Did They Call It, or Did They Just Post It?

A Formal Account of Why Crypto Track Records Fail,
and a Commitment-Based Remedy

Jonathan Quintero
Founder & CEO @ Aevum Protocol

November 2026

Abstract

Crypto runs on performance claims: a screenshot of a winning trade, a win rate in a Telegram bio, a backtest in a pitch deck. Almost none of them can be checked. We give a short formal account of why. A track record fails in four ways: it can be *selected* from many records (Proposition 1), *edited* by deleting the losing calls (Proposition 2), *timed* after the outcome is known (Proposition 3), or summarized by the *wrong metric* (Proposition 4). We show that under modest assumptions each failure produces records that look excellent while carrying no information about skill, and that a reader holding only the published record cannot tell the difference. We check the theory against a snapshot of a public leaderboard, where 47 of the week's 100 best-performing wallets were still losing money overall, and against reported fraud losses. We then describe a remedy: commit to the call before the market moves, grade it by a fixed rule on an independent price, and keep every result, including the wrong ones. We prove that the protocol removes editing and timing as failure modes (Theorem 1), state plainly the failure modes it does not remove, and report our own early results, which were wrong.

1 Introduction

The question every follower should ask a trader is not "how good are you?" It is "how would I know?" In most of finance, the answer involves audited statements, regulated disclosures, or at least a broker statement. In crypto, the answer is usually a screenshot.

This would be a curiosity if the stakes were small. They are not. The FBI's Internet Crime Complaint Center recorded $11.366 billion in reported losses from 181,565 cryptocurrency-related complaints in 2025, of which cryptocurrency investment fraud alone accounted for $7.228 billion [2]. Chainalysis observed at least $14 billion flowing to crypto scams on-chain over the same year and estimates the full figure at $17 billion [3]. Not all of this involves a fake track record. But an investment scam sells one thing: an expectation of returns. And expectations of returns are built from claims about past performance that nobody verified.

Our contribution is modest and practical. First, we name the four ways a record can mislead and put numbers on each. Second, we show that from the outside the misleading record and the honest record look the same, which is why "just be careful" is not advice. Third, we describe a fix that removes two of the four failure modes outright and makes the other two visible. Fourth, we are honest about what it does not fix.

2 The Setup

Definition 1 (Call)**.** A *call* is a tuple $c = (a, d, t_0, \tau)$: an asset a, a directional statement d (for example, "higher" or "above $80,000"), a start time t_0 and a window length τ. The call is *correct* if d holds at time $t_0 + \tau$ under a stated price rule, and *wrong* otherwise.

Definition 2 (Record)**.** A trader's *record* is the set $R = \{c_1, \ldots, c_n\}$ of calls the trader has made. The *published record* $P \subseteq R$ is the set the trader shows. A reader observes only P.

Definition 3 (Commitment)**.** A *commitment* to a call c is a pair $(h,\ s)$ where $h = H(c)$ is a cryptographic fingerprint of the exact text of c and s is a timestamp from a source the trader does not control. A commitment is *ex ante* if $s \leq t_0$, i.e. it exists no later than the moment the window opens.

Everything below is about the gap between R and P, and about when P was written.

3 Failure Mode 1: Selection

Suppose the trader has no skill. Each call is a fair coin: correct with probability $\frac{1}{2}$, independently.

Proposition 1 (A perfect record is expected from noise)**.** *Let N traders each make k calls with no skill. The expected number of traders with a perfect record is $N2^{-k}$, and the probability that at least one exists is $1 - (1 - 2^{-k})^N$.*

Proof. Each trader is perfect with probability 2^{-k}, independently. Linearity of expectation gives the first claim; the complement of "nobody is perfect" gives the second. □

Example 1. Take $N = 10{,}000$ people posting calls, each with a run of $k = 10$. The expected number of perfect ten-for-ten records is 9.77, and the probability that at least one exists is 0.99994. Lower the bar to eight or more correct out of ten and the expected number of such records is about 547.

For a market this size, ten thousand people posting calls is a modest number. The screenshot of a perfect streak is therefore not evidence of skill. It is the expected output of a large population with none, filtered by who chooses to post. The reader sees the winner, not the 9,990 others.

4 Failure Mode 2: Editing

A single trader can run the same filter on themselves.

Proposition 2 (Deletion inflates the win rate)**.** *Let a trader's true hit rate be p, and suppose each wrong call is deleted from P with probability δ, while correct calls are kept. Then the expected published win rate is*

$$\hat{p} \;=\; \frac{p}{p + (1-p)(1-\delta)} \;\geq\; p,$$

with equality only if $\delta = 0$.

Proof. Of n calls, pn are correct and kept, and $(1-p)n$ are wrong, of which a fraction $1 - \delta$ survive. The ratio follows; it exceeds p whenever $\delta > 0$ because the denominator is below one. □

Example 2. A coin-flip trader ($p = 0.5$) who quietly deletes four out of five losing calls ($\delta = 0.8$) publishes a win rate of 83%. Deleting half of them gives 67%. No single post is false. Every published call really happened.

This is the uncomfortable part. Editing does not require lying about any individual trade. It only requires forgetting some of them, and posts can be deleted.

5 Failure Mode 3: Timing

The cheapest way to be right is to decide what you predicted after you know what happened.

Proposition 3 (Indistinguishability without a timestamp)**.** *Let c be a call published at time s, and let the outcome become known at time $t_0 + \tau$. If the reader has no timestamp for s from a source the trader does not control, then a call written before the outcome and a call written after it produce the same published record.*

Proof. The published record contains only the text of c and the trader's own claim about when it was made. Both can be produced at any time. Without an independent s, no function of P separates the two cases. □

Platforms show post times, but posts can be edited, backdated in screenshots, or deleted and reposted, and private groups show nothing at all. "I called this last week" is, as a matter of information, the same sentence whether or not it is true.

6 Failure Mode 4: The Wrong Metric

Even a complete, honestly timed record can mislead if it is summarized by the wrong number. The favorite number in crypto is the win rate.

Proposition 4 (Any win rate below 100% can lose money)**.** *Let a strategy win with probability $p < 1$, with average win $W > 0$ and average loss $L > 0$. Its expected result per trade is $E = pW - (1-p)L$, which is negative whenever $L > \frac{p}{1-p} W$. In particular, for every $p < 1$ there exist W, L with $E < 0$.*

Proof. Direct computation; choose L above the stated bound. □

Example 3 (Illustrative numbers)**.** Trader A wins 9 of 10 trades at +\$100 each and loses one at −\$1,200: a 90% win rate and a net result of −\$300. Trader B wins 4 of 10 at +\$300 and loses 6 at −\$100: a 40% win rate and a net result of +\$600. These numbers are made up to show the arithmetic.

A high win rate is often a sign of the opposite of skill: taking small profits quickly and letting losers run. A record that reports the win rate without the average win, the average loss and the worst trade is reporting the least informative number available.

7 Why the Honest Trader Loses Too

It is tempting to treat this as a problem for followers only. It is at least as much a problem for honest traders, and this is the part that keeps the market broken.

Proposition 5 (Unverifiable skill is priced at the average)**.** *Suppose a fraction q of traders offering signals have real skill, worth $V > 0$ per period to a follower, and the rest have none. If followers cannot distinguish the two from published records, a rational follower will pay at most qV to any trader. A skilled trader whose cost of providing the service exceeds qV withdraws, which lowers q, which lowers what followers will pay.*

Proof. If records carry no information, every trader looks identical, so the follower's expected value from any one of them is qV. A skilled trader participates only if paid at least their cost k; if $k > qV$ they leave. Since unskilled traders have no comparable cost of being wrong, they stay, and the skilled share falls. □

This is Akerlof's market for lemons [1], applied to track records. The honest trader with a real edge cannot prove it, so they are paid like everyone else, and the best of them stop selling. What remains is a market where the loudest records win. Verification does not only protect followers; it is the only way an honest trader can be paid for being honest.

8 A Note on Backtests

Backtests deserve their own warning, because they are where serious people get fooled. Bailey, Borwein, López de Prado and Zhu [4] prove that high backtest performance is easy to achieve after trying only a modest number of strategy configurations, that the more configurations are tried the more likely the result is overfit, and that because the number of trials is rarely reported, investors usually cannot tell how overfit a backtest is. This is Failure Mode 1 again, applied to strategies instead of people: try enough variants and one will look brilliant by chance. The only clean test of a strategy is a forward one, in which the strategy is fixed before the data it is judged on exists.

9 Evidence

9.1 A leaderboard snapshot

On October 28, 2026 we took a snapshot of Hyperliquid's public leaderboard [5]: 47,760 wallets, filtered to the 11,445 that were active (at least $100,000 traded in 30 days and an account value of at least $1,000). We then compared each group's recent returns with its whole history. We kept only aggregates; no names or addresses were stored.

The 100 best wallets of the week had a median gain of 109% in seven days, the kind of week that becomes a screenshot. Forty-seven of them had still lost money over their whole history, and 56 were losing before the month began. Roughly half of the week's best traders had not made money; they had made a week.

The caveats matter and we state them: one exchange, one snapshot, leaderboard wallets rather than wallets tied to particular social posts, and return figures as calculated by the exchange. None of this shows that anyone lied. It shows that one great week tells a reader very little, which is what Propositions 1 and 4 predict.

Group	n	Median 7-day return	Still down overall	Losing before this month
All active wallets	11,445	−0.4%	41.7%	35.9%
Top 100 by 7-day return	100	+109%	47%	56%
Top 500 by 7-day return	500	+52.8%	43.2%	49.6%
Top 100 by 30-day return	100	+23.1%	31%	47%

Table 1: Hyperliquid leaderboard, snapshot of Oct 28, 2026. "Still down overall" means all-time profit below zero. "Losing before this month" means all-time profit minus the last 30 days' profit below zero.

9.2 The scale of the problem

The fraud figures in the introduction are not track-record fraud specifically, and we do not claim they are. They describe the size of the market that unverifiable performance claims feed: $7.228 billion in reported crypto investment-fraud losses to the FBI in a single year [2], and at least $14 billion observed flowing to scams on-chain worldwide [3]. These are reported figures: they count only the losses that victims chose to report.

10 The Remedy: Commit, Grade, Keep

The four failure modes share a structure. Selection and editing work on the gap between R and P. Timing works on when P was written. The wrong metric works on how P is summarized. A remedy has to close all three. The protocol we built, Aevum, does it with three rules.

Rule 1. Commit before. The exact text of the call is fingerprinted and the fingerprint is recorded on Ethereum before the window opens. The timestamp is the block, which the trader does not control.

Rule 2. Grade by rule. When the window closes, the call is graded by a rule fixed in the locked text, on a price the trader does not choose. For Bitcoin up/down calls the price is a signed Atlas Oracle (CoinMarketCap) price recorded on-chain; for other coins and for target/stop calls, Aevum grades on Coinbase's public one-minute prices; for tokenized stocks, on Kraken's public trades. The price rule is part of the locked text; for calls graded by Aevum, so is what happens if no price exists (the call is void and the deposit is returned).

Rule 3. Keep everything. Every commitment from an address stays on that address's public record: correct, wrong and abandoned. An abandoned call counts as wrong. Each address can make at most three commitments in any 30-day tracking period, enforced by the contract, and a summary score is shown only after five graded calls.

Theorem 1 (Editing and timing are removed)**.** *Under Rules 1–3, for any address: (i) the published record equals the full record of commitments from that address, so $P = R$ and $\delta = 0$ in Proposition 2; and (ii) every call in P carries an ex-ante timestamp, so the indistinguishability in Proposition 3 does not arise.*

Proof. (i) Commitments are recorded on a public blockchain and cannot be deleted by the trader or by us; the record shown for an address is read from that chain, so nothing in R can be

omitted from P. (The contract owner can mark an issued certificate as revoked, but that act is itself public and nothing is erased.) (ii) The window opens in the same block that records the commitment, so the timestamp is no later than the window start. Anyone can recompute the fingerprint of the revealed text and compare it with the one on-chain, so a different call cannot be substituted. □

The wrong-metric failure is addressed by presentation rather than proof: records show correct and wrong calls side by side with the size of each move, and our free wallet check shows a trader's win rate next to the average win, the average loss and the worst trade, and flags when the average loss is bigger than the average win.

11 What This Does Not Fix

A paper about unverifiable claims should not make one. Here is what the protocol does not do.

- **Selection across addresses.** Nothing stops one person from opening many addresses and showing the best. The rate limit and the five-call minimum make this slower and more expensive, and every address's record stays complete, but a determined person can still run Proposition 1 on themselves with enough wallets. The defense is to weight a record by its length and age, and to be suspicious of new addresses with short perfect runs.
- **Trust in the grader.** Results are attested by an approved attestor, currently Aevum. For Bitcoin up/down calls the price itself is oracle-signed and stored on-chain; for everything else the price is computed by Aevum from public exchange data that anyone can recheck, but it is not oracle-verified.
- **Grading on time.** For Bitcoin calls, the closing price is the first signed oracle price fetched after the window closes. Our grader refuses to run more than ten minutes late unless deliberately overridden, and when it runs late the delay is written into the on-chain result. A grader that waited could, in principle, choose a better moment; the record would show that it waited.
- **Execution.** A locked call proves the call was honest. It does not prove the trader actually traded it at the stated size and price. Tying calls to real exchange fills is on our roadmap, not live.
- **Small samples.** A complete record of three calls is still three calls. Completeness removes bias; it does not create information that isn't there.
- **It is not an audit.** An audit-ready record is useful and far lighter than a formal audit, but it is not one, and we do not describe it as one.

12 Our Own Record

It would be convenient to report that our own first calls were right. They were not. Our founder's first live Bitcoin call, walked through in Section 13, was wrong. A team member's first call was wrong too, down 2.16%. Both are on the public record, permanently, which is the point: a system that only works when the results are flattering is a marketing tool, not a verification tool. In October we also locked the full rules of our own Bitcoin bot on Ethereum for a 30-day

window ending November 26, 2026, with no code changes and no manual trades allowed. We will publish the result either way.

The contracts have been live on Ethereum mainnet since September 3, 2026, and are owned by a 2-of-3 multisig that includes a hardware wallet.

13 A Worked Example

To make the protocol concrete, here is how one real call moved through it.

1. **The call.** "Bitcoin will be higher in 7 days than at the start." The exact text, including the grading rule, was fingerprinted.
2. **The lock.** The fingerprint was recorded on Ethereum on October 13, 2026 (UTC). The signed Atlas Oracle price recorded with it was $85,532.63.
3. **The window.** For seven days nothing about the call could be changed. Not by the caller, and not by us.
4. **The grade.** Six minutes after the window closed, a fresh signed price was fetched and recorded on-chain: $82,839.10. The call was *wrong*, by 3.14%.
5. **The record.** The result was written to the caller's public record, next to every other call from that address, where anyone can read the locked text, check it against the fingerprint, and compare both prices with the chain.

The caller was our founder. Nothing about the outcome was convenient, and nothing about the process depended on it being convenient.

14 A Checklist for Readers

The propositions above translate into five questions a follower can ask before trusting any record, whether or not it uses Aevum.

1. **Was each call written down before the move, somewhere the trader can't edit?** If not, Proposition 3 applies: you cannot tell a prediction from a story.
2. **Is this the whole record, or a selection?** Ask how many calls were made in total. A record that cannot answer is subject to Proposition 2.
3. **How many calls is it?** Three right out of three is noise. Ten out of ten, from a population of thousands, is also noise (Proposition 1).
4. **What are the average win, the average loss and the worst trade?** A win rate alone says almost nothing (Proposition 4).
5. **Who graded it, and on what price?** A result the trader computed on a price the trader chose is a claim, not a record.

If a trader declines all five, that is also an answer.

15 Conclusion

The question in the title has a precise answer. A reader holding a screenshot cannot tell whether a trader called a move or posted it afterward, whether the record is complete or curated, or whether the one number shown is the one that matters. Propositions 1–4 show that each of these gaps is large enough to make a skill-less record look excellent. The remedy is not to trust harder. It is to change when and where the record is written: before the outcome, somewhere the trader cannot edit, with every result kept.

We do not sell winning calls. We sell calls that cannot be hidden when they lose.

References

[1] G. A. Akerlof, "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," *Quarterly Journal of Economics* 84(3), 488–500, 1970.

[2] Federal Bureau of Investigation, Internet Crime Complaint Center, *2025 Internet Crime Report*, pp. 52–53 ("Complaints Involving Cryptocurrency" and "Cryptocurrency Fraud Trends"). ic3.gov.

[3] Chainalysis, *2026 Crypto Crime Report: Scams*, "Record $17 Billion Estimated Stolen in Crypto Scams and Fraud in 2025," January 13, 2026. chainalysis.com/blog/crypto-scams-2026.

[4] D. H. Bailey, J. M. Borwein, M. López de Prado and Q. J. Zhu, "Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance," *Notices of the American Mathematical Society* 61(5), 458ff, May 2014.

[5] Aevum Protocol, "Wallet study: Hyperliquid public leaderboard," snapshot of October 28, 2026. Source data: Hyperliquid public leaderboard (stats-data.hyperliquid.xyz). Aggregates only.