be certified.

THE APPARATUS · INSTRUMENT 1

The retraction register — 54 findings we overturned ourselves

The most unusual institution in this project is a file containing every statement that has proven **false**. Not as a footnote — as a standalone, machine-monitored register with four mandatory fields per entry.

Why it exists

In July 2026 an already-refuted finding cost a full working day. The correction was present in the authoritative file — but an automatically loaded summary carried the false statement onward. A version comparison would never have caught this: **both files were "current."** Only a register of retracted *statements* finds something like that.

Since then the rule is: when a finding is refuted, it enters the register **in the same work step** — not later, not "at the next cleanup."

The four mandatory fields

FIELD	CONTENT
`pattern`	A search pattern that finds the **false** statement anywhere in the project — notes, code comments, reports.
`truth`	What holds instead, in one sentence.
`retracted`	When and by what means it was retracted.
`cost`	**What the error cost** — so the lesson does not remain abstract.

A checker runs across all files at every project start. If a pattern still appears anywhere as an assertion, it fires. **The past cannot creep back in.**

The first 22 entries

22

RETRACTED FINDINGS

caught in the same work step	**8**
felled by a counter-check	**6**
corrected by the founder	**4**
found only days later	**4**

Most expensive single error: a full working day — an entire explanatory chain was built on a false premise *and* a rebuild was designed that would have replaced a working component.

What an auditor reads into this

A company that keeps an error register does not make more errors than others — it **finds** more. For a certification body that is precisely the decisive difference between a measurement laboratory and a marketing department.

The uncomfortable number: of the first 22 retracted findings, **five had already been reported to the founder** before the counter-check ran — including a "first genuine out-of-sample proof" and a "first tool to clear its pre-registered hurdle." From this came the project's hardest rule: *no verdict before the counter-check — and a positive result inside a chain of negative findings deserves more suspicion, not less.*

CASE STUDIES

Four falsifications in detail (1 of 2)

Selected not for drama but for instructional value: each shows a different way in which an apparent advantage arises where there is none.

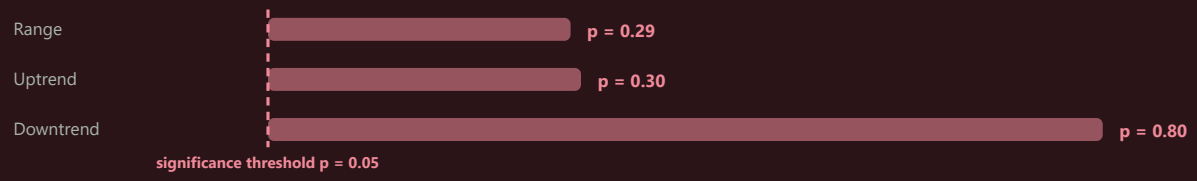
Case 1 — "The advantage is regime-conditional"

falsified · Jul 15, 2026

The hypothesis. The model shows no advantage globally — but perhaps it works very well in certain market phases and very poorly in others, so the two cancel out in the pooled sample. This is a good, plausible idea and a common one in the literature.

The test. All observations were evaluated separately by market phase — against *ground truth*, i.e. the price path that actually occurred, not against the system's own assessment. Sample: **26,940 observations**.

The result.



The further a bar extends to the right, the better the result fits the assumption "pure chance." All three market phases lie far beyond the threshold.

The consequence. A better regime label does not recover an entry advantage. This ended several weeks of work on a "regime router" — and with it the temptation to tune its thresholds.

Case 2 — "The model is learning, the curve is rising"

retracted · Jul 28, 2026

The claim. A learning curve showed a rise in hit rate from **0.8 % to 18.1 %** across several sessions. This number had sat in the header of the analysis tool since July 2026 and was printed on every run. It was the **only progress indicator** of the frozen learning phase.

The error. Over the same period the share of sideways outcomes fell from **96.1 % to 60.0 %**. The theoretically attainable ceiling therefore rose with it — *independently of the model*. Computed against the correctly determined, day-dependent chance line, **4 of 5 sessions are negative**.

The lesson. A progress indicator that does not compute against a *moving* null line reports market change as learning success. It did exactly that for a year.

What Case 1 cost

Several sessions of work on a regime router the data did not support — plus the temptation to tune its thresholds.

What Case 2 cost

A year of a progress indicator reporting market change as learning — the only one the learning phase had.

What both produced

The obligation to compute every null line from raw data instead of assuming it. Now enforced in every tool.

The common denominator of both cases: in each, the measurement was technically correct and the code free of bugs. What was wrong was the **frame of reference** — what it was compared against. This is why testing software is not enough; you must test the *question*.

CASE STUDIES

Four falsifications in detail (2 of 2)

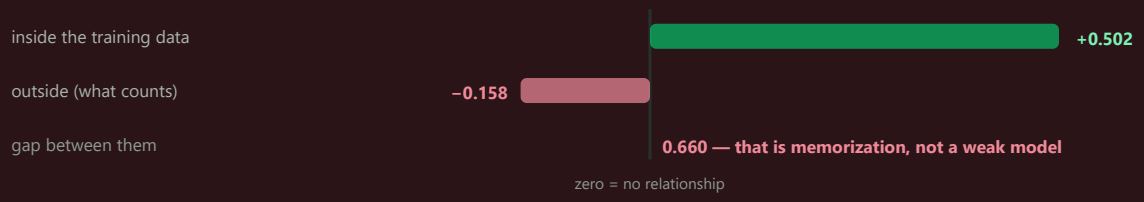
Case 3 — The third prediction head, answered after eight months

negative · Jul 29, 2026

The open question. Since November 2025 a release condition had been standing: the model's third head — which estimates a trade's *outcome* from the market state at entry — could only be activated once its predictive power **outside the training data** was demonstrably above zero.

Why it took eight months. Not because the computation was hard. Because there was no *general* test method — every question of this kind was answered individually, by hand, and slightly differently each time. Only when this became a reusable library was the question settled in minutes.

The result. Measured across **295 usable trades**:



Rank correlation between model estimate and actual trade outcome. **p = 0.96 · 0 of 10 data splits positive.**

The two side findings. The same test for the simplified question "does the trade end in profit?" yielded `AUC 0.421`, and for the proposed entry-gate head `AUC 0.382` — both below the chance level of 0.5. **On the positive side:** the technical precondition — that the data store persists across program restarts — was proven in the same run.

Case 4 — "The relationship has merely inverted"

pre-registered refutation · Jul 29, 2026

The last excuse. When a model performs *negatively* outside the training data, two explanations are available. The flattering one: the market has changed, the relationship exists, it has merely flipped sign. The boring one: there never was a relationship, and the model memorized the training data.

The decisive manoeuvre. The decision rule was fixed in writing *before* the computation — with the **boring explanation as the default**. Only a clearly defined pattern could have supported the flattering one.

The result. All three target variables lie **forwards and backwards** below their own null line: `0.382 / 0.337` · `−0.158 / −0.151` · `0.421 / 0.391`. Had a relationship inverted, at least one direction would have to be positive. **It was never there.**

On the timescale: the question in Case 3 had been open since November 2025 and went unanswered for eight months — not for want of data, but because every test of this kind was built by hand. Only once it became a **reusable test library** with 23 self-tests of its own was the answer a matter of minutes. That same library then settled Cases 3 and 4 plus two further questions in a single evening. **Tool-building beats case-by-case work — by orders of magnitude.**

Why Case 4 is the most valuable: it is the proof that pre-registration works. Without it, one would have gone looking for an explanation after the negative result and would most likely have found the flattering one — there is always a data split that supports it. **The rule existed beforehand. So there was nothing to go looking for.**

THE APPARATUS · INSTRUMENT 2

Pre-registration: the rule precedes the calculation

The procedure comes from clinical research. There it is mandatory, ever since it emerged that studies with unwelcome results were systematically going unpublished. In financial software it is virtually unknown — and therefore the largest available lever.

The problem it solves

Every analysis involves dozens of legitimate choices: which time window? which minimum sample? which metric? which outliers stay in?

Make those choices *after* seeing the data and you will almost always find a combination that looks good — entirely without bad intent. The technical term is the **"garden of forking paths."**

Pre-registration closes that garden: all choices are fixed in writing, versioned and dated **before** the analysis runs.

How it is implemented here

1. **A versioned specification file**
   holds every requirement — metrics, minimum samples, hurdles, abort criteria.
2. **A checker enforces it.**
   If an analysis tool deviates from the specification, it does not run.
3. **Exactly one confirmatory endpoint.**
   All other analyses are explicitly exploratory — they may generate hypotheses but may not establish any.
4. **Minimum samples counted in trading days,**
   not in observations. A single day yields thousands of correlated data points — but only *one* independent observation.

What pre-registration enforced immediately

On its introduction on July 29, 2026, **three defects** in an already-finished analysis tool surfaced in the same moment:

· the minimum sample was set too low at 30 → raised to **150**
· no primary time horizon had been fixed → added
· the metric was a difference value rather than a normalized skill score → replaced

In addition an entire test stage was **locked**, because its statistical power stood at only 28 % — it would very likely have missed a real result while still counting as "tested."

The hardest self-imposed constraint

From the gating protocol, in substance:

Forbidden: changing thresholds after a result has been seen, so that a candidate qualifies.

Forbidden: releasing on a point estimate — always on the lower bound of the confidence interval.

Forbidden: skipping the shadow phase and going live directly.

Each of these three rules arose from a specific, documented error of our own.

Translated for investors: pre-registration costs nothing but discipline and is the difference between a result that survives due diligence and one that disintegrates during it. It is also the reason a future positive result from this project would be **certifiable at all** — a retrospectively discovered advantage never is.

THE APPARATUS · INSTRUMENT 3

The placebo control — the test that cost us the most

In medicine a control group receives an inert sham treatment. If the real drug does not perform markedly better, the observed improvement was not evidence. The same principle transfers to market data — and almost nobody applies it.

The case that forced the rule

On the evening of July 28, 2026 a tool was reported as "*the first to clear its pre-registered hurdle.*" Its job was to determine whether an aggressive buyer or seller stood behind a given trade — the foundation of virtually all order flow analysis.

Hit rate: **75.4 %**. Against a naive benchmark of 8.9 % that was a factor of 8.5. It looked like the first genuine breakthrough.

Then the placebo ran. The same rule, applied to a data point **50,000 trades away** — that is, guaranteed to be unrelated.

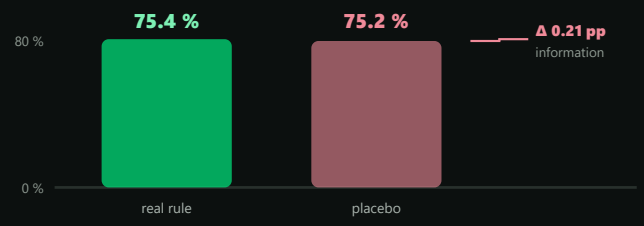


Figure 6: The rule was not measuring the market — it was measuring its own construction.

Why the rule matched almost always

The distribution analysis explained it immediately:

· **87.7 %** of all trades involve only a **single contract**.
· With a tolerance of ±1, the condition effectively read "any change between 0 and 2" — which is almost always true.
· On *large* trades of 25 contracts and up, the hit rate was **0.0 %** (n = 314).

The rule was blind to exactly what it was meant to find.

The second find of that same night

Same file, same data, only the order of evaluation steps reversed:

Variant A: **+136,155**
Variant B: **−133,141**

The sign of the overall result flips purely from the evaluation order. That is not a market finding, that is a construction defect — visible only because someone reversed the order as a test.

The benchmark

A simple rule of thumb, known for decades, reaches **86.43 %** on the same contracts in a peer-reviewed evaluation.

Our own 75.4 % therefore sat **below the trivial benchmark** — something nobody would have noticed without the placebo control, because the comparison had been against a *self-chosen* benchmark.

The rule that followed: placebo control, size distribution and duplicate rate have been **mandatory lines** in every classification tool since July 29, 2026. A coverage rate without a null hypothesis is not evidence — it is a number.

THE APPARATUS · INSTRUMENT 4

The decay curve — separating describing from predicting

Perhaps the most elegant test in the whole apparatus, because it answers a question that is otherwise never asked: **how long is a signal valid for?**

The principle in one sentence

You measure the same hit rate not only "now" but at increasing temporal distance — after 5 events, after 50, after one second. If it drops to chance level immediately, the signal was a **description** of the present, not a prediction of the future.

Two measurements, two consequences

Microprice — half-life 90 milliseconds

Measured across **936,000 de-duplicated trades**. Immediate hit rate: **86.19 %** against a majority class of 50.06 %. After just five events (about 225 milliseconds): **chance**.

Consequence: no second measurement channel. The test stage was closed and explicitly marked "do not repeat."

Touch imbalance — dead after 1 second

The information content fell from a skill value of 1.05 to 1.00 at one second of delay.

Consequence — and this is the valuable part: the existing data path delivered this value once per second with up to two seconds of age. At entry time it was therefore *mostly already dead*. In response a dedicated direct data channel was built that writes every **200 milliseconds** — and the staleness tolerance was cut from 2,000 to **500 milliseconds**.

Figure 7: Both signals are *real* — and both are worthless by the time of entry if the data path is too slow.

Why this is an engineering question

The decay curve moves the problem from statistics into engineering. If a signal lives for 90 milliseconds, success is decided not by the model but by the **latency of the data chain**.

That is good news: latency is solvable. A non-existent relationship is not.

What the curve rules out

A signal that is real, strong and useless — because it has already expired by the time an order can act on it.

What it turns into a spec

Half-life becomes a latency budget. 90 ms of signal life sets a hard ceiling on how slow the data path may be.

What it saved here

An entire test stage was closed on the evidence and marked "do not repeat" — instead of being pursued for weeks.

Mandatory since July 2026: every reported metric carries a decay column. Without it the finding counts as incomplete — on the same footing as a hit rate without a null hypothesis.

THE APPARATUS · INSTRUMENT 5

Scope — who does this number actually belong to?

The most frequent class of error in this project, appearing five times in different disguises: two individually correct numbers describing **different slices** of the same reality — and therefore not comparable.

MANIFESTATION	WHAT WAS COMPARED	CONSEQUENCE
Two session windows	One metric was anchored at 18:00 New York time, the second at 09:30. Result: `+13,225` versus `−16,816`.	Looked like a sign error in the trading logic. Was a display question.
Two sampling grids	A computation ran every 5 seconds, but its log line was written only every 30 seconds — i.e. every sixth one.	63 % of the time windows contained a state the log never showed.
Two accounts	A daily P&L mixed two trading accounts. That number drove a *shutdown rule*.	A manual trade on the second account ended the automated trading day.
Two time zones	Timestamps already carried local time but were read as universal time.	Two hours of offset. An entire analysis invalidated.
Two definitions of "usable"	Two verification contracts were both called "citable" and measured different things: one reported 4 of 9 days, the other 1 of 9.	The intersection was **empty** — both values correct, jointly useless.

The countermeasure: a scope guard

A dedicated verification module ensures, before every analysis, that the categories used **exhaust the entire population**. A textbook example from our own operations: `2,529 + 1,126 ≠ 9,906` — more than 6,000 cases were silently falling into no category at all.

The guard has **ten self-tests of its own**, and its counter-check is precisely this real case: it must find it, or it counts as defective.

The question that resulted

Before any statement of the form *"this number drives something,"* the question is now:

Who does this number belong to?

Which account, which time window, which instance, which time zone, which definition? Only once all five are answered may the number trigger a decision.

And: the guard itself once fell into this trap — it flagged a writer process legitimately idle after market close as a defect. Corrected with counter-checks for *both* window edges.

Why this is central to certification: an auditor will not ask "is the number correct?" but "what does it refer to?" A system that answers that question automatically for every metric is auditable. One that does not, is not — regardless of how good its result looks.

THE APPARATUS · INSTRUMENT 6

Tools that test themselves — and guards proven to be watching

A checking program that always passes is worthless. You cannot tell the difference from its output — only from whether it **can still turn red at all**. That is why every tool here has two built-in parts: a self-test and a counter-check.

170	50	22	17
ANALYSIS TOOLS IN INVENTORY	WITH A BUILT-IN SELF-TEST	STEPS IN THE EVENING RUN	OF THOSE WITH A DOMAIN ACCEPTANCE TEST

The difference between technical and domain acceptance

Technical acceptance — "did it run?"

The program completed without error, wrote a file, returned zero. **That says nothing about whether the result is correct.**

A tool can run flawlessly while measuring the wrong quantity — which is exactly what happened in all five traps on page 5.

Domain acceptance — "is the result right?"

A substantive question answerable only with correct data. The textbook example from our own operations:

"Does the reconstructed price path of a full loss ever touch its stop price?"

Answer on the first run: **35 %**. After repairing the data path: **100 %**. **And the final result flipped in the process.**

The case that shows why this is necessary

An analysis of the question "would an earlier exit have been better?" initially reported an improvement of **+6.96 R** at a threshold of 2 ticks. Domain acceptance revealed that the underlying price path was too coarse — it consisted of roughly 20 points per trade rather than the actual path.

Re-cut from the raw data stream (**median 1,447 points per trade**) the result inverted: the 2-tick threshold now *loses* 2.67 R, and the best value is an 8-tick threshold at +1.93 R. That is still not a verdict — the sample covers 30 trades over 2 days. **But the coarse path demonstrably pointed the wrong way.**

The four forms of "documented but not enforced"

1 · Rule without a guard
Written in the docs, checked by nobody.

2 · A path that stays silent
Does *nothing* on failure — instead of reporting.

3 · A guard that passes
Runs, but no longer tests its own claim.

4 · A guard without a self-test
Cannot prove it is able to turn red.

The working rule derived from this: whoever writes down a trap must, **in the same work step**, define the automated test that will catch it in future. Otherwise the documentation is merely a note describing how the same mistake will be made again next time.

STATE OF RESEARCH

What is established today — the balance after twelve months

This is the most candid page of the report. It separates cleanly what has been **measured**, what has been **refuted**, and what remains **open**.

Established negative — tested with adequate statistical power

HYPOTHESIS	TEST	FINDING
The neural engine has a global directional advantage	3× powered	**No.** Matthews correlation ≈ 0.02 — effectively chance
The advantage exists within specific market phases	n = 26,940	**No.** p = 0.29 / 0.30 / 0.80
The regime label carries directional information	v1 *and* v2	**No.** 15/16 cells below baseline, ~19 % wrong sign
The 41-feature state is learnable at all	tick-exact	**No.** Out-of-sample rank correlation −0.158, 0/10 splits positive
The microprice is a second, independent channel	936k trades	**No.** Half-life 90 ms — descriptive, not predictive
The size evidence identifies aggressors	placebo	**No.** 75.4 % against a 75.2 % placebo = 0.21 pp of information
Model confidence marks good trades	33 trades	**Inverted.** Winners 53.2 %, losers 57.4 % confidence
The AI position manager beats the simple rules	shadow run	**No.** 48 % against 64 % — remains disabled

Established positive

The timing finding. Across 422 trades from 23 trading days:
· trade reaches its first price target → **100 % win** (n = 100)
· trade does not reach it → **25 % win** (n = 322)
· **40 % of all losers are dead within 30 seconds**

This is the project's most robust insight, because it **shrinks** the research question: no longer "which way is the market going" but "is this moment viable."

Also established — technically

· The learning store **persists across program restarts** (stage 2 of the third head proven)
· Recording has been **duplicate-free** since July 29 (one writer elected per process)
· The data chain is accelerated to **200 ms** and its staleness tolerance is empirically justified
· The installed software is verified against its checksum on disk — the project rule "the user runs the installed build" has been **measured** since July 29 rather than assumed

The summary statement, as plainly as possible: from the market state captured today, the future price direction **cannot** be predicted better than chance using any of the methods tested so far. What is emerging is something different and smaller — that *entry moments* may be separable by their probability of survival. That is the current lead.

CURRENT LEAD

July 29, 2026 — a lead, explicitly not a verdict

On the day before this report was written, the CORTEX-rebuilt strategy ran for the first time with a repaired market-speed sensor. The result is striking enough to report — and weak enough that we do **not** call it a finding.

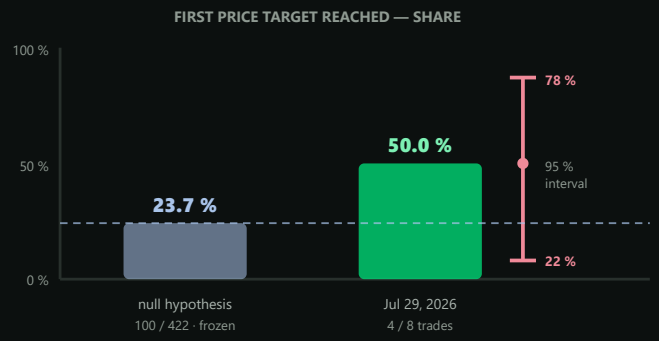


Figure 8: The confidence interval of the new value **still includes the null hypothesis**. That is precisely why this is a lead and not a result.

The mechanism reproduces

What is notable is not the rate but the **structure** of the day — it repeats exactly the finding from 422 comparison trades:

OUTCOME	COUNT	RESULT
First price target reached	4	**4 of 4 won** — all with a secured entry
Not reached	4	**1 scratch, 3 full losses** after 18 / 26 / 26 seconds
Day result	**8**	**+$522.64 · 62.5 % hit rate**

Direction did not improve — entries simply survive the first half-minute more often.

What made the day possible at all

Not a new model, not a new rule — the repair of a sensor that had been dead for months. The market-speed value was queried in a program window where it does not exist, and was silently filled with the fallback "quiet market."

Consequence: the *market too quiet* block rejected **1,228 of 1,228** book-confirmed decisions. After the repair it fell to **zero**.

The lesson now stands as a governing rule in the code: sending nothing is better than sending a fallback value. A missing value looks like a defect — a disguised fallback looks like a measurement.

Why we are not celebrating

1. Too small. One-sided: P(≥4 targets reached) = `0.097`. For the overall hit rate even `0.217` — which carries nothing at all.

2. The day formally does not count. On July 29 the trading logic was reloaded seven times; **five different builds** ran. A day with more than one build is excluded by pre-registration — even when the result is flattering.

3. A limiter engaged. The daily cap of 8 trades was reached; from here on that is the bottleneck, not signal quality.

The ledger that makes any tampering visible

Since July 29 a running register writes the frozen null hypothesis into **every line** (100 / 422). If anyone changes it later, the change is visible in the version history.

Two built-in controls:
· **Placebo** — null hypothesis against itself: $p = 0.519$ ✓
· **Counter-direction** — 50 % at n = 120: $p = 3.8 \cdot 10^{-10}$ ✓

So the checker can do both: not fire too early, and recognize a genuine signal.

The target metric has changed

No longer the hit rate, but the **target-reach rate**. The reason is purely statistical: moving from 23.7 % to 50 % requires roughly **40 trades** for a defensible verdict — about **five trading days** instead of twenty.

THE OPEN CHAIN

Nine stages — the cascade has fallen, the endpoint has moved

The July research plan consisted of nine consecutive test stages with exactly **one** confirmatory endpoint. On September 1 stage 5 fell — and with it, as the pre-registration states verbatim, everything above it. That is not a breach of the plan but its intended outcome: the rule preceded the calculation.

STAGE	QUESTION	STATUS	FINDING · AS OF SEPTEMBER 23
0	Is the recording complete and unambiguous?	done	Gap scan per day, part of the evidence chain since August 24
1	Do book, trades and log agree?	done	Acceptance 10/10, ladder coverage 98.90 %
2	Can the aggressor be determined cleanly?	discarded	Placebo control: 0.21 pp of information
3	Does the microprice carry predictive power?	no	Half-life 90 ms — explicitly do not repeat
4	Does order flow imbalance measure the move?	holds	Median R^2 **0.893** (42 runs) — it *describes* the concurrent move almost completely
5	Does it predict the next move?	fallen	**Sep 1:** skill 0.995 against a hurdle of 1.25, 19 tape days, all four horizons — Chapter 24
6a	Do sweeps act as triggers?	retired	Revision 13 — moot without stage 5
6b	Does absorption at walls act?	retired	Revision 13; the wall question lives on in the highway test (Chapter 24)
7	Are stop hunts detectable?	retired	Revision 13; the price-target join survives as a tool
8	**Does the OFI entry gate improve the result?**	replaced	**New endpoint "Bracket-MinP" (Sep 5) — the look is still due**

Why exactly one endpoint remains confirmatory

The endpoint moved; it did not multiply. The new pre-registration asks: *does the fourth CORTEX head disagree more often on losing entries than on winning ones?* — 10 trading days, **two-sided** (a result below zero establishes a *harmful* gate and is worth the same finding), with a mandatory counter-check that excludes the first day.

The unusual part is the field **prior sighting**. Day 1 was known before signing (−4.8 percentage points, *against* the hypothesis) and is disclosed in the document, included in the fingerprint. The next pre-registration (dead zones) is drafted — its claim right is released only **after** this verdict. Two open endpoints would raise the family error rate without anyone noticing.

The brake — and its deliberate breach

Since July 30 the rule has been: no recompilation of the trading logic. Between September 1 and 4 it was **broken four times** (v0.9.5 to v0.9.9) — each time on the founder's instruction, each time documented in three places: the pre-registration field, the project status, and the source comment at the gate.

Since then the counter of homogeneous days has **never counted: 0 of 20** — the source changed on September 12, 15 and 22, each time on record. It is no longer a binding anchor.

On the state of play: the apparatus stands, the rules stand, the tools test themselves. What is running is **data collection** — since release 0.5.30 (September 23) once more from zero.

SCHEDULE

The road to a verdict — what gets decided when

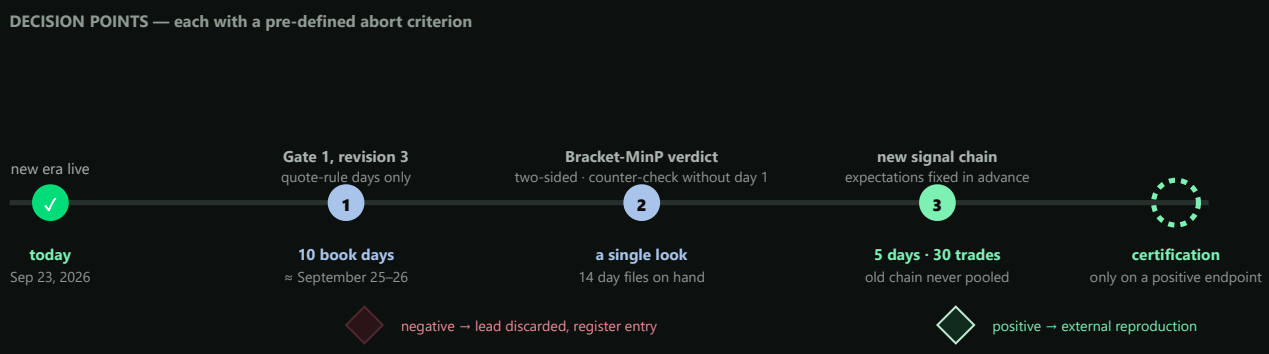


Figure 9: Every milestone has two exits, and both are defined in advance. A schedule without a defined failure mode is a wish list.

What happens if it comes out negative

The finding enters the retraction register, the lead is closed, and the next candidate on the list moves up. **The product is unaffected** — since July 2026 it explicitly sells no edge claim.

Cost of a negative outcome: a few weeks. Cost of an *unverified positive* outcome: the credibility of the entire company.

What happens if it comes out positive

Then, for the first time, a result exists that satisfies **all six rungs of the ladder of evidence**: pre-registered, placebo-controlled, tested outside the training data, checked against a null distribution, gathered under frozen conditions and **recomputable by third parties** from archived raw data.

From that point the certification process begins — described in Part V.

Why the schedule is counted in trading days rather than weeks: a calendar date can be met by lowering the requirements. A counter of homogeneous trading days can only be satisfied by waiting — **and it resets to zero on any change.** That is the only form of scheduling immune to one's own impatience.

One consequence worth stating plainly: between today and the endpoint there is nothing left to build. That is unusual for a software company and uncomfortable for everyone involved — the natural instinct is to improve something while waiting. Every such improvement, however, resets the counter and pushes the verdict further away. **In the first week of September that discipline was broken four times, deliberately and on record** — the price is a counter back at zero, and from September 8 the most valuable activity is once again doing nothing.

FIVE DAYS IN SEPTEMBER

Three null findings in five days — and what they prevented

Between September 1 and 5, three pictures that felt right fell against their time-of-day-matched null line. None was a setback: **each prevented a rebuild** that would have cost weeks and devalued data. All three were pre-registered; all three are in the retraction register.

PICTURE	BASE SET	MEASUREMENT AGAINST CONTROL	PREVENTED
Order-flow direction (stage 5) OFI sign → next mid move	19 tape days 77,972 windows (5 s)	Hit rate **49.76 %**, skill **0.995** (CI 0.986–1.004) against a hurdle of 1.25 — on all four horizons, bootstrap over days	the planned OFI entry gate; per the pre-registration the whole cascade 5–8 falls with it
Liquidity marker (LP) "almost always shows at the turn"	693 markers 19 days	≥ 4 ticks within 60 s: **50.6 %** against 49.9 % control (95th percentile 65.0 %) — all four cells fall	a 42nd feature for CORTEX — and with it wiping 27 sessions of learned state
Trend continuation the "highway" model, eleven cells	19 days 436,088 RTH seconds	continuation after a pit stop **46.5 %** against 58.8 % — better on **0 of 19 days**; walls hold (42.9 % against 49.5 %)	a phase module in the strategy; instead a fade pre-registration, stage A running
Exit on counter-signal **"leave immediately on a counter-signal"**	**61 trades 16 days**	**every rule beats the actual result (+$1,067 to +$3,582) — and loses against random exit times (−$296 to −$2,811)**	**an exit knob that only looked good because 79 % of trades end at the stop**

What holds: stage 4 — and a signature

The same instrument that does *not* predict direction measures the *concurrent* move almost completely: median R^2 of **0.893** across 42 runs. The pre-registration calls the confusion of describing with predicting "the most common and most expensive error in this literature" — Chapter 17 turned it into a test criterion.

All three null findings carry the same signature as the timing axis of Chapter 26: **flow ⇒ reversion** within 30 to 600 seconds, with a lead. That is not confirmation — it is the reason the fade pre-registration was signed *before* anyone turned a threshold.

The fourth head — misjudged twice in one evening

In early September CORTEX received a fourth prediction head ("does the trade reach 8 ticks before it loses 10?"). Its first metric read like "twice the base rate": accuracy 0.631 next to a positive rate of 0.305. The correct comparison line for accuracy is **0.695** — the head sat **below** it. Its Brier score of 0.448 looked poor against 0.21; its naive line is 0.408.

Both errors are in the register. Consequence: since app 0.5.9, metrics are written **together with their comparison line**, not beside it. And before any head is judged, the question is: **who reads it at all?** In the strategy at the time: nobody.

Six cases of one class in two days: a reader discarded 2,941 usable rows because the stream wrote "UP/DN" and it expected "LONG/SHORT" — exit code green. A bridge reset a new field to zero on every message (one trading day, −$448). A column had never been filled for weeks (0 of 27,979 rows). In all six cases something reported "nothing" where "does not fit together" would have been right. Since September 5 a guard checks the value inventory of every data stream against the values its readers branch on (evening chain, step 7) — positive control run on the real prior state.

THE MEASUREMENT LIMIT

The delta microscope — 56.7 % of volume has no certain sign

Every delta display on the market claims to know who the aggressor of a transaction was — buyer or seller. In August we measured **how often two equally legitimate classification rules disagree on the same raw data.** The result changes what an "accurate delta" can even mean.

The finding

Two rules, one raw data stream, 14 days

Every transaction from 14 fully recorded days was classified **twice**: once by the tick rule (comparison to the last price), once by the quote rule (position relative to bid/ask). On **56.7 % of traded volume** the two contradict each other — the sign of those shares depends on the *choice of rule*, not on the data.

The disagreement is not noise — it has a shape

Sorted by the age of the most recent quote, the contradiction rate falls **monotonically from 51.6 % to 3.2 %**. The fresher the quote, the more the rules agree. The uncertainty is therefore **measurable, explainable and quantifiable per transaction** — a property of the data stream, not a vendor's bug.

The uncomfortable side finding

The measurement also revealed that **our own software ran two delta paths with two different rules** — chart and heatmap could give the same moment different signs, and nobody knew. The finding is in the retraction register; unification is part of the next shipped build.

The consequence

"The most accurate delta" is an unprovable claim

If more than half the volume changes sign depending on the rule, **no** vendor can claim to show "the correct" delta — on this data there is no observable ground truth. What *can* be proven is something else: a delta that **declares its own error band** — one that says, for every value, which share of it is rule-invariant and which hangs on a convention.

That is precisely the product property that follows from this measurement: **not "the most accurate", but "with declared uncertainty"** — a property no competitor offers today, and one that cannot be offered without this measurement.

And the named product risk

Customers compare our display live against vendors that silently run the *other* rule. On days with many stale quotes the two displays can point in opposite directions — **both consistent with the raw data**. Without a declared band that looks like a defect. With it, it is a measurement.

Why this matters for every study downstream

Any research keyed on the delta sign — including several stages of our own evidence ladder — inherits this uncertainty. A study that treats the sign as ground truth on the ambiguous 56.7 % is measuring its own rule convention, not the market. The disagreement is logged per transaction from the archived tapes, so every downstream aggregate can carry its error budget instead of hiding it. That is also why the directional anchors of Chapter 39 exist: a conclusion resting on an unsecured sign is not a conclusion.

The number that remains: No vendor can make the 56.7 % smaller — the raw data does not give more. One can only **declare it or conceal it**. We chose to declare it, because it is the only verifiable position.

THE SECOND SEARCH ROUND

Four empty reader classes — and one axis that holds

Gate 1's "NOT LEARNABLE" explicitly applied only to a **linear, pointwise reader**. The obvious question: do stronger model classes see more on the same data? In late August it was answered systematically — with a twofold result.

The stronger readers' answer: empty

1. **Four reader classes** — from tree ensembles to a sequential network — on the identical data space (12 features, 8 steps, 600-second horizon): **none delivers a replicable signal.**
2. **The one "LEARNABLE" hit refuted itself.** A recurrent network initially reported learnability — the preregistered replication then measured **AUC 0.504**: a coin flip. Without the replication requirement this would have been celebrated as a breakthrough.
3. **Gate 1 now stands broader than ever:** NOT LEARNABLE at **effN 2,279** on **22 of 22** citable recording days (verdict of August 27, signed off on the 28th). The upper bound from Chapter 23 remains valid unchanged.

What the verdict now covers — and what it does not

Covered: linear *and* the four tested stronger classes on the pointwise feature space. **Not covered:** other feature spaces and other time resolutions — exactly where the weekend study of the next chapter kept searching.

The side finding with substance

The timing axis: flow ⇒ reversion in the mid window

Across the reader classes a directional structure appeared: **strong aggressive flow predicts a counter-move in the 30–600 second window** (mean reversion), not continuation. The effect survives the bounce control, the circular shift and the per-day split (**9 of 19 days** individually significant, no day opposite) — and our own recorded trades are convergent with it (correlation −0.27, n = 48).

Status: candidate — with a date

The effect is measured **without trading costs** and not yet replicated. It therefore carries the status **HYPOTHESIS CANDIDATE** with a preregistered replication window around **September 9**. Only if it reappears there does the axis become a template for the confirmatory path — until then it is an observation.

The difference between "discovery" and "candidate" is a date: that of the preregistered replication. Anything that does not wait for that date is a story about the past — this project has had enough of those.

Why the reader classes' null is valuable: It rules out the cheap explanation that we merely used the wrong model. If the next chapter finds a candidate, it is **not** because a stronger model could read more into the data — but because a *different data space* was opened. That distinction decides where search effort goes next: into resolution and context, not into ever-larger models.

THE WEEKEND STUDY

Fifteen million configurations — one mirage and one candidate

On August 29 two exhaustive searches ran across the entire recorded data inventory. The first proved that our validation funnel **would let a real edge through**. The second found a candidate that passes every first examination — and along the way exposed a second one that passed none.

Run 1 — the book at one-second resolution

631,800 configurations, zero finalists

Six signal families × a fine exit grid, exhaustive rather than sampled. Discipline declared up front: chronological 13/4/4 split, **test block touchable only once**, plateau requirement, 10,000 permutations. Result: **no book feature carries anything beyond pure price action in this space.**

The positive control — proof that the null counts

A known edge was planted into 21 synthetic days and sent through the same funnel: **582 → 60 → 49 → 20 finalists, 20 of 20 pass the one-shot test** (separate ledger, removed afterwards). The funnel lets real edges through. **Run 1's null is therefore a statement about the data — not about the tool.**

The mirage

One cluster ("quote pull") looked brilliant in validation: **+$51 to +$64 per trade**. The one-shot test block showed **−$10 to −$78** — only 1 of 12 exit variants positive, a knife's edge. Without the one-shot discipline exactly this would have been celebrated as a discovery. **That is the apparatus at work.**

Run 2 — the raw tapes at quarter-second resolution

14.9 million configurations on 103 million events

A 250-ms grid built directly from the raw quote tapes (21 sessions — including three days the app stream had lost, carried gaplessly by the indicator's recording), conservative limit-order fills, 16 context gates, confluence pairs, iterative refinement — **81.7 minutes of compute**, declaration before the run.

The survivor: "flush absorption"

Price plunges ≥ 12 ticks in 10 s **while** the 2-second order flow shows strong aggressive *buying* → enter long via limit order into the plunge. Training **+$17–23** per trade (n ≈ 111) · validation **+$15–19** · one-shot test **+$27–43 at a 75 % hit rate — all 7 exit variants positive** (a plateau, not a spike), p ≤ 0.003 over 10,000 permutations. Microstructurally coherent: the classic absorption bounce — and the confluence form of the timing axis from Chapter 26.

The honesty that goes with it: The test block carried only **n = 16** trades (Wilson lower bound ~51 %). The status is therefore not "edge" but **HYPOTHESIS CANDIDATE with a passed first examination** — frozen (any change = a new candidate, counter reset to zero) and since August 29 in a **daily forward test** on every newly recorded session. Threshold: **n ≥ 50 and Wilson lower bound ≥ 55 %. Addendum, September 23:** the forward test has decided against the candidate. At n = 90 it hits 65.6 % (Wilson 55.3 %) — and loses **$5.17 per trade**. Discarded since September 18. On September 14 the threshold was briefly met (n = 52); the rule knows no stopping at the first crossing. Lesson: a hit rate without an expected value is not an edge.

TEN DAYS IN SEPTEMBER · I

The replay machine — would the strategy have won the other way round?

On September 9 and 10 all **105 replayable strategy trades** since July 29 were replayed tick-exact on our own tape — with the real price path, stop, target and costs (0.64 ticks per contract). Four pre-registrations, one look per variant. The question behind it was uncomfortable: **is the fault in the management — or in the entry itself?**

QUESTION	VARIANTS	NET TICKS PER CONTRACT	VERDICT
Calibration original, unchanged	the actual rule set	**−1.76** · hit rate 35.2 %	reference
Reversed direction "trade it the other way"	every entry mirrored, own stop	**−1.73** with 1 tick of slippage (−1.04 without) · 39.0 % against break-even 50.4 %	does not hold
Structure time stop, T1, break-even lock	no time stop · T1 = stop · no early lock · all three	**−1.36 to −1.72** — all four lose	does not hold
Scale are 8/10 ticks too small?	16/12 and 24/12 ticks, each with and without lock	**−2.01 · −1.56 · −1.39 · −0.89** — all below break-even; features AUC ≈ 0.50	does not hold
Scaling brackets **in fast markets**	**factor 1.5 and 2.0 at full throttle**	**−1.86 · −1.87 against −1.64; at full throttle itself worse (−1.46 against −0.11)**	does not hold

What follows from it

The entries themselves carry no edge. With a reward-to-risk ratio close to 1, an entry without information is a coin toss minus costs — and that is exactly what every variant looks like. No management, no scale and no direction rescues it.

Consequence: the strategy was **not** reversed, the brackets were **not** enlarged. The work shifted to the question of *who* sets the direction — and whether the measured quantity underneath is right at all (Chapters 29 and 30).

The trap the apparatus caught

The quick approximation "flip the sign of the results" gave **+0.95 ticks** and a 57.3 % hit rate — a reversal strategy would have felt like a discovery. But the mirrored trade has its own stop, and it sits exactly where price ran beforehand. On the real price path nothing of it remained. Retracted on September 10 — before a single line of code had changed.

Why a replay can only kill, not support: the variants were derived from the exit reasons of the original — that is, from the outcome. The pre-registration therefore explicitly calls the run *descriptive*: a positive result would only have been a reason for a forward test of the same variant, not evidence. A negative result for *every* variant, by contrast, says exactly one thing — and says it unambiguously.

TEN DAYS IN SEPTEMBER · II

The exchange as referee — 97.5 % instead of 69.9 %

Since September 10 we have had **real CME order data** — every single order with its side. For the first time it was possible to check whether our rule, which assigns an aggressor to every trade, is right: not against itself, but **against the exchange**. Every quantity derived from the delta — absorption, exhaustion, CVD, divergence, order flow — depends on it.

QUESTION	SAMPLE	RESULT AGAINST EXCHANGE TRUTH	CONSEQUENCE
Which rule identifies the aggressor?	7 days 94–96 % coverage	quote rule **97.5 %** of volume correct, tick rule only **69.9 %** — daily delta with the wrong sign on 4 of 7 days	product decision: delta on the quote rule
Locked quote bid = ask	7 days	neighbouring quote hits 40.2 % and 14.3 % — *below* chance; explains 104 % of the residual deviation	"side not determinable" instead of guessing
Absorption and exhaustion real or rule artefact?	7 days	signs agree only 74.1 % of the time; the tick version reports exhaustion in 41–49 % of seconds, the truth in 34 %	largely an artefact
Does *true* exhaustion carry?	3 days 7,723 events	AUC 0.510 against 0.510 — the correct rule makes it true, but not useful	no signal lost
Big prints as an AI feature? side of the ≥ 50-lot prints at the entry level	30 days	AUC 0.503; above the null band on 12 of 30 days; the sign flips from day to day	stays a display, not a feature
Confirmation day after deployment	**Sep 11 780,123 fills**	**product = recomputation on 100 % of fills · 97.0 % correct · per minute r 0.985**	**shipped: indicator, app, AI**

Why this is the most important finding of the month

Almost a third of the volume stood on the wrong side. Every study of the tick era computed on this basis — that many of them came up empty may partly come down to exactly that. This proves nothing in the other direction, but it enforces a rule: **old and new days are never pooled.** Gate 1 therefore restarted on September 12, on days under the quote rule only.

The error before — measured circularly

In July the project recorded that the tick rule beat the quote variants. The number behind it (r = 0.965) measured the agreement of the tick rule with the indicator's *own* fields — an instrument reading itself. Against the exchange: quote r = 0.983, tick r = 0.677. Retracted on September 10, in the same work step.

The consequence, with a price tag: the new rule changes the meaning of the book features CORTEX learns on. A model that learns on two definitions learns their difference. That is why the learned state was restarted — for the founder on September 11, for everyone with release 0.5.30 on September 23.

THE NEW ERA

ONE CORTEX decides — and first we measured what slows it down

Between September 14 and 22 the signal chain was rebuilt: the direction comes **from CORTEX alone**, a CORTEX counter-signal can close the trade before the first target, and the strategy's own filters now only log. On September 23 the new era went out to all customers as **release 0.5.30** — with a learned state that starts at zero.

MEASUREMENT	RESULT	CONSEQUENCE
Feature inventory Sep 22, before the restart	**6 of 49** inputs practically dead (thresholds above the real distribution); no feature separates up from down on its own (AUC 0.47–0.51)	pre-registered, not silently fixed
Signal chain book → signal → order	three throttles (3,000 / 1,500 / 2,000 ms) ⇒ **3,290 ms**; afterwards **≈ 208 ms** — computed from cycle times, not measured end to end	throttles removed; measurement point at entry still missing
Compute time	45 ms on the cold model — but **348 ms** in live trading, because training ran inside the cycle	training in its own process: warm ≈ 200 ms, target < 100 ms missed
Entry	CORTEX ≥ 55 % · book ≤ 200 ms old, imbalance ≥ 10 % in the same direction · target-odds gate only from 3,000 labels	remaining filters in shadow only
CORTEX exit	counter-signal ≥ 55 % for ≥ 300 ms before T1 ⇒ flatten; first day: 1 of 4 trades	verdict after 5 days and 30 trades
HUD events **four pre-registrations**	**reversal signal 57.0 % (lower bound 46.8 %) against 48.2 % · takeover 58.5 % (n = 41) · absorption price holds 48.6 % against 46.5 % · strongest events 45.7 % against 50.1 %**	**all "does not hold" — they remain displays, with an honest legend**

The breach — documented three times

The rebuild happened deliberately *before* the verdict of its pre-registration — on the founder's instruction, recorded in the pre-registration, the project status and the source code. The expectations were fixed beforehand: more than 6.6 entries per day, exit share 10–30 %. **Not built** was what had measured too weak in advance: a bracket width from the bracket head (AUC 0.54 over 28,986 samples).

What the restart costs — and why it is right

With 0.5.30 every model starts at zero — at the customer's as at the founder's. No pretrained model, no third-party trades. The price is the learning history so far. The gain: a model that learns on **one single definition, checked against the exchange** — and a result that then clearly belongs to CORTEX, not to a rule set around it.

Two findings a customer could have seen: the spoof detector was on in practically every second — it measured a direction, not a share; the spoof gate went into shadow. And the wall tracker counted cancellations as absorption. Both surfaced while pre-registrations were being drafted, before the actual question had been computed.

CERTIFICATION

Why certification is possible — and who could issue it

The question is not "is there an authority that certifies trading advantages?" — there is none. The question is: **which verifiable properties must a result have for established audit bodies to be able to attest to it?** To that there are clear answers.

Four real routes, each with a clear scope

ROUTE	WHAT IS ATTESTED	PRECONDITION — AND WHETHER IT IS MET TODAY
Public accountant Agreed-upon procedures (ISAE 3000 / AT-C 105)	That a defined computation on defined raw data yields **exactly the stated result**.	**Achievable.** Raw data is archived, every tool is open within the project, and every result carries a SHA-256 fingerprint over both result *and* data provenance. An auditor recomputes — if the fingerprint matches, nothing was altered after the fact.
Academic review peer review / preprint	That the **methodology** meets the state of the art and that the conclusion follows.	**Achievable.** Pre-registration, permutation null distributions, cross-validated robustness and placebo controls are precisely the elements reviewers examine. The negative result is no obstacle here — **published negative findings in this field are rare and in demand.**
ISO/IEC 42001 AI management system	That the **development and testing process** for the AI system is traceable, documented and monitored.	**Largely prepared.** Change tracking, a risk register, release criteria and a procedure for refuted findings already exist. Formal role and responsibility definitions would need to be added.
Independent recomputation third party, raw data + protocol	That an **outside team** obtains the same result from the same raw data.	**Achievable.** This is the strongest form and the actual target state (rung 6 of the ladder of evidence). It requires that the raw data can be shared — which is the case for self-recorded market data.

What is explicitly *not* certifiable

Future returns. No body anywhere attests that a method will work tomorrow. Anyone offering that is not serious.

What is certifiable is always only: *this method, on this data, under these pre-specified conditions, produced this result, and it was computed correctly.*

But that is precisely the decisive difference from everything the market offers today.

The three properties that make it all possible

1 · Tamper-evident. Every result document carries a cryptographic fingerprint over result and data provenance. Recomputing yields the same hash — or exposes a change.

2 · Reproducible. Raw data, tool and parameters are archived and versioned.

3 · Pre-registered. The criteria were fixed before the data was seen — demonstrable via dated entries in the version history.

The core in one sentence: certifiability does not arise from a good result but from a verifiable **path** to the result. That path exists in full at OrderFlowAi today — it is merely waiting for something positive to stand at the end of it.

CERTIFICATION

The dossier — what already exists today

A certification dossier cannot be assembled retroactively; it has to **run alongside**. That is why the chain of evidence was built as a machine from the outset rather than as a document. This page lists what an auditor would find today.

1 · The edge validation certificate

A program generates a complete document from the current study — as text, as styled HTML and as PDF. It contains:

· **Data provenance** — which trading days, how many, per market phase · **Feature health** — which inputs carried information at all · **The pre-registered verdict** under a specified clause · **Cross-validated robustness distribution** (CPCV) · **p-value against the null distribution** · **Leakage check** across a parameter range · **Scope caveats** — what the statement does *not* cover · **SHA-256 fingerprint** over result and provenance

2 · The strategy change certificate

The counterpart for rule changes: expectancy before versus after, trade counts, bootstrap significance of the improvement, a full census of all suppressed signals with an approximate counterfactual — **and the same caveats and the same fingerprint.**

3 · The data integrity certificate with history

A daily traffic light across six independent quality guards — with a maintained **history**. The reasoning behind it is the important part: if an analysis looks odd in August, the question will be "what was the data quality on the days that fed it?" Without a dated record that cannot be reconstructed after the fact.

4 · The running ledgers

Append-only registers, each carrying the frozen null hypothesis in **every** line:

· **Target-reach ledger** — the current lead
· **Head ledger** — every prediction head with its verdict
· **Learning ledger** — model age and performance per session
· **Gate ledger** — every filter rule with its measured effect
· **Recording ledger** — completeness per trading day
· **Book coverage ledger** — usable days

5 · The retraction register

Fifty-four entries, machine-monitored. For an auditor this is **the most valuable part of the entire dossier** — it is the evidence that the system can not only make errors but find and withdraw them.

6 · The evening protocol

Twenty-two steps, every trading day, automated. From verifying the fingerprint of the installed software through recording completeness to the ledgers and the nightly study.

Seventeen of the 22 steps carry a domain acceptance test — they check not merely *whether* something ran but *whether the result is right*. The missing five are named and deliberately declared open rather than calibrated away.

Generated, not written

Every document above is produced by a program from live artefacts. Nobody assembles it by hand, so nobody can quietly omit an inconvenient line.

Append-only

The ledgers only ever grow. A revised figure appears as a new dated line beside the old one, never in place of it.

Fingerprinted

Result and provenance are hashed together. Recomputation either reproduces the hash or exposes the change.

The difference from a conventional backtest report: a backtest report is a snapshot somebody produced. This dossier is a running process that nobody can halt without it being noticed. **It cannot be dressed up — it can only be switched off, and that would be visible.**

CERTIFICATION

The claim to being first — what exactly would be "the first time"?

A claim to primacy must be phrased precisely or it will not survive scrutiny. This page phrases it narrowly enough to be defensible — and states what it does **not** assert.

The defensible formulation

"The first commercially available order flow analysis software for retail users whose predictive component has been subjected to a pre-registered, placebo-controlled and independently recomputable test procedure — with the result published, regardless of whether that result was positive or negative."

What this claim includes

- **Pre-registered** — criteria dated before the analysis
- **Placebo-controlled** — with a declared control arm
- **Recomputable** — raw data + tool + fingerprint
- **Published** — including on a negative outcome
- **Commercially available** — not a research prototype
- **For retail users** — not institutionally walled off

What it explicitly does not assert

- **Not**: "the first AI in trading" — that would be false
- **Not**: "the first with an advantage" — unproven
- **Not**: "the most accurate" — not comparatively measured
- **Not**: any statement about future returns
- **Not**: that institutional firms do not do this internally — they probably do, only **not verifiably in public**

Why the narrowness of the claim is its strength

A broad claim ("the world's best AI") is worthless because it is unverifiable and collapses at the first critical question. A narrow claim is defensible — and therefore **citable**: in trade publications, in sales conversations, before regulators, in a due diligence.

Moreover, the narrow claim is already **half fulfilled**. The apparatus exists, the tests are running, the results are documented — including the negative ones. What is missing is only the completion of the current test cycle.

The historical parallel

In medicine, introducing pre-registered trials did not produce better drugs. It produced, **for the first time, knowledge of which ones work** — and the market cleaned itself up within a few years.

Retail financial software has yet to take that step. Whoever takes it first defines the standard by which everyone else is subsequently measured.

Demonstrable today

Pre-registration dated in the version history · placebo controls as mandatory lines · a fingerprint per result · 22 published retractions

Demonstrable after the confirmatory endpoint

The confirmatory endpoint under frozen conditions — positive or negative, citable either way

Demonstrable after external review

Recomputation by a third party from the same raw data — rung 6 of the ladder of evidence

The strategic point: even if the current lead ends negatively, the claim to primacy stands — it rests on the **procedure**, not on the result. That makes it the one part of this project that cannot fail.

THE EVOLUTIONARY STEP

The autonomous co-founder — what happens when the AI thinks along instead of executing

This project is built by **one person** and **one AI agent**. In 2026 that alone is no longer remarkable. What is remarkable is the operating mode: the agent is configured not as a tool but as a **co-founder under an obligation to take initiative**.

The four rules of the mode

1. **Context routing.**
   Depending on the area being worked on, the agent autonomously loads the relevant domain knowledge — market microstructure and quantitative methods for the analysis engine, web architecture and conversion optimization for marketing.
   Without being asked.

2. **Architect rather than executor.**
   The agent draws its own conclusions from the project state and searches continuously for a statistical or technical advantage — including when the task at hand was something else.

3. **A mandatory block at the end of every answer.**
   Every substantive unit of work ends with a fixed section titled `[Autonomous Project Evolution]`, which *must* contain two things: a concrete improvement impulse for the area just worked on, and a
   feature proposal
   aimed at the long-term goal.

4. **No code without approval.**
   Proposals are implemented only after an explicit "go." The human retains the decision; the AI takes over thinking ahead.

Why rule 3 is the actual lever

An agent that only processes instructions never thinks beyond the instruction. The enforced closing block inverts this: **at the end of every unit of work, one step further must have been thought.** Across 395 sessions that yields several hundred unprompted improvement proposals — a substantial share of which flowed into precisely the tools described in Part III.

The retraction register, the mandatory placebo line, the scope guard and the acceptance inventory **all** originated in that block — none of them was an instruction.

The persistent memory

The agent maintains a file-based knowledge base, today holding **808 entries** — user preferences, project decisions, technical traps, refuted assumptions, each with its reasoning and cross-references.

The effect: **an error is made at most twice.** The second time produces an entry, and the third time an automated checker fires.

This is also the basis of the self-correction capability documented throughout this report.

The self-binding that grew from it

The direction of development is notable. Across sessions, the agent has imposed **increasingly strict rules** on itself:

· no verdict before the counter-check
· every rate only with a null hypothesis
· every metric only with a decay curve
· every guard only with a self-test
· every trap immediately paired with an automated checker

None of these rules was specified externally. All arose from analysing our own errors — and all make the work slower and the result more defensible.

The honest limitation

The same agent also produced the 54 retracted findings. The operating mode generates **more** hypotheses — correct and incorrect alike. Its value lies not in a higher hit rate but in the fact that the verification machinery grows with it.

THE EVOLUTIONARY STEP

The compounding effect — measured, not asserted

The thesis: because the project is co-developed by an AI agent, its development speed and quality grow with every model generation. That thesis is testable — and we measured it against our own version history.

Figure 10: Newly created analysis tools per month, counted from the version history. The green portion is tools with a built-in self-test.

The honest interpretation

These numbers do **not** demonstrate a model effect alone. July also brought the shift in method: after the break (page 11) the work moved from product building to research, and research inherently requires more measurement tools.

What the numbers do demonstrate: the move from 0 % to 45 % self-test coverage happened *without* additional people, *without* more time and *without* external requirement. It happened because the rule "every guard needs a self-test" emerged during that month — and was then applied consistently.

Why this compounds

The effect is non-linear, because each new tool raises the **error detection rate** of all future work. An example from July:

The scope guard found an error → that produced the acceptance inventory → which found five steps without domain verification → one of which revealed that the installed software had never been verified against its checksum.

One tool produced three more — and along the way exposed a project rule that had been in force for a year and never measured.

What this changes about the role model. The bottleneck of a research company used to be the number of people who can think carefully at the same time. Shift part of that into an agent that does not tire, forgets nothing and turns every error into a permanent automated checker, and the bottleneck moves to something else: **the number of trading days that must elapse.** That is exactly where this project stands today — it is not waiting on development work, it is waiting on the calendar. That is a good place to be waiting.

The forecast, clearly labelled as a forecast: if each model generation raises the quality of autonomous contributions, the limit of what a one-person company can achieve in research depth moves further out. The trajectory measured so far is consistent with that expectation. **It does not prove it** — two months are not enough for that. But precisely this distinction between "consistent with" and "proven" is why this report carries weight at all.

VALUE

The value chain — five stages, four of them independent of the edge

The decisive strategic move of July 20, 2026 was **decoupling**: the product sells what demonstrably works, and the research runs separately. As a result, company value does **not** hinge on the outcome of a single experiment.

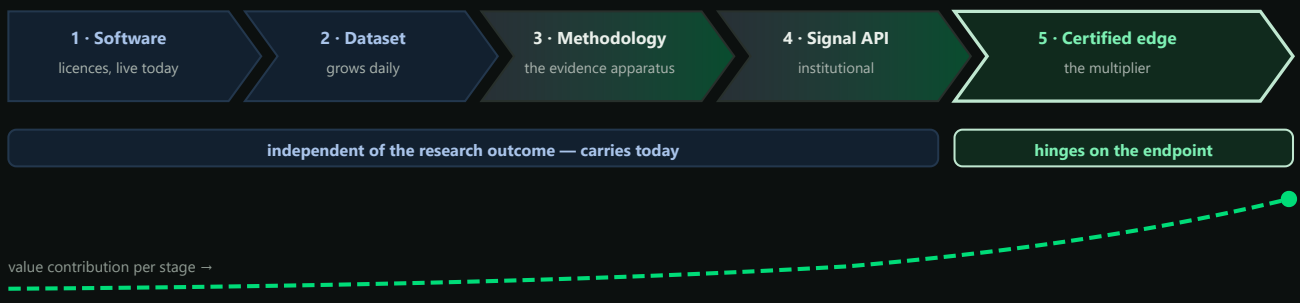


Figure 11: Four of five stages carry regardless of whether the edge proof succeeds. Stage 5 is not a foundation but a multiplier.

STAGE	WHAT IS SOLD	WHY IT CARRIES	STATUS
1	**Software licences** Pro · Trial · Institutional	Level 2 visualization, heat map, footprint, DOM analysis, trading journal. Customers buy **visibility**, not prediction — and that visibility demonstrably works. An official NinjaTrader ecosystem listing serves as external confirmation.	live
2	**The dataset**	Fully recorded order book and trade data with per-line provenance, continuous since 2025. **Not obtainable retroactively** — its value grows with every day a competitor fails to record.	growing
3	**The methodology**	The evidence apparatus itself — 170 tools, pre-registration, certificate generation, retraction register. Transferable to any quantitative question. Conceivable as a licence, a consulting service, or a standalone product.	option
4	**Signal interface**	Machine-readable access to the analysis engine for institutional customers. Technically prepared; commercially sensible only once there is a defensible proof.	prepared
5	**The certified edge**	**On a positive endpoint: a substantiated, externally recomputable statement about a statistical advantage. The price of a tool that demonstrably works is not the price of a tool that merely looks good.**	open

Decisive for valuation: an investment in OrderFlowAi is not a bet on stage 5. It is the acquisition of an operating software business with a proprietary, non-replicable dataset and a research apparatus that is exploitable as a standalone asset — **with a free option on stage 5 on top.**

VALUE

Risks, stated honestly

A report that documents 22 of its own errors and then falls silent on risk would lack credibility. This page names what can go wrong — and what has already been done about it.

RISK	SEVERITY	ASSESSMENT AND MITIGATION
The edge does not exist	high	This is the single most likely outcome — the evidence to date is predominantly negative. **Buffered by the decoupling:** four of five value stages carry independently of it. A negative outcome costs weeks, not the company.
Key person dependency	high	One founder, one agent. This is simultaneously the strongest accelerator and the greatest vulnerability. **Partly buffered** by unusual documentation depth: 808 knowledge entries, every decision with its reasoning, every trap with an automated checker. A new team would find an unusually well-handed-over state — **but that does not replace a person.**
Time required by the test chain	medium	The confirmatory endpoint requires 20 trading days *under unchanged conditions*. Any change resets the counter. This is methodologically compulsory and practically inconvenient — **and it is exactly the discipline that makes the later proof defensible.**
Market change invalidates findings	medium	An advantage that exists today may vanish tomorrow. **Addressed** by the stationarity test, which checks forwards *and* backwards, and by the principle of reporting findings together with their half-life.
Competitors copy the methodology	low	Technically possible, practically unlikely. The methodology is uncomfortable: it requires withdrawing your own success announcements. A vendor under sales pressure will not do that. **The competitive advantage is cultural, not technical** — and therefore hard to copy.
Regulatory requirements	low	The software is an analysis tool with no investment advice and no asset management. Risk disclosures are implemented across all customer-facing surfaces. The EU AI regulatory framework has been taken into account; the management system is prepared for ISO/IEC 42001.
Data quality at the instrument	low	Historically the most frequent source of error — three of the last five retractions. **Today the most heavily safeguarded area:** per-line provenance, duplicate detection, completeness checks before market open, and a daily integrity certificate with history.

The largest residual risk in one sentence: it is not that the edge does not exist — the company is prepared for that. It is that someone shortens the test chain out of impatience and carries an unverified result outside. **The entire apparatus described in Part III was built against exactly that** — including the rule that any change to the criteria remains visible in the version history.

On how these severities were assigned: they are judgements, not measurements, and are labelled as such. Where a measurement does exist it is cited in the right-hand column — the predominantly negative evidence base, the 28 % power of the locked stage, the three-of-five retraction share for data quality. Where none exists, the entry is an assessment. **That distinction is itself part of the method:** a risk table that presents opinion in the same typeface as evidence is exactly the kind of document this report argues against.

POSITION

Where we stand today — in numbers, not intentions

This chapter is the compass. It answers three questions: **Where do we stand? What is missing before a verdict?** And: **what does the strategy base its decisions on?** All figures are measurements taken on September 23, 2026; where an older measurement point applies, its date is stated.

The four running chains of evidence

1. The confirmatory endpoint — "Bracket-MinP"

Signed September 5, two-sided. 14 day files are in, the hurdle of 10 days has been reached — **the one pre-registered look is still outstanding**. Context: the calibration of September 10 showed that the head does not separate (top decile predicted 78 %, actual 33 %); since September 23 it learns from zero. The anchor has **never counted** (0 of 20).

2. Gate 1 — the order book as a second input, revision 3

Restarted on September 12, **only on days under the quote rule** (Chapter 29). Status: **8 of 10** book days, verdict around September 25–26. The previous version read **NOT LEARNABLE** on 61 days of the tick era — it is no longer pooled.

3. CORTEX — the learning model

ONE CORTEX: 49 features, five heads, since release 0.5.30 **solely** responsible for the direction and since September 23 with a fresh learned state. The last defensible verdict on the old generation (July 29: **NO EDGE**, MCC 0.0201) remains the baseline. Measured before the restart: 6 of 49 inputs practically dead (Chapter 30).

4. The shadow filters

They report but do not intervene. `DELTA-MISMATCH` has grown stronger: flagged entries win **29 %** (23 of 80) against **50 %** for unflagged ones, `p = 0.0001`, 27 days. The figure is **exploratory** — it pools several versions and the switch to the quote rule. No bound has been registered.

The apparatus underwriting all of it

Tools with self-tests (348 green)
350
Individual assertions
4,829
Pre-registrations
43
Self-retracted findings
54
Working sessions
471

What changed since the last edition (September 5)

The replay machine (Chapter 28): 105 trades replayed tick-exact — reversal, structure, scale and scaling brackets all lose. The entries carry no edge.

The exchange as referee (Chapter 29): the quote rule identifies the aggressor 97.5 % of the time, the previous tick rule 69.9 %. The product has been switched since September 11.

The new era (Chapter 30): CORTEX decides alone, the signal chain is cut from 3.3 s to about 0.2 s, and the learned state starts at zero for everyone. And the last candidate from the weekend study has been discarded by rule (Chapter 27).

The uncomfortable line. The *confirmatory* headline result remains **negative**, and all three forward candidates have since been discarded. What is new is not a find but the measurement basis: **for the first time the delta has been checked against the exchange**.

THE ROAD TO A VERDICT

What is missing — and what the strategy decides on now

The programme allows **two admissible outcomes**, and both count as a result: **A — a demonstrated edge. B — a defensible no.** A third outcome ("we still don't know") would be the only real failure. This page states what each of the two still requires.

Outcome A — the edge is demonstrated

1. **The one look at "Bracket-MinP"**
   — the days are in. After that the claim right of the next pre-registration is released; never two open endpoints at once.
2. **`DELTA-MISMATCH` gets a registered bound**
   and is tested on new days — only under the quote rule, so that the strong signal is not assembled from two different measured quantities.
3. **Independent replication**
   by a third party on the raw data. The dossier for it exists: pre-registration, retraction register, 4,829 self-tests, every figure with its measurement point.

Outcome B — the defensible no

1. **An upper bound, not a null finding.**
   "No edge found" is worthless without stating which edge would still have *fit*. For order-flow direction it exists (skill below 1.016); for Gate 1 it is being recomputed under the quote rule.
2. **The scope must be explicit.**
   The "no" now covers the pointwise feature space, order-flow direction, trend continuation — and since September also direction, structure and scale of the entries (replay).
3. **The measured quantity must be right.**
   Since September 11 the delta has been anchored against the exchange (97.0 %) — the largest single sign safeguard in the programme.

What the strategy actually decides on now

Since strategy 0.9.14 (September 22) the chain is short:

CORTEX ≥ 55 % → book ≤ 200 ms old and aligned (imbalance ≥ 10 %) → target-odds gate (only from 3,000 resolved labels) → risk gates: quiet market, 12:00–14:00 ET, 300 s minimum spacing, order flow against the signal, full throttle.
Exit: CORTEX exit · time stop (120 s before T1, runner 600 s) · flat at 15:40 ET.

On September 4 the bottleneck was still the rule set *downstream* of the model — the inventory counted 120 blocks per entry. Now the model decides, and it starts at zero. That is the more honest experimental set-up: **a result then clearly belongs to CORTEX.**

The order in which we harden

1. Secure the recording — without citable days every analysis is worthless. done
2. Check the measured quantity against the truth — a wrong delta invalidates everything above it. delta anchored
3. Prove the filters individually — each against its own null hypothesis. in shadow, bound missing
4. Only then sharpen the model — a head trained on unsecured inputs learns the noise. restart Sep 23

Why this order is not negotiable: each stage measures the ones beneath it. Sharpening the model before the inputs means optimising against a measurement error — and noticing only once months of data are unusable. In September exactly that became visible: a third of the volume stood on the wrong side.

OUTLOOK

What happens next

The next four weeks — as of September 23

1. **Gate 1, revision 3 — verdict around September 25–26**, on 10 book days under the quote rule. If a day drops out of the book contract, the verdict moves back a day — the hurdle does not.
2. **The one look at "Bracket-MinP"**, two-sided, with the mandatory counter-check excluding day 1. After that the claim right of the next pre-registration is released — never two open endpoints at once.
3. **The new signal chain measures itself:** verdict after 5 trading days and 30 trades against the expectations fixed in advance; old and new chain are never pooled. Plus the missing measurement point: signal age at entry.
4. **Bring filters and inputs to maturity.** `DELTA-MISMATCH` gets a registered bound; the six dead feature inputs get fixed under pre-registration — at the next restart that is needed anyway.
5. **On a positive endpoint:** begin the certification process, starting with independent recomputation by a third party.

The long-term goal, unchanged since day one

The world's best order flow analysis AI — defined not by feature count but by a property no vendor currently claims:

that every statement it makes is verifiable.

The road there does not run through more features. It runs through more proven statements — and through the courage to label the unproven ones as unproven.

What this report set out to establish

Not that we have found an advantage. Rather:

· that we know **how to find one**
· that we know **how one deceives oneself** in the attempt
· that we caught ourselves doing so **54 times**
· and that this produced an apparatus capable of making a future result **certifiable**

That is a foundation a company can be built on. An unsubstantiated success announcement is not.

A request to every reader assessing this project: do not ask for the hit rate. Ask for the null hypothesis, the sample size, and when the criteria were fixed. Anyone unable to answer those three has not measured an advantage — they have found a number. **We wrote this report so that all three answers appear on every page.**

Closing statement. Thirteen months, 471 working sessions, 54 retracted findings, one negative confirmatory headline result, one measured quantity checked against the exchange, one restart of the learned state — and a measuring apparatus this market has not seen before. We are scratching the surface of a data space that we alone record systematically. **Whether an advantage is hidden there, we do not know. But we are the only ones who could prove it — and the only ones who would admit it if there is none.**

OrderFlowAi

OrderFlowAi Research
orderflowai.io/research