When the Signal Was Right and the Strategy Still Failed

A retrospective on a self-directed quantitative research project in prediction markets

信号是对的，策略照样死

一个自主量化研究项目的完整复盘

Author	Xuemin (Thomas) Zhang
Affiliation	University of Toronto Mississauga
Project	`polymarket-quant`, a private research repository
Field period	28 June 2026 to 16 July 2026
Document date	September 2026

The project did not find a profitable trading strategy. It produced a verdict, a cost ceiling that held, and a set of decision rules that turned out to be more portable than the strategy itself. This document reports all three.

本项目没有找到能赚钱的交易策略。它产出的是一份判决、一个守住了的成本上限，以及一套比策略本身更能迁移的决策规则。本文报告这三样东西。

Prepared as a research writing sample for graduate study in management. All figures are drawn from the project's own dated archive. Appendix B records where each figure comes from.

Contents

Reader’s note / 阅读说明

This report is organised in two parts. Part I is the English text and Part II is the Chinese text of the same report, with matching section numbers, matching tables, and matching figures. Neither part is a summary of the other.

The project ran for nineteen days and generated a large archive: a private repository with a dated commit history, a set of pre-registered verdict scripts with their recorded outputs, append-only trade ledgers for every strategy variant, and a running research log that was updated the same day any finding appeared. Every number quoted below sits somewhere in that archive. Where a figure came from an outside source rather than from my own measurement, the sentence that uses it says so.

A note on how the work was done. The system was built in a continuous working pair with an AI assistant, which wrote most of the code under my direction. I set the research questions, decided which strategies were built and in what order, wrote or approved every verdict rule before it was committed, and made every decision that involved money. The archive records the relationship running in the other direction as well, since a number of the errors reported below were caught by me and corrected against the assistant’s own stated conclusions, including the sample contamination in section 3.2 and the objection that produced the snapshot rule in section 3.3. I state this at the front because the division of labour is part of the method rather than a footnote to it.

本报告分为两部分。第一部分为英文，第二部分为同一份报告的中文文本，章节编号、表格与数据完全对应，两部分互不为对方的摘要。

项目历时十九天，留下的档案包括一个带完整提交时间戳的私有代码仓库、多份预先注册并留有执行记录的判决脚本、每条策略腿的只增不改交易账本，以及一份当天发现当天入档的研究日志。下文引用的每一个数字都能在该档案中找到出处。凡是来自外部来源而非我本人实测的数据，我会在使用它的那句话里注明。

关于工作方式的说明。本系统是在与一个 AI 助手持续结对的方式下建成的，大部分代码由它在我的指导下写出。研究问题由我设定，建哪些策略、按什么顺序建由我决定，每一条判决规则在提交之前由我书写或批准，所有涉及资金的决定都由我作出。档案同样记录了相反方向的关系：下文报告的若干错误是我抓出来的，并且是在与助手已经写下的结论相反的方向上被纠正的，其中包括第 3.2 节的样本污染，以及催生第 3.3 节快照规则的那次反对意见。我把这一点放在开头说明，是因为分工本身属于方法的一部分，而不是一条脚注。

Part I English Report

1 Introduction and research question

Between 28 June and 16 July 2026 I designed, built and operated an automated research system for Polymarket, a prediction market on which a contract settles at one dollar if the stated event occurs and at zero if it does not. A quoted price of 0.08 therefore reads as an eight percent probability, which is what makes these venues attractive to study: the instrument states its own implied belief, so a claim about mispricing is unusually easy to state and unusually easy to test.

The question I set myself was deliberately narrow. Given a small account, full automation, only public data, and no private information, is there a repeatable edge on this venue that survives real fees, real fills and real settlement?

My answer, reached on 16 July and executed the same morning, was that there is not, for the configuration I could actually reach. The interesting part is not the answer. It is that almost every intermediate result pointed the other way before it was examined properly, and that the examination, rather than the strategy design, turned out to be where the work was.

I am presenting this project rather than a successful one because the record of a falsification carried out under a cost ceiling shows more about how I make decisions than a profitable backtest would. A backtest that works can be produced by accident. A project that kills its own thesis on a rule written before the data arrived cannot.

2 What an edge would have to mean on this venue

Two properties of prediction markets shaped the first strategy families I tried.

The first is that prices are interpretable as probabilities, which lets forecasting accuracy and trading profit be discussed in the same units (Wolfers and Zitzewitz, 2004). The second is the favourite longshot bias, documented in parimutuel betting by Thaler and Ziemba (1988): low probability outcomes tend to be overpriced and high probability outcomes underpriced relative to their realised frequencies. If that bias held here, then systematically selling longshots, or buying near certain favourites and holding them to settlement, should pay.

Both families require one piece of arithmetic that I did not take seriously enough at the start. For a contract bought at price p and held to settlement, a win returns $1 - p$ per share and a loss costs p per share. Break even therefore requires a win rate w satisfying $w(1 - p) = (1 - w)p$, which reduces to $w = p$. The required win rate is simply the entry price. Buying a favourite at 0.99 obliges you to be right ninety nine times in a hundred before fees, and Polymarket's taker fee, charged as a rate multiplied by $p(1 - p)$, raises the bar further.

This identity explains why a high win rate is not evidence of anything on its own, and it became the standing test that several of my own strategies later failed. One of them, a sniper that bought near certain favourites in the last thirty seconds of a window, reached a win rate of 96.92 percent over 260 settled trades against a break even line of 97.81 percent. It was winning almost always and losing money

anyway.

3 Research design: the rules that bound the decisions

Most of what I would defend about this project lives in this section rather than in the strategies. The rules below were adopted in sequence, each one after a specific error made it necessary, and each one was written into the project log with the incident that caused it.

3.1 Verdicts were pre-registered as executable code

Every strategy variant was born together with a verdict script that was committed to the repository before its sample was full. The script stated the sample size at which judgment would occur, the statistic to be computed, and the action to be taken on each branch, including the branch that killed the strategy. When the sample reached the line, the script ran and its output was executed without renegotiation.

The reason for the mechanism is the one the preregistration literature in the experimental sciences gives: a hypothesis generated from the data and a hypothesis tested on the data are different things, and the moment of judgment is exactly when hindsight makes them hardest to tell apart. Writing the verdict as code moved the decision from a moment when I would be looking at a profit and loss curve to a moment when I was not.

Verdict scripts were written for every strategy that mattered, and the archive records at least nine of them firing at their pre-registered line. Four ended the strategy they governed: an instrumentation bundle whose incremental value clause failed, a market neutral pairing strategy at a sample of 150, a sniper on near certain favourites at 260, and the live directional strategy at 32. That last one came due on 16 July against a line of thirty written days earlier, and the leg was retired from real money the same morning.

The verdict scripts were themselves tested. Before any real sample reached them I fed them synthetic data covering every branch. That exercise found a genuine hole in one of them, a case in which a strategy that clearly beat the incumbent would have been filed under a default branch and lost, and the clause was repaired while the strategy's sample was still empty so that the repair could not be accused of being data driven.

3.2 Comparisons had to be paired within the same window

Two strategies that trade at different moments cannot be compared by cumulative profit, because the totals mostly report which opportunities each one happened to take. I learned this by getting it wrong in writing. I recorded that one variant was better than another on the basis of totals of $+21.4$ against $+17.3$ over a similar number of trades. Redone as a paired comparison restricted to windows both had entered, the difference was $+0.08$ per pair with a confidence interval spanning zero. The two were statistically indistinguishable, and the apparent gap had been an artefact of coverage.

What makes the episode worth reporting is that the paired comparison rule was already in the project's own discipline list before I made the error, and I had restated it in my notes more than once, and I

still violated it in the next analysis. Knowing a rule and executing it are separate skills, and the fix was procedural rather than intellectual: from that point, any cross strategy profit comparison had to be prefixed by three stated answers, whether the comparison was paired, how large the sample was, and whether the number had stopped moving.

3.3 A number sampled mid stream is a snapshot, not a finding

Roughly halfway through the project I objected to my own research log on the grounds that it was accumulating precise figures that changed within hours. The rule that replaced the practice is that any mid sample estimate records direction and ordering only, and never enters a decision at its face value.

Two measurements justify it. The estimated marginal contribution of one exit layer read -111 in the evening and -64.5 a few hours later on the same growing sample. A coverage statistic that governed whether a variant lived or died read $+42.3$ one night, which was past the kill threshold of 33, and $+5.3$ the following morning. Because the reading had been filed as a snapshot rather than a conclusion, the variant was not killed, and it went on to pass its verdict cleanly.

3.4 One variable at a time, and never edit a running strategy

Editing a live strategy destroys the sample it has accumulated and makes its verdict meaningless. New ideas were therefore born as new strategy legs whose entry logic was byte identical to the incumbent's, verified by diffing the compiled gate functions, so that the only difference under test was the one being tested. Where a change had to reach several legs at once, all of them were restarted in the same second so that paired comparisons across them stayed single variable.

3.5 Nothing touched real money until a paper version had survived its own verdict

The progression was fixed in advance: paper simulation, then a pre-registered verdict, then a single one dollar live round trip to prove the plumbing, then a batch of small real orders whose only purpose was to measure execution quality, and only then a supervised real money strategy under hard caps. The project reached the last stage on 13 July and left it on 16 July.

4 What was built

The system was written in Python using only the standard library, which was a constraint I set to keep the runtime environment reproducible and the dependency surface at zero. After a mid project cleanup the codebase held sixty files, down from eighty two, with retired components moved to an archive directory rather than deleted so that dead ends stayed auditable.

At its widest the system ran nine concurrent paper strategies on a home machine and three to five more on a rented server, in Virginia at first and in Ireland from 8 July. The relocation followed a measurement rather than an assumption: the venue's matching infrastructure turned out to sit in Europe, the United Kingdom is geoblocked, and Ireland was the nearest jurisdiction that was not. An early low latency

reading from the Virginia box was traced to a content delivery layer rather than to the matching engine, which is one of several occasions on which a flattering number turned out to be measuring something other than what it claimed. The two machines ran identical code, so the pair also served as a speed experiment, with the home machine roughly 150 milliseconds from the venue and the server far closer to it.

Underneath the strategies sat a shared core: a dual WebSocket data layer with starvation watchdogs, the fair value mathematics, a settlement machine with a fixed ordering to prevent double settlement, per token fee retrieval, and file logging. Around them sat the verification infrastructure, which by the end was larger than the strategies. It included a behavioural health checker that ran every thirty minutes and tested assumptions rather than uptime, permanent ballistic test suites that had to pass before any deployment touching the money layer, append only ledgers, an automated daily digest, and a nightly ledger backup.

Risk governance on the real money leg was layered and stated in advance: six dollars maximum per order, twenty five dollars maximum exposure, five dollars maximum loss in a day, a wallet floor, and a kill file whose presence blocked all new entries. The caps were never breached.

Table 1: Project phases and what each one settled

Dates	Phase	What it settled
28 to 30 Jun	Build and broaden	Full paper stack in one day, then a second day of strategy proliferation, then the move from polling to real time data
1 to 2 Jul	Measurement repair	A currency basis defect was found to have poisoned the direction signal. All earlier conclusions from that signal were voided and the data resampled
3 to 7 Jul	Adversarial testing	Pre-registered verdict protocol adopted. Three strategy families judged and killed, and the passive entry direction closed permanently
8 to 11 Jul	Execution measurement	First real order round trip, then a batch of fifty nine real orders whose only purpose was to measure fill rate, slippage and queue time
12 to 15 Jul	Supervised live operation	Real money strategy under caps. The divergence between simulated and realised performance measured directly
16 Jul	Verdict and wind down	Death test fired at the pre-registered line. All systems retired, infrastructure deleted, full account audit filed

5 Findings

5.1 Repairing a pipeline does not repair the data it already collected

The first strategy family rested on measuring the favourite longshot bias from resolved markets. The sampler that collected those measurements was supposed to read prices from seven days before resolution. Because the venue’s published end date is frequently much later than the moment a market actually resolves, and because the sampler carried a tolerance window, it was in fact reading prices recorded at or after resolution, when they have already collapsed to zero or one. Forty four percent of the sampled markets had prices that simply contained the answer.

Fixing the sampler was the obvious response and was not sufficient. The two hundred markets already

collected with the defective sampler stayed in the dataset and kept contaminating every analysis that touched it. The size of the contamination is easy to state: the measured hit rate for longshots was thirty two percent on the old data and eleven percent on data collected after the repair, a factor of three.

The operating rule that came out of this is that a repaired instrument creates a boundary in time, and results are only trustworthy on the data collected after it. The same principle appeared twice more later in the project, once when the fee model was corrected and once when the price feed was changed, and on both occasions the data was explicitly divided into eras rather than pooled.

5.2 The small sample mirage

One strategy showed a hundred percent win rate and a measured edge of $+7.89$ percent. The sample was four trades. Applying standard shrinkage to the win rate, which is the ordinary correction for a proportion estimated from very few observations, turned the edge negative immediately. Four wins out of four is consistent with a very wide range of true win rates, and the zero variance that makes such a result look clean is precisely the sign that no losing observation has arrived yet.

The same shape returned at a larger scale on 13 July, when the real money leg produced six consecutive wins and then two full stake losses that erased them within an hour. For strategies that buy favourites, the profit distribution is many small gains and occasional large losses, so the sample size at which a mean becomes informative is much larger than intuition suggests. This is the reason the verdict lines were set at one hundred paired observations wherever the sample allowed it.

5.3 A measurement artefact that looked like a market regime

On 1 July the direction strategy bought the downside in ten consecutive trades, and the underlying move measure clustered in a narrow band of -0.09 to -0.17 percent. The cause was that the current price came from a blend of two venues, one quoted in a dollar stablecoin and one in dollars, while the opening price came from the stablecoin venue alone. The stablecoin was trading about 0.14 percent below a dollar at the time, which injected a permanent negative offset of roughly 0.07 percent into a measure whose genuine values were mostly smaller than that. Small real rises were being reported as falls.

The market had not moved down. The instrument had. Every conclusion previously drawn from that signal, including a confident statement that the market was efficient over one hundred and thirty seven samples, was voided and the data recollected. The rule recorded was that a ratio between two quantities requires both endpoints to come from the same source and the same unit of account, so that any common offset cancels.

5.4 Costs that were assumed rather than measured

The system’s fee model used a static fee rate of seven percent taken from documentation, applied through the venue’s formula in which the charge equals that rate multiplied by $p(1-p)$. Querying the venue’s own fee endpoint per token returned a base fee of one thousand basis points, which is ten percent. Every

simulated result in the system had therefore been under charging fees by roughly thirty percent. Because both sides of any internal A/B comparison used the same wrong model the comparisons remained valid, but every absolute profit figure carried an asterisk until the model was corrected on 5 July and the data split into eras.

This finding arrived by an indirect route that is worth recording. A separate implementation written by someone I know reported strong arbitrage profits and described the venue as zero fee. Reading that code showed the fee lookup was present and correct in structure but parsed the wrong key from the response, returning nothing, defaulting to zero, and caching that zero permanently. The author's own comments elsewhere in the file showed he understood the fee arithmetic perfectly well. His instrument was hiding from him the one input his own logic depended on. Checking his claim is what sent me to check mine.

5.5 Simulated performance and realised performance diverged, and the divergence was structural

This is the finding that ended the project, and it only became visible when real orders were placed.

The paper system had assumed that an order could transact at the price visible in its own copy of the order book. Measurement against real orders showed three separate problems with that assumption. First, in fast moving windows the price the paper system recorded as a fill did not exist by the time a real order could reach the venue. A cross check at one window found the paper strategies recording fills between 0.818 and 0.832 while the real leg's verification probe returned 0.92. Second, real resting orders join a queue and do not always fill: the measured fill rate for the exit order resting at 0.99 was seventy four percent over twenty seven real orders, against the hundred percent the simulation had assumed. Third, and largest, the real strategy simply transacted far less often than the simulation, because the verification step aborted most attempts when the real price no longer matched the signal. The early measurement was an abort rate of eighty seven percent, a real transaction cadence of roughly one tenth of the simulated one.

Taken together these effects meant the simulator's absolute profit was inflated by roughly an order of magnitude. The internal comparisons it had been used for were still sound, since every strategy was flattered equally. Its absolute numbers were not a forecast of anything.

Table 2: The same momentum signal, simulated and realised

	Paper simulation	Real money
Win rate, hold to settlement	80 percent (n = 270)	34 percent on filled orders
Assumed fill on a resting exit at 0.99	100 percent	74 percent (n = 27)
Transaction cadence	baseline	about one tenth
Abort rate at price verification	not modelled	87 percent (early measurement)
Absolute profit and loss	inflated about tenfold	the reference

5.6 Attribution before celebration

At thirteen real fills the leg was net positive by $5.33, which was the point at which it would have been easy to conclude that the strategy worked. Decomposing the result by source showed that the entry itself

carried a margin of -0.131 per share and that the whole of the profit came from the exit management layer, which cut losses and locked in gains. A profitable total was concealing an unprofitable thesis.

The decomposition mattered because the two components have different futures. An exit layer that rescues a negative entry to approximately break even is not an edge, it is a well built brake. Scaling it does not produce more profit, it produces more trades that end near zero, and each of those trades still pays fees.

5.7 Eliminating the remaining options

Once the entry thesis was in doubt, the sensible move was not to tune it but to check whether any other route on the same venue was open. Between 13 and 16 July each remaining option was tested and closed.

Table 3: Options tested and the reason each was closed

Option	Outcome
Directional momentum	Signal correct in simulation, unfillable in the windows where it was correct
Fading the signal	Also dead. The naked hold to settlement variant won 80 percent of 270 trades, so a blind fade wins 20 percent, and no pre-signal feature separated the winners from the losers
Intra platform arbitrage	A scanner was built and run across all mutually exclusive event groups. No executable opportunity found
News, earnings and macro	Scanned roughly 2,100 active markets against eighteen keywords. No market on these topics settled quickly enough to trade the way the system was built to trade
Speed	Measured rather than assumed. Order construction and signing took eight milliseconds and the best case for the full pipeline was about half a second. A first reading suggested most of that was self inflicted and removable. A second and more careful measurement corrected it: the placement path was already fast and the remainder was not cuttable in the way the first reading implied, so measuring before rebuilding avoided a pointless change
Passive market making	Tested to death earlier in the project. A market making entry filled preferentially when the market was about to move against it, which is the classic adverse selection result
Copy trading	The wallets available to copy were exit driven, and the profile of consistently profitable wallets showed they were category specialists rather than general operators
Domain knowledge	I had none that the market did not already have

6 The terminal finding

The three sentences I wrote into the archive when the project closed are the honest summary of what nineteen days produced.

The signal was right and the strategy died anyway. The same momentum signal produced an eighty percent win rate when held to settlement in simulation and a thirty four percent win rate on the orders that actually filled. The difference is not in the signal. It is in which windows are reachable.

The gap is adverse selection at the moment of the fill. The windows in which the signal was correct repriced within about a hundred milliseconds, so an order aimed at them arrived after the opportunity had gone. The windows in which an order rested calmly and filled were disproportionately the windows about to reverse. What determines whether a trade wins is therefore not a property of the signal at the

moment of the decision. It is a property of whether the price was still there when the order landed, which is information the strategy does not possess at decision time.

This maps directly onto the standard microstructure result that a quoting party loses systematically to better informed counterparties, and that the loss shows up as a spread the quoter must charge to survive (Glosten and Milgrom, 1985). In my case I was on the wrong side of that relationship without being able to charge for it. The spreads on the relevant contracts ran between eight and seventeen cents while the mispricing I was chasing was worth a few cents, so a round trip that crossed the spread was structurally negative before any question of forecasting skill arose.

The defect is not reachable by tuning. It cannot be removed by adjusting parameters, because the parameters govern the decision and the problem is downstream of the decision. It cannot be removed by reversing the signal, because fading was tested and failed for the same reason. It cannot be filtered out, because the feature that would need to be filtered on, whether this particular window is about to reverse, is absent from every pre-signal observable I recorded.

What remained after that was a statement about the combination rather than about any component: small capital, high speed, full automation, this venue, and no information advantage do not compose into a workable position. Each element is individually defensible. Together they leave nothing to exploit.

7 Risk governance and what it cost

The financial outcome of the live phase is small enough to state completely. The real money strategy placed thirty five orders and ended cumulatively down $5.62, with the wallet holding no open positions and nothing unclaimed. Server infrastructure cost nothing in the end, since the recorded usage of 13.65 US dollars was covered by promotional credit. The caps that had been set in advance, six dollars per order, twenty five dollars of exposure, five dollars of loss per day, were never breached.

I report this not as a good financial result but as evidence about a process. The strategy was wrong, and the loss from being wrong was bounded to roughly the size of one restaurant meal, because the bound had been fixed before the strategy was allowed to touch money and because the instrument that enforced it was mechanical rather than discretionary.

The governance layer produced its own findings, and they were more operational than statistical. Settled winnings did not redeem automatically, so twenty dollars of realised profit sat in an intermediate token form and was invisible to a cash based accounting check. A race condition between order cancellation and order fill discarded shares that had actually been purchased, creating positions with no exit management and no ledger entry. A cancellation call that reported success was found by a deliberately safe test order to be a silent no operation, which explained a slippage pattern on stop orders that had been attributed to market behaviour. None of these are strategy defects. All of them would have caused real losses at any larger size, and all of them were found because the system was instrumented to check its own assumptions rather than merely to report that it was running.

8 Six revisions to my own thinking

From finding an edge to falsifying one. I started looking for a profitable pattern and finished having built an apparatus whose purpose was to destroy candidate patterns as cheaply as possible. The second activity turned out to be the one with transferable output.

From win rate to the payoff adjusted break even line. The identity that the required win rate equals the entry price is elementary, and I still had to be taught it by a strategy that won ninety seven times in a hundred and lost money.

From signal quality to execution quality. For most of the first two weeks I treated the question as forecasting. The evidence says the binding constraint was never the forecast. It was whether the forecast could be transacted.

From direction to sign dependence. Passive entry was poison and passive exit was safe, on the same venue, with the same mechanism. Adverse selection is directional, and a mechanism that harms you in one direction can help you in the other. I had been treating market making as a single technique with a single verdict.

From tuning to arena selection. Late in the project I stopped asking which parameter to change and started asking whether the venue, at my capital and my latency, contained a reachable edge at all. That question should have been asked in the first week, and asking it is what produced the answer.

From failure to verdict. The project reads as a failure if the output is defined as a profitable strategy. Defined as a research question answered under a cost ceiling, with the answer reproducible from a dated archive, it produced what it was built to produce. I would rather present a documented negative result than an undocumented positive one.

9 What transfers

The strategies do not transfer. The procedures do, and I would use all of the following again in a management research or analytics setting.

1. **Write the decision rule before the data arrives, and make it executable.** A rule that is merely written down is renegotiated when the moment comes. A rule that is code, with a stated sample threshold and a stated action, is not.
2. **Compare like with like or do not compare.** Cumulative totals across units that faced different opportunity sets measure the opportunity sets. Pairing on a shared unit of observation is the minimum defence.
3. **Treat mid stream estimates as direction only.** Numbers that are still moving belong in a log as ordering information, not in a decision as values.
4. **Date your instruments and split your data at the repair.** Fixing a measurement process creates a boundary, and pooling across it reintroduces the defect you just removed.
5. **Decompose a favourable total before believing it.** A positive result assembled from a negative core and a compensating mechanism will not scale, and the decomposition is usually cheap.

6. **Simulations validate relative ranking, not absolute levels.** Whenever a simulated quantity is used as a forecast of a realised one, the mapping between them needs its own measurement.
7. **Mathematical elimination permits an early decision without discretion.** When the remaining sample cannot change the outcome, stopping early is arithmetic rather than impatience. The sniper strategy was judged at 260 observations rather than 300 because 40 further wins could not have reached the break even line.
8. **Instrument the assumptions, not the uptime.** Every one of the operational defects above was found by a check that tested whether a stated assumption still held, not by a check that confirmed the process was alive.

10 Limitations

The real money sample is small. Thirty five orders is enough to fire a pre-registered rule that was written for that size and not enough to support a general claim, and the thirty four percent realised win rate is an estimate with a wide interval around it. The verdict I am defending is about one configuration on one venue at one capital level, not about prediction markets as a class.

The paper to live divergence was measured on a single strategy family. I would expect the direction of the effect to generalise, since it follows from queue mechanics rather than from anything specific to momentum, but I did not test it elsewhere.

Two lanes were reasoned about and not closed with real money. A cross venue arbitrage against a second exchange was designed and then cut on my own decision before implementation, and a market neutral pairing strategy was killed on simulated evidence at a sample of one hundred and fifty rather than on live evidence. I state both as open rather than as answered.

Finally, the project has one operator and one account. There is no replication, and the research log, although written the same day as the events it records, is written by the person whose decisions it evaluates.

11 Relevance to graduate study in management

Four threads in this project sit inside management research rather than beside it.

The first is commitment devices and the escalation of commitment. The pre-registered kill criterion is an anti escalation mechanism aimed at a known failure of individual and organisational decision making, which is the tendency to increase investment in a failing course of action in proportion to what has already been spent. I built one because I expected to need it, and on 16 July I did.

The second is measurement validity. A simulated key performance indicator and its realised counterpart differed by an order of magnitude while remaining perfectly useful for ranking. Any organisation that steers on a modelled metric faces the same distinction, and the mistake of reading a level off an instrument that was only ever calibrated for comparison is not confined to trading.

The third is information asymmetry and adverse selection, which reached me through market microstructure but is the same mechanism that organises insurance underwriting, procurement auctions and plat-

form design. Meeting it as an operational fact rather than as a model made the literature legible to me in a way that reading alone had not.

The fourth is survivorship in self reporting. The most uncomfortable finding in the project was not about the market. It was that my own sense of how the account had performed was assembled from the episodes that had survived, and that a published dashboard figure and the underlying cash movements disagreed because they measured different things. Correcting that required a full reconciliation rather than an argument. Organisations produce the same artefact at scale, and I would like to study it properly.

For a master’s programme in management my intention is to work on decision processes under uncertainty, specifically on how rules fixed in advance change the quality of decisions made later by the same people who wrote them. This project is a single case study of that question in which I was both the subject and the experimenter, which is a weakness as evidence and the reason I want to study it with better methods.

12 Closing assessment

The system was retired on 16 July by the rule that had been written for it, and the archive was closed rather than tidied, so that the dead ends remain visible next to the decisions that killed them. Nothing about the outcome was surprising by the end. The surprising part had all happened in the middle, where a corrected pipeline kept producing corrupted answers, where a clean hundred percent win rate turned out to be four observations, and where a market that appeared to be falling was an instrument reading itself wrong.

If the project is worth anything to a reader, it is as a demonstration that a research question can be closed honestly at small cost, and that the closing is worth writing down.

Part II　中文报告

1　研究缘起与问题

2026 年 6 月 28 日至 7 月 16 日，我独立设计、搭建并运行了一套针对 Polymarket 的自动化研究系统。Polymarket 是一个预测市场，合约在事件发生时按一美元结算，不发生时归零，因此 0.08 的报价可以直接读成百分之八的概率。这一性质使该类市场在研究上有吸引力：标的自己报出隐含信念，关于定价偏误的主张既容易陈述，也容易检验。

我给自己设定的问题范围很窄：在小额本金、完全自动化、只使用公开数据、且不具备任何私有信息的条件下，这个平台上是否存在一个能够在真实费率、真实成交与真实结算之下存活的、可重复的优势。

我在 7 月 16 日得到的答案是：就我实际能够达到的配置而言，不存在，并且当天上午就执行了退出。真正值得说明的不是这个答案，而是几乎每一个中间结果在被认真检验之前都指向相反的方向，以及整个项目的工作量主要落在检验上，而不是落在策略设计上。

我选择提交这个项目而不是一个成功的项目，是因为在成本上限之内完成一次自我证伪的记录，比一条漂亮的回测曲线更能说明我如何做决定。能跑通的回测有可能是偶然产生的，而一个按照数据到来之前就写好的规则杀死自己论点的项目不会是偶然。

2　在这个平台上，优势必须意味着什么

预测市场的两个性质决定了我最初尝试的策略族。

其一是价格可以直接解释为概率，这使得预测准确度与交易盈利能够放在同一套单位里讨论（Wolfers and Zitzewitz, 2004）。其二是 Thaler 与 Ziemba（1988）在赛马彩池市场中记录的冷门偏误：低概率结果相对其实际发生频率倾向于被高估，高概率结果倾向于被低估。如果这一偏误在此处同样成立，那么系统性地卖出冷门，或者买入接近确定的热门并持有至结算，都应当有利可图。

这两族策略都要求一个我在开始阶段没有足够重视的算术。以价格 p 买入并持有到结算，赢一次每股收益为 $1-p$，输一次每股损失为 p。保本要求胜率 w 满足 $w(1-p)=(1-w)p$，化简后得 $w=p$。也就是说，所需胜率恰好等于入场价格。以 0.99 买入热门，意味着在不计费用的情况下必须每一百次对九十九次，而 Polymarket 的吃单费按费率乘以 $p(1-p)$ 收取，会把这条线进一步抬高。

这个恒等式解释了为什么单看胜率不构成任何证据，并且后来成为我自己几条策略未能通过的常设检验。其中一条在窗口最后三十秒买入接近确定的热门，在 260 笔已结算交易上取得 96.92% 的胜率，而保本线是 97.81%。它几乎每次都赢，却仍然在亏钱。

3　研究设计：约束决策的几条规则

这个项目中我最愿意为之辩护的部分在本节，而不在策略本身。下面这些规则是依次形成的，每一条都出自一次具体的错误，并且都连同引发它的事件一起写进了项目日志。

3.1 判决以可执行代码的形式预先注册

每一条策略腿在出生时都配有一份判决脚本，该脚本在样本尚未填满之前就提交进仓库，其中写明在多大样本量上开庭、计算哪个统计量，以及每个分支对应的动作，包括判死策略的那个分支。样本到线时脚本自动运行，其输出不经重新协商直接执行。

采用这一机制的理由，与实验科学中预注册文献给出的理由相同：由数据生成的假设与用数据检验的假设是两回事，而判决那一刻恰恰是事后偏见最容易把两者混淆的时刻。把判决写成代码，等于把决策从我正盯着盈亏曲线的那一刻，挪到我还没有看见它的时候。

凡是重要的策略都配有判决脚本，档案中至少记录了九份在各自预注册线上落锤。其中四份终结了它所管辖的策略：一套仪表组件在增量条款上失败，一条市场中性凑对策略在样本 150 判死，一条针对接近确定热门的狙击策略在 260 判死，以及真钱方向策略在 32 判死。最后这一份于 7 月 16 日在数日之前就写好的判决线 30 上到线，该腿当天上午即退出真钱。

判决脚本本身也接受检验。在任何真实样本到达之前，我用合成数据打穿它们的全部分支。这项演练在其中一份里查出了一个真实漏洞：一条明显战胜在位者的策略会被归入默认分支而被判输。该条款在这条策略样本仍为零时完成修补，因此这次修补不可能被指为由数据驱动。

3.2 比较必须在同一窗口内配对

两条在不同时刻交易的策略无法用累计盈利比较，因为累计值主要反映的是各自碰到了哪些机会。这一点是我用白纸黑字写错一次学会的。我曾记录某个变体优于另一个，依据是在交易笔数相近的情况下总净值 $+21.4$ 对 $+17.3$。改用仅限双方都进入过的窗口做配对比较后，差值为每对 $+0.08$，置信区间跨越零。两者在统计上无法区分，表面上的差距来自覆盖范围的差异。

这件事值得写进来的原因在于，配对比较这条规则在我犯错之前就已经写进项目的纪律清单，我还在自己的笔记里不止一次重申过它，结果仍然在下一次分析里违反了它。知道一条规则和执行一条规则是两种不同的能力，因此修补是程序性的而非认知性的：自那以后，任何跨策略的盈利比较都必须先回答三个问题，是否配对、样本多大、数字是否已经稳定。

3.3 采样过程中的数字是快照，不是结论

项目进行到大约一半时，我对自己的研究日志提出了反对意见，理由是里面正在堆积一批几小时之内就会改变的精确数值。取而代之的规则是：任何中途估计只记录方向与排序，绝不以其面值进入决策。

有两项测量支持这条规则。某个出场层的边际贡献估计值在傍晚读数为 -111，几小时后在同一个仍在增长的样本上变为 -64.5。另一个决定某变体生死的覆盖率统计量，前一晚读数为 $+42.3$，已经越过 33 的判死阈值，次日清晨回落到 $+5.3$。由于该读数当时是按快照而非结论入档的，这个变体没有被杀掉，后来干净地通过了自己的判决。

3.4 一次只动一个变量，并且绝不修改运行中的策略

修改一条正在运行的策略会毁掉它已经积累的样本，使其判决失去意义。因此新想法一律以新腿的形式出生，其入场逻辑与在位者逐字节一致，通过对编译后的闸门函数做差异比对来验证，

从而保证受检验的差异只有一个。当某项改动必须同时到达多条腿时，这些腿在同一秒内重启，以维持跨腿配对比较的单变量性质。

3.5 在纸面版本通过自己的判决之前，不动真钱

推进顺序是预先固定的：纸面模拟，预注册判决，一笔一美元的真实往返以验证管道，一批仅用于测量执行质量的小额真单，最后才是硬性帽子之下有人值守的真钱策略。项目在 7 月 13 日进入最后一个阶段，并在 7 月 16 日离开它。

4 实际搭建了什么

系统用 Python 编写，只使用标准库。这是我自己设的约束，目的是让运行环境可复现，并把依赖面积降为零。项目中期清理之后，代码库从 82 个文件缩减到 60 个，退役组件移入归档目录而非删除，以便走过的死路仍然可查。

系统最宽的时候，在家用机上并行运行九条纸面策略腿，另有三到五条运行在租用的服务器上，服务器起初位于弗吉尼亚，自 7 月 8 日起改在爱尔兰。这次迁移依据的是实测而不是假设：该交易所的撮合设施实际位于欧洲，而英国属于地理封锁区，爱尔兰是最近的未被封锁的辖区。弗吉尼亚机器上一个很好看的低延迟读数，后来被追溯为内容分发层造成的，而非撮合引擎的真实距离，这是本项目中若干次“讨喜的数字其实在测量别的东西”里的一次。两台机器运行同一份代码，因此这一对机器同时充当速度实验，家用机距交易所约 150 毫秒，服务器则近得多。

策略之下是共享内核：带饿死看门狗的双 WebSocket 数据层、公允价值数学、带固定次序以防重复结算的结算机、按 token 拉取的费率、以及文件日志。围绕它们的是验证基础设施，到项目后期其体量已超过策略本身，包括每三十分钟运行一次、检验假设而非检验存活的行为体检器，任何触及资金层的部署之前都必须通过的常设弹道测试套件，只增不改的账本，自动生成的每日简报，以及每晚的账本备份。

真钱腿的风险治理是分层的，并且事先声明：单笔最高 6 美元，总敞口最高 25 美元，单日最大亏损 5 美元，钱包地板，以及一个只要存在就阻断全部新开仓的 KILL 文件。这些帽子在整个过程中从未被突破。

表 1　项目阶段与各阶段解决的问题

日期	阶段	解决了什么
6/28 至 6/30	搭建与铺开	一天内完成整套纸面系统，次日策略数量迅速扩张，第三天从轮询转向实时数据
7/1 至 7/2	修复测量	查出一个货币基差缺陷毒化了方向信号，由该信号得出的既有结论全部作废并重新采集数据
7/3 至 7/7	对抗性检验	确立预注册判决机制，三个策略族被审判并判死，被动挂单入场方向被永久关闭
7/8 至 7/11	执行质量测量	完成首笔真单往返，随后下达 59 笔真单，唯一目的是测量成交率、滑点与排队时长
7/12 至 7/15	有人值守的真钱运行	在帽子之下运行真钱策略，直接测得模拟绩效与实际绩效之间的差距
7/16	判决与收摊	死亡测试在预注册线上落锤，全部系统退役，基础设施删除，完成账目终审

5 研究发现

5.1 修好管道并不能修好它已经收集的数据

第一个策略族依赖于从已结算市场中测量冷门偏误。负责采集这些测量值的采样器本应读取结算前七天的价格。由于平台公布的结束日期经常远晚于市场实际结算的时刻，加之采样器带有一个容错窗口，它实际上读到的是结算当时或之后记录的价格，而那时价格已经钉死在零或一。被采样市场中有 44% 的价格直接包含了答案。

修好采样器是显而易见的回应，但并不充分。此前用有缺陷的采样器收集的两百个市场仍留在数据集中，继续污染每一次触及该数据集的分析。污染的规模很容易陈述：冷门的实测命中率在旧数据上为 32%，在修复之后收集的数据上为 11%，相差三倍。

由此得到的操作规则是，一次仪器修复在时间轴上划出一条边界，只有边界之后收集的数据才值得信任。同一原则在项目中又出现过两次，一次是费率模型被修正时，一次是价格源被更换时，两次都把数据显式划分为不同纪元，而没有合并处理。

5.2 小样本幻觉

有一条策略一度显示 100% 胜率与 $+7.89\%$ 的实测优势，样本量是四笔。对胜率施加标准的收缩处理，也就是对由极少观测估计出的比例所做的常规校正之后，这个优势立刻转负。四战四胜与相当宽的一段真实胜率取值都不矛盾，而让这类结果显得干净的零方差，恰恰是尚未出现任何一次失败观测的标志。

同样的形状在 7 月 13 日以更大的尺度重现：真钱腿连续赢了六笔，随后两笔满注亏损在一小时之内把它们抹平。对于买入热门的策略而言，盈亏分布是大量小额盈利加上偶发的大额亏损，因此均值变得有信息量所需的样本量，远大于直觉的估计。这也是判决线在样本允许的情况下一律设在一百对配对观测的原因。

5.3 一个看起来像行情状态的测量假象

7 月 1 日，方向策略连续十笔买入下跌方向，而底层的涨跌幅度量挤在 -0.09% 到 -0.17% 的窄区间里。原因是当前价格取自两个交易所的混合价，其中一个以美元稳定币计价，另一个以美元计价，而开盘价只取自稳定币交易所。当时该稳定币比一美元低约 0.14%，于是在一个真实取值大多小于该幅度的度量里注入了约 0.07% 的永久负偏移。真实的小幅上涨被报告成了下跌。

市场并没有下跌，是仪器在动。此前由该信号得出的全部结论，包括一项关于市场在一百三十七个样本上有效的自信陈述，全部作废并重新采集。记录下来的规则是：两个量之间的比值要求两端来自同一来源、同一计价单位，任何共同偏移才能相消。

5.4 被假设而非被测量的成本

系统的费率模型使用了来自文档的静态费率 7%，按平台的公式计费，即收费等于该费率乘以 $p(1-p)$。按 token 查询平台自己的费率接口，返回的基础费率为一千个基点，也就是 10%。这

意味着系统里全部模拟结果对费用的计提都低了约三成。由于任何一次内部 A/B 比较的两侧使用的是同一个错误模型，比较结论仍然成立，但所有绝对盈利数字在 7 月 5 日模型被修正、数据被划分纪元之前，都必须带一个星号。

这项发现是通过一条间接路径到达的，值得记录。我认识的一个人另行实现了一套系统，报告了可观的套利利润，并称该平台零费率。阅读那份代码发现，费率查询在结构上是存在且正确的，但解析了返回结果中错误的键，取不到值，默认为零，并把这个零永久缓存。该作者在同一文件其他位置的注释显示，他对费用算术的理解完全正确。他的仪器向他隐藏了他自己的逻辑所依赖的那一个输入。去核对他的说法，正是促使我核对自己的原因。

5.5 模拟绩效与实际绩效发生分叉，且分叉是结构性的

这是终结项目的那一项发现，而它只有在真实订单被下达之后才变得可见。

纸面系统假设订单可以按其自身订单簿副本中可见的价格成交。用真实订单做对照测量后，这一假设暴露出三个彼此独立的问题。第一，在行情快速移动的窗口里，纸面系统记录为成交的那个价格，在真实订单能够到达交易所时已经不存在；某个窗口的交叉核对显示，纸面策略记录的成交价在 0.818 至 0.832 之间，而真钱腿的验价探针返回的是 0.92。第二，真实的挂单需要排队，并不总能成交：挂在 0.99 的出场单在 27 笔真单上实测成交率为 74%，而模拟假设的是 100%。第三，也是影响最大的一点，真实策略实际发生交易的频率远低于模拟，因为验价环节在真实价格不再匹配信号时中止了大多数尝试。早期测得的中止率为 87%，真实交易频次约为模拟的十分之一。

这些效应叠加起来意味着模拟器的绝对盈利被放大了大约一个数量级。它此前被用于的内部比较仍然成立，因为每条策略受到的美化是等同的。但它的绝对数字不构成对任何事情的预测。

表 2　同一个动量信号在模拟与真实中的表现

	纸面模拟	真钱
持有至结算的胜率	80%（n = 270）	已成交订单 34%
挂在 0.99 的出场单成交率	假设 100%	实测 74%（n = 27）
交易发生频次	基准	约十分之一
验价环节中止率	未建模	87%（早期测量）
绝对盈亏	放大约十倍	参照系

5.6 先做归因，再谈成绩

在十三笔真实成交时，该腿净盈利 5.33 美元，这正是最容易得出策略有效这一结论的时点。按来源拆解后可以看到，入场本身的边际是每股 −0.131，全部利润来自出场管理层，也就是负责止损与锁定收益的那一层。一个为正的总数掩盖了一个为负的论点。

这项拆解之所以重要，是因为两个组成部分的前景并不相同。把一个为负的入场救回到大致保本的出场层不是优势，而是一套做得不错的刹车。放大它不会带来更多利润，只会带来更多结果接近于零的交易，而每一笔这样的交易仍然要付费用。

5.7 把剩余选项逐一排除

入场论点一旦存疑，合理的动作不是继续调参，而是检查同一平台上是否还有别的路可走。7 月 13 日至 16 日之间，剩余的每一个选项都被检验并关闭。

表 3　被检验的选项与各自关闭的理由

选项	结果
方向性动量	信号在模拟中正确，但在它正确的那些窗口里成交不上
反向操作	同样不成立。裸持有至结算的变体在 270 笔上胜率 80%，因此盲目反做的胜率是 20%，而且没有任何信号前的特征能把赢单与输单分开
平台内套利	建立并运行了扫描器，遍历全部互斥事件组，未发现可执行的机会
新闻、财报与宏观	扫描约 2100 个活跃市场与十八个关键词，这些题材上没有任何市场的结算速度快到能被本系统的交易方式利用
速度	采取实测而非假设。构建并签名订单耗时 8 毫秒，整条管道最优约 0.5 秒。第一次读数提示其中大部分是自身造成、可以削减的，第二次更细致的测量作出了修正：下单路径本身已经足够快，剩余部分并不像第一次读数所暗示的那样可以削减，因此先测量再动手避免了一次无意义的改造
被动做市	在项目早期已被检验至死。挂单入场在市场即将向不利方向移动时优先成交，这正是典型的逆向选择结果
跟单	可供跟随的钱包属于主动卖出型，而持续盈利钱包的画像显示他们是单一品类的专才，而非通吃的操作者
领域知识	我不具备市场尚未掌握的领域知识

6　终局发现

项目关闭时我写进档案的三句话，是对这十九天产出的诚实总结。

信号是对的，策略照样死。同一个动量信号，在模拟中持有至结算的胜率为 80%，而在真正成交的订单上胜率只有 34%。差别不在信号，而在哪些窗口是够得到的。

这道缝隙的性质是成交时刻的逆向选择。信号正确的那些窗口在大约一百毫秒之内完成重定价，因此瞄准它们的订单到达时机会已经消失。而订单能够安稳挂住并成交的那些窗口，不成比例地正是即将反转的窗口。决定一笔交易输赢的，因此不是决策时刻信号的某个属性，而是订单落地时价格是否还在原处，而这恰恰是策略在决策时不掌握的信息。

这一点直接对应市场微观结构中的标准结果：报价方会系统性地输给信息更充分的对手方，其损失体现为报价方为存活所必须收取的价差（Glosten and Milgrom, 1985）。在我的情形中，我处在这一关系的不利一侧，却无法为此收费。相关合约的买卖价差在八到十七美分之间，而我所追逐的定价偏差只值几美分，因此一次跨越价差的往返在任何关于预测能力的问题出现之前，结构上就已经为负。

这个缺陷不是调参能够触及的。它无法通过调整参数去除，因为参数管辖的是决策，而问题发生在决策的下游。它无法通过反向操作去除，因为反做已经被检验，且因同一原因失败。它也无法被过滤掉，因为需要据以过滤的那个特征，即这个特定窗口是否即将反转，在我记录过的每一个信号前可观测量里都不存在。

在此之后剩下的，是一个关于组合而非关于任何单一部件的陈述：小额本金、追求速度、完全自动化、这个平台，以及没有信息优势，这五项无法组合成一个可行的位置。每一项单独看都站得住。放在一起则没有可供利用的空间。

7 风险治理及其成本

真钱阶段的财务结果小到可以完整陈述。真钱策略共下达 35 笔订单，累计亏损 5.62 美元，结束时钱包内没有持仓，也没有待领取的款项。服务器基础设施最终没有产生费用，因为记录在案的 13.65 美元用量被推广额度覆盖。事先设定的帽子，即单笔 6 美元、敞口 25 美元、单日亏损 5 美元，全程未被突破。

我报告这一点，不是把它当作一个良好的财务结果，而是把它当作关于流程的证据。策略是错的，而错误带来的损失被限制在大约一顿饭的量级，因为这个界限在策略被允许接触资金之前就已固定，并且执行这个界限的工具是机械的而非自由裁量的。

治理层自身也产出了发现，且这些发现偏操作而非偏统计。已结算的赢款不会自动兑付，因此二十美元的已实现利润停留在一种中间代币形态，对基于现金的账目检查完全不可见。订单撤销与订单成交之间的竞态条件丢弃了实际已经买到的份额，造成了既没有出场管理也没有账本记录的持仓。一个报告成功的撤单调用，被一笔刻意设计为安全的测试订单证明是静默的空操作，这解释了此前被归因于市场行为的止损滑点模式。这些都不是策略缺陷。它们在任何更大的尺寸上都会造成真实亏损，而它们之所以被发现，是因为系统被设计成检查自身假设，而不只是报告自己仍在运行。

8 我自己认知上的六处修正

从寻找优势转为证伪优势。我以寻找一个能赚钱的模式开始，最后建成的是一套以尽可能低的成本摧毁候选模式为目的的装置。后一件事才是产出可迁移的那一件。

从胜率转为赔付调整后的保本线。所需胜率等于入场价格这个恒等式是初等的，而我仍然是被一条百战九十七胜却在亏钱的策略教会的。

从信号质量转为执行质量。前两周的大部分时间里，我把问题当作预测问题处理。证据表明，约束条件从来不是预测，而是预测能否被成交。

从机制转为方向依赖。在同一个平台上，用同一套机制，被动入场是毒药，被动出场却是安全的。逆向选择是有方向的，一个在某个方向伤害你的机制，在另一个方向上可能帮助你。我此前一直把做市当成一种单一技术，给它一个单一判决。

从调参转为选择战场。项目后期我不再问该改哪个参数，而是开始问：以我的本金和我的延迟，这个平台上是否存在够得到的优势。这个问题本该在第一周就提出，而正是提出它给出了答案。

从失败转为判决。如果把产出定义为一条能赚钱的策略，这个项目是失败的。如果定义为在成本上限内回答一个研究问题，且答案可以从带时间戳的档案中复现，那么它产出了它被建造出来要产出的东西。比起一个没有记录的正面结果，我更愿意提交一个有记录的负面结果。

9 可以迁移的部分

策略不可迁移，程序可以。下列做法我在管理研究或数据分析的场景中会全部再用一次。

1. **在数据到来之前写下决策规则，并让它可执行**。只写在纸上的规则在时刻到来时会被重新协商。写成代码、带明确样本阈值与明确动作的规则不会。

2. **要么同类比较，要么不比较**。跨越不同机会集的累计值测量的是机会集本身。在共同观测单位上做配对，是最低限度的防御。
3. **把采样中途的估计只当作方向**。仍在移动的数字应当作为排序信息进入日志，而不是作为数值进入决策。
4. **给仪器标注日期，并在修复处切分数据**。修复一个测量过程会划出一条边界，跨越它做合并等于重新引入刚刚被去除的缺陷。
5. **在相信一个有利的总数之前先拆解它**。由一个为负的内核加上一个补偿机制拼出的正结果不会随规模扩大，而拆解通常成本很低。
6. **模拟验证的是相对排序，不是绝对水平**。只要一个模拟量被当作对实际量的预测使用，两者之间的映射就需要它自己的测量。
7. **数学淘汰允许在不动用自由裁量的情况下提前判决**。当剩余样本已不可能改变结论时，提前停止是算术而非不耐烦。那条狙击策略在 260 个观测上被判决而非等到 300，正是因为剩下的 40 笔即使全赢也够不到保本线。
8. **给假设装仪表，而不是给存活装仪表**。上文提到的每一个操作缺陷，都是被那种检验既有假设是否仍然成立的检查抓到的，而不是被确认进程还活着的检查抓到的。

10 局限

真钱样本很小。35 笔订单足以触发一条为该量级书写的预注册规则，却不足以支撑一般性的主张，34% 这个实测胜率是一个区间相当宽的估计。我所辩护的判决是关于一个平台上、一种资金量级下、一种配置的判决，不是关于预测市场这一类别的判决。

纸面与真钱之间的分叉只在单一策略族上被测量过。我倾向于认为该效应的方向可以推广，因为它源自排队机制而非动量策略本身的性质，但我没有在别处检验。

有两条路线只做了推演而没有用真钱关闭。针对第二家交易所的跨平台套利完成了设计，随后在实现之前被我自己砍掉；一条市场中性的凑对策略是在样本一百五十的模拟证据上被判死的，而不是在真实证据上。这两项我都标注为未决而非已答。

最后，本项目只有一名操作者和一个账户，没有复现，而研究日志尽管是在事件发生当天写下的，其书写者正是被它评估的那些决策的作出者。

11 与商科研究生学习的衔接

这个项目中有四条线索位于管理学研究的内部，而不是它的旁边。

第一条是承诺装置与承诺升级。预注册的死刑判据是一个反升级机制，针对的是个体与组织决策中一个已知的失效模式，即按照已经投入的规模继续追加投入到一个正在失败的行动方案里。我建造它是因为我预期自己会需要它，而 7 月 16 日我确实需要了。

第二条是测量效度。一个模拟的关键绩效指标与它对应的实际指标相差一个数量级，同时在排序用途上仍然完全可用。任何依据一个建模指标来掌舵的组织都面对同样的区分，而从一个只为比较而校准过的仪器上读取绝对水平这种错误，并不局限于交易领域。

第三条是信息不对称与逆向选择。这条线索是通过市场微观结构到达我这里的，但它同样是组

织保险核保、采购拍卖与平台设计的机制。把它作为一个操作事实而非一个模型遇见，使相关文献对我变得可读，而这种可读是单纯阅读没有带来的。

第四条是自我报告中的幸存者偏差。项目里最令人不适的发现不是关于市场的。它是，我对自己账户表现的感觉，是由那些幸存下来的片段拼出来的，而一个公开显示的仪表数字与底层资金流动之所以互相矛盾，是因为两者测量的并非同一件事。纠正它需要的是一次完整的对账，而不是一场争论。组织会以更大的规模生产同一种产物，我希望正式地研究它。

就管理学硕士阶段而言，我的意向是研究不确定条件下的决策流程，具体地说，是研究事先固定的规则如何改变写下这些规则的同一批人在此后所作决策的质量。本项目是该问题的一个单案例研究，其中我既是被试也是实验者。这作为证据是一个弱点，也正是我希望用更好的方法去研究它的原因。

12　结语

系统在 7 月 16 日按照为它书写的规则退役，档案被封存而不是被整理，以便那些死路仍然与杀死它们的决策并排可见。到最后，结果本身没有任何令人意外之处。令人意外的部分全部发生在中途：一条已经修好的管道仍在源源不断地产出被污染的答案，一个干净的百分之百胜率原来只有四个观测，一个看起来在下跌的市场其实是一台读错了自己的仪器。

如果这个项目对读者还有价值，那价值在于它演示了一个研究问题可以在很小的成本下被诚实地关闭，而这个关闭的过程值得写下来。

Appendices　附录

Appendix A. Timeline / 附录 A：项目时间线

Date	Event	事件
28 Jun	Project starts. Full paper stack built in one day: paper trading engine, a slow strategy, an arbitrage scanner, a dashboard and a continuous server loop	项目启动。一天内完成整套纸面系统：纸面交易引擎、一条慢速策略、套利扫描器、看板与常驻服务循环
29 Jun	Strategy proliferation. Position sizing with a circuit breaker and append only journals are written, and become the skeleton of every later strategy	策略扩张。带熔断的仓位模块与只增不改账本写成，成为此后所有策略腿的骨架
30 Jun	Move from polling to real time dual WebSocket data. Self audit fixes three real defects and one core logic flaw	从轮询转向双 WebSocket 实时数据。自审修复三个真实缺陷与一处核心逻辑错误
1 Jul	Currency basis defect found. Price source changed. All earlier conclusions from the direction signal voided. Research log created	查出货币基差缺陷，更换价格源，由方向信号得出的既有结论全部作废。研究日志建立
2 Jul	Region measured: the matching infrastructure is in Europe, and an earlier low latency reading was a content delivery artefact	实测撮合设施位于欧洲，此前的低延迟读数是内容分发网络造成的假象
3 Jul	Health checker and live monitor born. Pre-registered verdict protocol adopted and the first A/B opened	体检器与活体监控上线。确立预注册判决机制并开启第一场 A/B
4 Jul	A market making entry strategy judged dead at n = 98, needing 86 percent and holding 77 percent. Passive entry closed permanently. Verdict logic committed as code	一条挂单入场策略在 n = 98 判死，需要 86% 而实得 77%。被动入场方向永久关闭。判决逻辑以代码形式提交
5 Jul	First verdict fires at n = 100. Fee model corrected from an assumed 7 percent to a measured 10 percent, and data split into eras	首份判决在 n = 100 落锤。费率模型由假设的 7% 修正为实测的 10%，数据按纪元切分
6 Jul	Second verdict fires. Exit design changes hands on tail metrics rather than on mean profit	第二份判决落锤。出场设计依据尾部指标而非均值盈利易主
7 Jul	Market neutral pairing strategy judged dead at n = 150, net expected value −207.6. Coverage tax discovery: one variant was missing 39 to 40 percent of the windows its siblings caught	市场中性凑对策略在 n = 150 判死，净期望值 −207.6。发现覆盖税：某变体漏掉了同族其他腿捕获窗口的 39% 至 40%
8 Jul	First real money round trip completed. Boarding configuration fixed by verdict	完成首笔真钱往返。登船配置由判决确定
9 Jul	Sniper strategy judged dead by mathematical elimination at n = 260. Maker orders found to require a five share minimum	狙击策略在 n = 260 因数学淘汰判死。发现挂单存在五股最小数量限制
10 Jul	Execution quality batch: 59 real orders. Resting exit fill rate measured at 74 percent against an assumed 100 percent. Settled winnings found not to auto redeem	执行质量批次：59 笔真单。挂单出场成交率实测 74%，此前假设为 100%。发现已结算赢款不会自动兑付
11 Jul	Cancel and fill race condition found discarding purchased shares. Supervision and restart policy hardened	查出撤单与成交的竞态条件会丢弃已买入份额。值守与重启策略加固
12 Jul	Configuration decision taken: middle tier exit set, capital held flat, live strategy parked until validated	作出配置决定：中间档出场组合，本金不加码，真钱策略在验证前停牌

Date	Event	事件
13 Jul	Live operation resumes under caps. Latency measured end to end. Real transaction cadence found to be about one tenth of simulated	在帽子之下恢复真钱运行。端到端延迟实测。发现真实交易频次约为模拟的十分之一
14 Jul	Attribution at thirteen fills shows entry margin negative and all profit from exit management. Adverse selection at the fill identified	十三笔成交时的归因显示入场边际为负，利润全部来自出场管理。识别出成交时刻的逆向选择
15 Jul	Fade, intra platform arbitrage and the news and macro corner each tested and found empty. Latency measured rather than assumed	反向操作、平台内套利、新闻与宏观题材分别检验，结果均为空。延迟改为实测而非假设
16 Jul	Death test fires at n = 32 against a line of 30. Live strategy retired, servers deleted, home machine stopped, full account audit filed	死亡测试在 n = 32 对判决线 30 落锤。真钱策略退役，服务器删除，家用机停机，完成账目终审

Appendix B. Data provenance / 附录 B：数据出处

Every quantitative claim in this report is traceable to one of four places in the project archive, all of which carry dates that were written at the time rather than reconstructed afterwards.

本报告中的每一项定量陈述都可以回溯到项目档案中的四类来源，这四类来源的日期都是当时写下的，而非事后重建的。

1. **The research log.** A single running document updated the same day any finding appeared, including the finding that a previous entry in it was wrong. Retractions were added next to the original statements rather than replacing them.
   研究日志。一份持续更新的文档，任何发现当天入档，包括此前某条记录出错这一发现。更正被写在原陈述旁边，而不是替换原陈述。
2. **Pre-registered verdict scripts and their recorded outputs.** One script per strategy that mattered, each committed before the sample it judged was full, and each leaving a frozen result file when it fired.
   预注册判决脚本及其执行记录。每一条重要的策略各有一份，每一份都在其所审判的样本填满之前提交，落锤时留下冻结的结果文件。
3. **Append only ledgers and execution logs.** One ledger per strategy, never edited in place, plus a separate log of real order intents and fills used for the execution quality measurements.
   只增不改的账本与执行日志。每条策略一份账本，从不原地修改，另有一份记录真实订单意图与成交的日志，用于执行质量测量。
4. **The repository commit history.** Several hundred commits across nineteen days, each carrying the decision it implemented and, where applicable, the instruction that prompted it.
   代码仓库提交历史。十九天内数百次提交，每一次都载明它所执行的决定，以及在适用时载明触发该决定的指令。

Three qualifications apply to specific figures. The reported industry wide latency figures for this venue, namely an average of about twelve seconds in 2024 narrowing to a few seconds by early 2026, come from third party on chain analyses that I read rather than from my own measurement, and I have not verified them independently. The 34 percent realised win rate is computed on filled orders only and

rests on a small sample. Figures described in the text as snapshots were recorded while their samples were still growing and are reported here for direction and ordering only, in keeping with the rule stated in section 3.3.

有三项限定适用于特定数据。文中提到的该平台全行业延迟水平，即 2024 年平均约十二秒、至 2026 年初收窄至数秒，来自我阅读的第三方链上分析，而非我本人的测量，我没有独立验证过。34% 这一实际胜率仅在已成交订单上计算，且样本量小。文中被称为快照的数字，是在其样本仍在增长时记录的，此处仅用于说明方向与排序，这与第 3.3 节陈述的规则一致。

References / 参考文献

Bailey, D. H., Borwein, J. M., López de Prado, M., & Zhu, Q. J. (2014). Pseudo-mathematics and financial charlatanism: The effects of backtest overfitting on out-of-sample performance. *Notices of the American Mathematical Society*, 61(5), 458–469.

Glosten, L. R., & Milgrom, P. R. (1985). Bid, ask and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, 14(1), 71–100.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11), 2600–2606.

Thaler, R. H., & Ziemba, W. T. (1988). Anomalies: Parimutuel betting markets: Racetracks and lotteries. *Journal of Economic Perspectives*, 2(2), 161–174.

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives*, 18(2), 107–126.

These five references are the only external sources cited. They are used for the established concepts the project's own findings map onto, namely backtest overfitting, adverse selection in quoted markets, preregistration as a method, the favourite longshot bias, and the interpretation of prediction market prices as probabilities. No claim about my own results depends on them.

以上五项是本文引用的全部外部来源，用于标注本项目自身发现所对应的既有概念，即回测过拟合、报价市场中的逆向选择、作为方法的预注册、冷门偏误，以及把预测市场价格解释为概率。本文关于自身结果的任何主张都不依赖于它们。