ZIMMËR TRADING / WORKING PAPER 04

Learning to Keep the Edge

The First Seven Weeks of Zimmër Trading and a Kalshi-First Research Agenda

De'Mondre Zimmerman · November 7, 2026 · Version 1.0

1. The ambition and the record

I began Zimmër Trading with a large ambition: build a research process that could eventually support a trading firm. The first seven weeks have supplied a more immediate problem. I need to distinguish an advantage from a favorable run, measure both correctly, and keep the cost of learning under control. This paper documents that progression from September 21 through November 7, 2026. It extends my earlier working papers rather than replacing their dated observations.

The initial project connected energy fundamentals, weather, market data, and a paper-broker account. Its question was whether physical-market information could improve a trade. A useful electricity-demand forecast does not automatically forecast a producer's earnings or identify an underpriced stock. Corporate hedges, market expectations, timing, and execution each require evidence. Working Paper 01 describes that initial agenda [1].

Two experimental records followed: small real-money Kalshi trades and simulated equity/ETF trading. The Kalshi commodity strategy V3 produced a profitable initial run before a sharp reversal. Meridian added a separate forecasting research program. Combining live, paper, and modeled dollar outcomes would produce a misleading track record.

At the November 7 review, the reconstructed Kalshi record contained 281 exchange fill records across 181 completed markets, with net trading P&L of -$41.8293 after recorded exchange fees. The fills ran from September 25 through October 12; the last settlement was October 13. This is a reconstruction from the desk's synchronized records, not an independently audited account return. Deposits, withdrawals, a capital denominator, taxes, and full operating costs have not been reconciled here. The equity figures reported later are simulated product-level price P&L, including marks on open positions.

My recollection that the commodity strategy worked for a while is supported in a specific sense: its early realized cashflows were positive. That does not establish that its expected return was positive or that its edge disappeared on an identifiable date. The records let me study what happened, identify weaknesses, and design a cleaner next test. They do not yet establish a durable profitable business.

I now want to learn one Kalshi market family deeply and earn the right to expand. Alex Gerko's XTX and Jim Simons inspire the ambition; their achievements are not a forecast for Zimmër. The next milestone is a reproducible advantage after costs.

2. What the live ledger actually says

A settlement row is not always the complete cashflow of a position. An early reduction can be represented by purchasing the complementary outcome. The resulting YES/NO pair releases one dollar. In the reviewed ledger, summing raw settlement P&L gave -$132.2893; including $90.4600 of offset-pair cash release gave -$41.8293. An independent calculation from all fills, economic entry costs, terminal payouts, and fees reached the same total. No correction is added again to an already reconciled result.

The following table uses one attribution rule: the strategy owning each filled order receives that fill's economic result. Every fill in the reviewed sample matched an owner and a terminal market result. Figures include recorded exchange fees, exclude operating overhead and taxes, and do not describe returns on account capital.

Live strategy	Completed markets	Net trading P&L
V3 commodity	46	-$39.8168
VSS4 commodity	10	-$14.7000
V3S commodity	5	-$4.7500
Original quiet commodity	33	-$3.5100
Tennis pre-event	25	-$2.6970
Weather market making	1	-$1.1684
NO-side	7	-$0.8100
Mentions Maker	9	+$0.9734
Macro market making	4	+$4.7478
ITF pre-event tennis	41	+$19.9017
All reviewed strategies	181	-$41.8293

The positive ITF result deserves further study, but selecting it after comparing strategies creates another research choice. The next test must include new eligible opportunities and rejected orders, not only its historical winners. Mentions Maker has a small positive sample. Macro's positive total is concentrated: removing one completed market's $5.55 contribution makes that sleeve negative. Across Macro, Mentions, and Weather, combined live P&L was +$4.5528; excluding that same market leaves -$0.9972. These are concentration diagnostics, not reasons to erase a legitimate gain.

V3's early record is particularly informative. Through October 6, 30 of 31 completed markets were profitable and cumulative realized P&L was +$72.1662. The full September 29–October 12 sample finished at -$39.8168 across 46 markets. Working Paper 03 reproduces those figures from its public ledger [3]. Paper 02's earlier 88.1% headline counted 37 positive closing lots out of 42, spanning 39 markets through October 9. The corresponding market count is 34 profitable out of 39, or 87.2%. Different observation units must never be treated as interchangeable [2].

3. Keeping a winning streak from setting the budget

Three consecutive V3 market losses totaled $128.22. Their significance is not that three losses are extraordinary. It is how much exposure the desk carried when they arrived. For a binary contract bought at 90 cents and held to a zero-or-one settlement, one win makes 10 cents and one loss costs 90 cents before fees. A high win rate is therefore compatible with a losing strategy. In the separate VSS4 experiment, seven wins and three losses produced -$14.70.

The public V3 reconstruction makes a useful counterfactual explicit. Holding historical selections, prices, exit proportions, outcomes, and proportional fees fixed, limiting each completed market to ten opening contracts changes the full-window result from -$39.8168 to +$2.8648. This calculation assumes those smaller positions would execute at the same terms. The cap was examined after seeing the outcomes, and several alternative caps produce different totals. It is evidence of sizing sensitivity, not proof that ten contracts is optimal.

Sizing was not the whole story. Equal-market mean net outcome per opening contract fell from +13.12 cents before October 7 to -18.40 cents over October 7–12. Those periods were selected retrospectively. Selection, entry prices, exits, and outcomes all differed. The evidence supports both exposure amplification and weaker realized unit outcomes; it does not isolate greed, a market regime change, or a particular signal defect as the sole cause.

The next risk design should separate three budgets: loss allowed on one event, combined loss allowed across related events, and loss allowed over the whole experiment. For a fully paid binary long held to settlement, worst-case trading loss is acquisition cost plus applicable fees. Pending orders must reserve that loss room too. Related commodity contracts and neighboring expiry windows can share the same price shock. Summing their worst-case losses is a conservative starting point when a reliable joint model is unavailable.

Profits should not automatically enlarge the experiment’s authorized loss allowance. Increasing size should require a scheduled review of fresh evidence, with a recorded reason and a separate operator decision. A high-water-mark rule can reduce new risk after a drawdown, but it cannot guarantee a realized loss ceiling when positions are already exposed. A stop or cancellation request is not a guaranteed fill or cancellation. Limits on exposure before entry do work that an emergency exit cannot promise to do.

I would freeze the signal and sizing policy together for the next test, record forecast and execution diagnostics alongside P&L, and define in advance what pauses new entries. Operational failures, stale prices, unresolved inventory, and exhausted loss room should have explicit handling. Statistical deterioration should be assessed under a predeclared monitoring rule; repeatedly redesigning after losses would contaminate the evaluation. No new risk threshold or trading control was activated for this paper. The objective is to make disciplined behavior a property of the system, rather than something dependent on how confident I feel after a good week.

4. Learning from Roll and Tide

The equity/ETF record covers 31 paper sessions from September 25 through November 6. Replaying the recorded opening handover, allocated simulated fills, and stored closing prices reproduced all 124 product-day P&L rows across four products within one cent per row. That checks the desk's price-P&L calculation. It does not establish live fill quality or a complete economic return including dividends, borrow, financing, and other charges. Alpaca explicitly describes its paper environment as a simulation [5].

Roll recorded +$291.69, compared with +$264.83 under the desk's model execution convention. Its cumulative daily-P&L series still suffered a $257.08 peak-to-trough drawdown. Its three best days contributed $230.30, leaving +$61.39 after their removal. This deletion is a descriptive concentration check, not an alternative strategy. A profitable final total and substantial interim pain can coexist.

Despite its name, Roll's reviewed implementation is a time-series momentum rule on UNG, a natural-gas ETF. It uses a 90-session lookback with a 10-session skip and volatility-based sizing. The target signal was short on 29 sessions and long on two. A positive result here does not demonstrate an independently measured futures-curve or roll-yield premium. UNG's instrument structure and shorting costs need their own treatment before any live inference.

Tide is a broader trend system across 12 ETFs, using a 252-session lookback, a 21-session skip, volatility weights, and portfolio exposure caps. It recorded -$1,477.00, with model price P&L of -$1,514.83. November 5 alone contributed -$1,260.10. Removing that day would still leave -$216.90; it cannot be dismissed as the only problem. On that day, SPY, IWM, QQQ, EFA, XLK, and XLF together contributed about -$1,176.56. Several symbols expressed overlapping equity risk.

There were useful positive components inside Tide: the reconstructed UNG sleeve contributed approximately +$235.18, XLF +$200.80, and XLU +$130.23. TLT contributed -$524.12; XLK and QQQ also lost. These are rounded symbol-level price-P&L reconstructions. Small differences from the sum of published daily figures arise because those figures are rounded by product and day. Dropping the losing symbols now would select on the same history used to judge success.

Roll and Tide therefore contain both a lesson and a trap. Both profited from a short UNG component, so they are partly exposed to the same underlying move. Their observed daily P&L correlation was about -0.10 over only 31 sessions, which is insufficient to establish durable diversification. Redundancy should be tested through shared exposures, simultaneous losses, and stressed periods, not inferred from product names.

The proposed equity research keeps each existing rule as a baseline. Roll gets a limited horizon-robustness and execution-cost study against a matched passive-short comparison. Tide gets an attribution study of common equity exposure and one prespecified challenger with tighter group risk limits. Both need untouched forward paper observations. A challenger should improve a declared objective after costs and add something at the portfolio level. This paper identifies experiments; it does not report a completed optimization or a newly validated winner.

5. Market making is a research question

Market making appeals to me because it combines pricing, software, execution, and repeated decisions. XTX describes using machine-learning forecasts to trade across venues and provide liquidity to clients [6]. The relevant lesson is the integration of forecasting and execution. Its scale does not establish that posting a bid and an offer will be profitable for a smaller participant, or that a method developed on Kalshi will transfer unchanged to another market.

A resting order is an invitation for another trader to choose when to trade with me. If the order fills mainly when my price has become stale, apparent spread income can be overwhelmed by adverse selection. Research on Kalshi motivates examining execution roles, product categories, and information effects, but findings from broader market samples do not validate my particular orders [7, 8]. Directional positions entered passively are also not automatically evidence of a profitable two-sided market-making process.

Zimmër's own research provides a warning. In the historical frozen FV-1 study, the row-weighted Brier score was modestly better than the market-midpoint baseline. Yet the associated maker simulation produced an average -4.072 cents per modeled fill across 23,921 fills. Its trade-through assumptions did not establish actual queue execution. In the later October 15–November 7 snapshot, the row-weighted Brier score was slightly worse than the market baseline, while equal-cluster weighting gave a small improvement. The corresponding prospective taker simulation lost about 1.463 cents per modeled execution. These are research outcomes, not real-money losses.

The next Kalshi experiment should isolate one family where the payoff rules, information arrivals, and trading hours can be documented. A baseline and one challenger should share the same eligible opportunities and risk budget. The log should retain quote arrival, decision, submission, acknowledgment, modification, cancellation, fill, and settlement times, plus rejected decisions and missing-data reasons. Queue position must be observed or treated conservatively as uncertain. Both sides of an intended round trip must be feasible.

The scorecard should separate spread earned, subsequent price movement, terminal inventory P&L, fees, incentives, and overhead. Markouts at short and longer horizons can show whether fills tend to precede unfavorable moves. A comparison of all opportunities with filled opportunities can identify selection patterns, but does not by itself identify a causal adverse-selection cost. The headline remains realized net results under the actual execution process, with uncertainty grouped by relevant event and day.

New prediction strategies should begin their forward execution tests in DEMO under the project's operating rules. Where a comparable contract or executable quote is absent, the record should show that gap. An optimistic simulated fill cannot fill it. DEMO then remains a limited operational test; a separately authorized small live test is needed to learn about real execution. Official designated-market-maker status is a further step: Kalshi describes approval, financial-resource review, and ongoing two-sided quoting obligations [4]. This paper neither applies for that status nor assumes associated fee benefits.

6. A record worth expanding

My immediate goal is to become competent at a narrow Kalshi task and document the evidence well enough that another researcher can understand why I continued, changed, or stopped. That requires two forms of redundancy. Operational redundancy concerns reliable data, recovery, and cancellation. Strategy diversification concerns different economic sources of return. Another variant of the same commodity favorite rule may improve measurement, but it does not necessarily diversify losses.

The first research priority is accounting. Live reports should reconcile fills, early reductions, settlements, fees, and remaining inventory under a single economic convention. Paper reports should separately identify price P&L and omitted cashflows. Every public statistic should carry a cutoff, observation unit, mode, cost basis, and source. A result can change when the window extends or an error is corrected, but the revision must say what changed.

The second priority is to preserve useful baselines. ITF and Mentions have positive live samples worth investigating; Roll has a positive paper sample. Those observations justify allocating research attention, not presuming profitable future trading. Each challenger should change one motivated feature, keep the baseline visible, and record all attempted variants. This reduces the temptation to remember only the latest winner. Backtest-selection research explains why that discipline matters [9].

The third priority is execution evidence. A study should preregister its horizon, minimum event coverage, comparison, cost assumptions, and stopping or review rules. Elapsed days alone are not enough, and hundreds of fills may arise from a handful of correlated events. The observation window should be extended when uncertainty remains large. A candidate should survive realistic latency and queue assumptions, leave-one-event diagnostics, and portfolio-level concentration checks. A positive outcome driven by one settlement is weaker evidence than a result that persists across distinct episodes.

The fourth priority is independent risk approval. A good research result should not silently increase size, enable a strategy, or replenish a loss allowance. Reviews should distinguish an implementation repair from a new signal, and distinguish improved expected return from permission to take more risk. Small controlled experiments make it possible to learn while retaining the ability to stop. A successful firm would need this discipline at a much larger scale, with reliable operations and economics after fixed costs.

Expansion into stocks, futures, or options should follow a specific hypothesis about what transfers. Data timing, experiment design, and ledger discipline can transfer. Contract payoffs, funding, borrow, margin, transaction costs, and feasible execution need market-specific work. Options additionally require explicit treatment of nonlinear exposures and changing volatility. Larger markets offer capacity and competition together. They are not an automatic reward for completing a Kalshi stage.

The early Zimmër story contains real gains, real losses, useful software, incomplete ideas, and corrected measurements. I want the next chapter to be narrower and more convincing: a rule specified before the outcome, fills I can reconcile, a risk budget I do not renegotiate after a streak, and a result that earns another test. That is the path toward a credible record. The scale of any eventual business remains something the evidence must support.

Sources, methods, and revision policy

Review cutoff: November 7, 2026. The live trade population was rechecked at publication preparation and remained 281 fills, with the latest fill on October 12. Equity reporting ends November 6. The Meridian figures refer to the fixed November 7, 16:42 UTC review snapshot; the research collector may continue afterward. Dollar amounts are USD. Live strategy totals are after recorded exchange fees and before overhead and taxes. Paper results and modeled executions are explicitly separate.

This is an author-operated case study, not an independent audit or peer-reviewed claim of investment performance. The source evidence consists of synchronized exchange records, product configurations, allocated paper fills, opening handover records, stored price bars, historical research receipts, and versioned code. A public aggregate data file supports the displayed totals. It intentionally omits account access details and raw private order records; it cannot independently authenticate the underlying exchange or broker records.

Working Paper 03's existing public replication package was rerun successfully. Its V3 total agrees with the all-fill strategy reconstruction. Equity reconciliation used opening handover quantities, each day's allocated fills, and stored closes. For each symbol, price P&L equals opening quantity times the change in closing price, plus the sum of signed fill quantity times closing price minus fill price. Daily product rounding explains minor differences between symbol totals and summed daily reports.

Consistency means using comparable definitions and documenting revisions. It does not mean preserving an old hypothesis after contrary evidence arrives. The accompanying research review records the correction to Paper 02's observation unit and cutoff, distinguishes Paper 01's historical design from later implementations, and supplies a checklist for future publications. New evidence may revise a conclusion; unexplained changes in scope or arithmetic should not.

AI tools assisted data extraction, reconciliation code, drafting, and editing. The author operated the desk. Version 1.0, published November 7, 2026. This paper reports research and proposed next tests; it does not authorize or announce any trading-control change.

1. Zimmerman (2026). Building a Research Process for Energy Markets. Working Paper 01. Source

2. Zimmerman (2026). How an 88% Win Rate Lost Money. Working Paper 02; correction dated November 7. Source

3. Zimmerman (2026). The Price of a Winning Streak. Working Paper 03, revision 1.1; public ledger and replication package. Source

4. Kalshi. How to Become a Market Maker on Kalshi. Accessed November 7, 2026. Source

5. Alpaca. Paper Trading. Simulation scope and omitted effects. Accessed November 7, 2026. Source

6. XTX Markets. Firm overview: forecasting and liquidity provision. Accessed November 7, 2026. Source

7. Bürgi, Deng and Whelan (2026). Makers and Takers: The Economics of the Kalshi Prediction Market. CESifo 12122. Source

8. Bartlett and O'Hara (2026). Adverse Selection in Prediction Markets: Evidence from Kalshi. Working paper, revised August 12. Source

9. Bailey, Borwein, López de Prado and Zhu (2015). The Probability of Backtest Overfitting. Working paper. Source