SEMANTIC EXPERT FOLLOWING IN PREDICTION MARKETS

Anonymous authors
Paper under double-blind review

ABSTRACT

We develop an expert-following strategy for binary prediction markets using public wallet histories. The central idea is that expertise is local to the question being traded. A fixed text encoder represents market questions, and a thresholded similarity kernel gives greater weight to a wallet's performance in closely related, settled markets. The strategy first waits for a market to meet observable volume, price, and scheduled-horizon conditions. It then aggregates qualified wallet flow, requires support from several distinct addresses, and controls both signal concentration and capital exposure. No predictive model is trained on trading outcomes. We evaluate four allocation configurations on a historical Polymarket archive with 41,891 candidate markets and 183.74 million valid public trade records. In a delayed, quantity-unconstrained price replay, the primary configuration returns 257.25% with a 13.64% daily maximum drawdown. A higher-exposure configuration returns 318.32% with a 19.02% maximum drawdown. The comparison shows how local expert selection and capital allocation interact within a common consensus-based strategy.

1 INTRODUCTION

Prediction markets transform dispersed beliefs about future events into prices that can be interpreted as probabilities (Wolfers & Zitzewitz, 2004). Their public transaction histories create a related empirical question: can a trader identify participants with relevant prior skill and profitably follow their subsequent order flow? Evidence from corporate prediction markets suggests that experienced participants and those with stronger past performance help correct market inefficiencies (Cowgill & Zitzewitz, 2015). More generally, forecast aggregation can improve when persistent positive contributors are identified rather than treating all opinions equally (Budescu & Chen, 2015; Atanasov et al., 2017).

Polymarket provides a useful setting for this question. Trades are associated with public proxy wallets, binary contracts usually settle at zero or one, and question text spans elections, sports, macroeconomics, cryptocurrency, and other domains. Historical success may therefore be context-dependent. A wallet that trades political markets well need not possess comparable information about sports or monetary policy. We operationalize this intuition by embedding market questions and weighting each wallet's settled-market record according to its semantic proximity to the current question. A thresholded, squared similarity kernel discounts weakly related histories and emphasizes nearby questions. Qualifying current trades are then aggregated into a skill–quantity weighted directional signal.

The timing of observation matters alongside the choice of experts. A newly listed question can attract speculative flow long before information becomes concentrated. We therefore use a rule-based market activation procedure: a market must accumulate sufficient observed dollar turnover, have recent trading activity and a sufficiently directional quantity-weighted average price, and approach its scheduled close. These conditions use the trade prefix at each decision time rather than the market's eventual trade count. Historical expert records are drawn from activity after a separate cumulative-volume admission rule and enter the score only after their settlement label is released in the replay.

Following high-price contracts creates an important risk problem. A strategy can settle profitably in almost every market and still suffer a substantial drawdown. High win rates are especially

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

uninformative when positions differ in size and a losing contract can forfeit its entire notional. Prediction-market probabilities also exhibit horizon-dependent calibration patterns (Page & Clemen, 2013), and average market accuracy does not remove the need to identify failure-prone contracts (Tai et al., 2019). Our strategy limits the largest wallet's share of same-direction flow, screens for related-history breadth, and caps exposure according to event type. Distinct addresses are not assumed to identify independent people.

We study four capital-allocation configurations on the same historical event stream. The primary configuration returns 257.25% with a 13.64% maximum drawdown; a balanced configuration returns 244.18% with an 11.95% drawdown, while a higher-exposure configuration reaches 318.32% with a 19.02% drawdown. These are retrospective, quantity-unconstrained price-replay results, not estimates of executable market capacity. The accompanying sensitivity analysis examines how the results change with expert-selection and allocation settings.

This paper makes three contributions. First, it combines local semantic conditioning of wallet skill with delayed replication of current expert flow. Second, it combines observable market activation with wallet-level consensus, related-history breadth, and event-sensitive exposure limits. Third, it examines how the speed and size of capital deployment change returns and drawdowns under a shared signal. The resulting framework is rule-based and requires no fitting of an outcome-prediction model.

2 Data and Sample Construction

2.1 Sources and analysis sample

We use market metadata from Polymarket's Gamma API and public transactions from its Data API, requested with `takerOnly=true`. A record identifies a proxy wallet, condition, outcome token, side, quantity, price, and timestamp. Token identifiers are matched to the ordered outcome-token mapping in market metadata; the feed's outcome-index field is not used when it conflicts with that mapping. YES denotes the first outcome token and NO the complementary token, including binary questions whose displayed outcome names are not literally YES and NO.

The archived candidate inventory contains 41,891 binary markets. It combines a 10,000-market inventory ranked by an archived metadata-volume snapshot with additional markets meeting a 250,000-USDC metadata-volume collection screen; one unsupported multi-outcome market is excluded. Within this retrospectively assembled inventory, trading eligibility is determined by cumulative turnover observed during replay. Appendix A.1 describes the collection procedure, and Section 7 discusses historical coverage.

Transaction collection covers September 18, 2023, 00:00:00 UTC through August 8, 2026, 06:42:30 UTC. Pagination continues to the API's exhaustion of that requested window rather than imposing a fixed per-market row limit. The archive contains 183,742,198 raw records; six invalid-price records are quarantined, leaving 183,742,192 valid records. Cumulative turnover is the sum of actual-token price times quantity from the observation boundary onward. It is therefore observed-window turnover, not necessarily lifetime turnover for markets that began earlier.

Table 1: Historical archive and strategy eligibility

Quantity	Count
Candidate binary markets	41,891
Valid public trade records	183,742,192
Markets with usable outcome labels by the window end	39,864
Markets with a recorded volume-only admission	4,785
Markets supplying nonempty admitted expert histories	4,458
Admitted wallet–market history records	13,685,164
Markets activated by the common trading rule	1,469

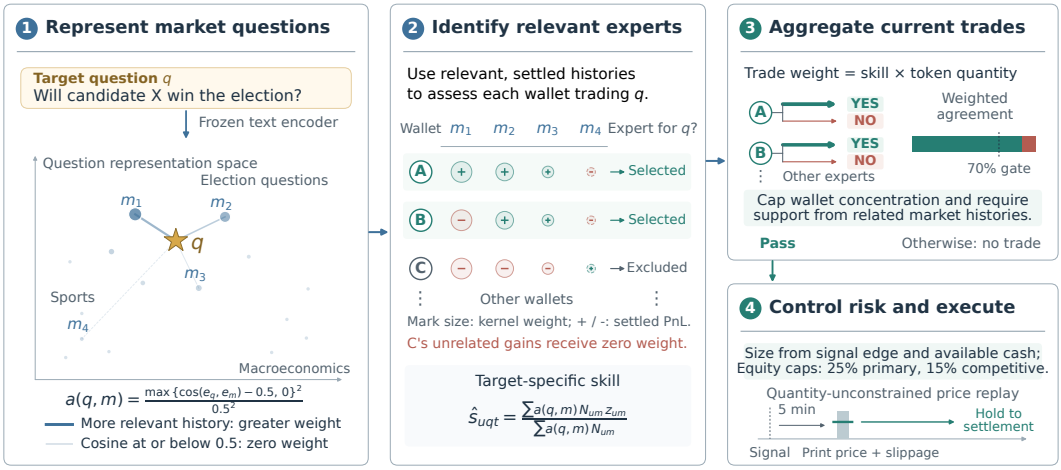


Figure 1: Semantic expert following after market activation. Question proximity weights settled wallet histories; qualified current trades form a directional consensus; wallet concentration, historical support, and position caps govern entry. The points and wallet records are illustrative, not an empirical embedding projection. The final panel depicts delayed price replay, not available order-book quantity.

Recorded volume admissions use the pre-resolution prefix when an explicit event boundary exists, and the observed window otherwise. A recorded admission without a usable outcome label cannot supply expert history.

2.2 Outcome timing and wallet records

We reconstruct outcomes and their release times from archived market metadata. A label requires a closed market, a resolved UMA status, an explicit resolution date, and an unambiguous terminal outcome-price proxy. Its event boundary is the latest valid explicit resolution date in the archived metadata, and the label is released one second later, after the complete timestamp group has been processed. Ambiguous or unavailable labels are excluded from expert histories. Appendix A.1 lists the metadata fields.

For historical market m, let b_m be the first second after a completed timestamp group raises observed cumulative turnover to one million USDC. Wallet histories use all trades in $[b_m, r_m)$, where r_m is the resolution event boundary. The crossing-second trades are excluded. The resulting score becomes visible only at the label-release time, and the target market is always excluded from its own experts' histories. This separates historical evidence admission from the stricter activation of a current trading opportunity.

Question representations are frozen 1,024-dimensional vectors from `BAAI/bge-large-en-v1.5` (Xiao et al., 2023); the checkpoint revision is recorded with the inputs. No text encoder or forecasting model is fitted to the backtest. Question wording and scheduled close dates use archived metadata versions. Public taker records omit maker activity, transfers, and inventory acquisitions outside the observation interval.

3 Semantic Expert-Following Strategy

The strategy answers four questions: when to observe a market, which wallets have relevant experience, whether their current trades agree, and how much capital to allocate. Figure 1 illustrates the expert-selection and allocation pipeline after market activation.

3.1 Observable market activation

The activation rule waits for a market to become sufficiently active and directional before collecting expert flow. At each completed UTC hour t, we compute observed cumulative actual-token turnover

$V_q(t)$, the number of prints in the past 24 hours $n_q(t)$, and their quantity-weighted YES-equivalent price $\bar{p}_q(t)$. A NO-token print at price p contributes $1-p$ to the YES-equivalent price, while its dollar turnover remains its actual price times quantity. With scheduled close T_q, activation is the first hour satisfying

$$V_q(t) \geq 10^6, \quad n_q(t) \geq 100, \quad \max\{\bar{p}_q(t), 1-\bar{p}_q(t)\} \geq 0.90, \quad 0 < T_q - t < 28 \text{ days}. \tag{1}$$

At least 24 hours must also have elapsed since the first observed print. All bars use trades strictly before their hour boundary. Activation occurs once: later fluctuations do not reset the market or erase accumulated flow. It only enables signal construction; it does not itself create an order or select a direction. Historical wallet scoring uses the volume-only admission rule in Section 2, not this full activation test.

3.2 Wallet settlement score

The first step measures expertise by the settlement value of observed trading, rather than a count of correct market calls. Repeatedly buying the winning token near one dollar has less replication value than reaching the same conclusion at a lower price. For wallet u in historical market m, let Q^Y_{um} and Q^N_{um} be net token quantities over the admitted history, C_{um} signed cash flow, and N_{um} absolute traded notional. Only records with $N_{um} > 0$ are retained. Buys enter cash flow negatively and sells positively. With first-outcome payout $\omega_m \in [0,1]$, define

$$\Pi_{um} = C_{um} + \omega_m Q^Y_{um} + (1-\omega_m) Q^N_{um}, \qquad z_{um} = \text{clip}\left(\frac{\Pi_{um}}{N_{um}}, -1, 1\right). \tag{2}$$

Here Π_{um} values trades in the admitted interval $[b_m, r_m)$ at settlement. A sale contributes its proceeds minus the tokens' settlement value, regardless of when the original inventory was acquired. Thus z_{um} measures the value per traded dollar of the included activity rather than the wallet's total return.

3.3 Semantic context-conditioned skill

Expertise may be local to a topic. Rather than assigning questions to discrete categories, we weight each wallet's settled histories by proximity to the current question. Historical notional weights the evidence, and a locality kernel suppresses weak similarities across unrelated questions.

Let e_m be the fixed BGE question vector and $s(q,m) = \cos(e_q, e_m)$. Fixed sentence vectors support efficient semantic comparison (Reimers & Gurevych, 2019). We use

$$a(q,m) = \left(\frac{\max\{s(q,m) - \theta, 0\}}{1-\theta}\right)^\gamma, \qquad \theta = 0.50, \quad \gamma = 2. \tag{3}$$

Thus a cosine similarity below 0.50 contributes zero, and the remaining weights grow quadratically toward one. At time t, only markets with released labels enter wallet u's history $\mathcal{H}_{u,t}$. Its target-specific skill is

$$\widehat{s}_{uqt} = \frac{\sum_{m \in \mathcal{H}_{u,t}} a(q,m) N_{um} z_{um}}{\sum_{m \in \mathcal{H}_{u,t}} a(q,m) N_{um}}. \tag{4}$$

A current observation qualifies when $\widehat{s}_{uqt} \geq 0.03$ and the denominator is at least 10 USDC. Empty weighted histories do not qualify.

3.4 Wallet-level consensus

The consensus signal combines the direction and size of qualified wallet trades. A cap on each wallet's share spreads support across addresses, while a separate history-support requirement favors experience across related markets.

Each qualified trade i observed after activation receives weight $w_i = \widehat{s}_{uqt_i} v_i$, where t_i is its arrival time and v_i is token quantity. Historical notional measures the amount of settled trading used to estimate skill; current token quantity measures exposure to the outcome payoff. Buying YES or selling NO supports YES; buying NO or selling YES supports NO. A wallet can contribute to both directions. We accumulate qualifying observations from activation through t, keeping each

observation's skill weight as measured when it arrived. There is no rolling expiration. Directional consensus is

$$R_{dqt} = \frac{\sum_{i \in d} w_i}{\sum_i w_i}, \qquad R_{dqt} \geq 0.70. \tag{5}$$

Let W_{udqt} be wallet u's accumulated weight in the chosen direction. The largest wallet may contribute at most 75% of this directional weight:

$$C_{dqt} = \frac{\max_u W_{udqt}}{\sum_u W_{udqt}} \leq 0.75. \tag{6}$$

For history weights $h_{um} = a(q, m)N_{um}$, effective historical support is $(\sum_m h_{um})^2 / \sum_m h_{um}^2$. Its w_i-weighted mean among observations supporting the chosen direction must be at least 1.5. This aggregate test rejects narrow collective experience without requiring every wallet to have the same number of past markets.

3.5 Semantic event risk and capital allocation

Signal selection and position size are separate decisions. Even a well-supported binary position can lose its entire notional. A transparent question-text rule identifies head-to-head sports contracts and questions about winning a named competition, and caps these competitive-event positions at 15% of equity. The classifier affects exposure, not direction; its exact keyword rules appear in Appendix A.3.

At submission, the chosen token must satisfy $0.60 < p \leq p_{\max}$, $R_d \geq p$, and a scheduled remaining horizon of $5 < \tau < 28$ days. Let E be marked-to-market equity, B cash, O open-position acquisition cost, and P pending requested notional. The remaining deployment target and allocation fraction are

$$G = \max\{0, 0.97E - O - P\}, \qquad f = \text{clip}\left(\lambda(R_d - p)\frac{365}{\tau}, 0.05, f_{\max}\right). \tag{7}$$

Here τ is measured in days, and f is the fraction of remaining deployment capacity G allocated to the order. The gap $R_d - p$ is a sizing heuristic; consensus is not a calibrated probability. The requested notional is

$$N = \min\{Gf,\ \max(0, B - 0.05E),\ cE,\ c_q E\}, \tag{8}$$

where c is the general position cap and $c_q = 0.15$ for competitive events (unbounded otherwise). Orders below 10 USDC are omitted. Pending requests reduce G but do not debit cash; actual fills remain cash-limited. Position caps apply at submission and are not continuous rebalancing rules.

Table 2: Four capital-allocation configurations. All other strategy settings are shared.

Configuration	λ	$f_{\max}$	General cap c	$p_{\max}$
A: Conservative allocation	0.25	0.35	25%	0.98
B: Faster allocation (primary)	1.00	0.50	25%	0.98
C: Balanced allocation	0.50	0.35	35%	0.98
D: Higher exposure	1.00	0.50	35%	0.97

Table 2 varies the speed of deployment and the admissible size and price of a position. B can deploy more rapidly than A without raising the 25% position cap. C combines a slower deployment rule with a larger cap. D combines faster deployment and a larger cap with a tighter entry-price ceiling. The configurations share expert selection and consensus; they need not fund exactly the same opportunities because available capital differs.

Each market permits one pending request at a time and at most one filled position, held to settlement. An order becomes eligible five minutes after submission. The next market print is converted to the chosen token's price, including the complementary price when the print is for the other outcome; the simulated buy price is $\min\{1, 1.001p\}$, including 10 basis points of slippage. Price bounds are checked again. A failed check cancels that request, although a later qualifying signal may submit another. There is no finite order-expiry rule or consensus recheck at execution. One print can fill the cash-funded request irrespective of its observed quantity; any unfunded remainder is not carried as a residual order. The horizon test applies at submission. There is no borrowing or external capital addition.

4 EVALUATION DESIGN

Every configuration begins with 10,000 USDC and uses the same September 18, 2023–August 8, 2026 event stream. Histories become available according to the label schedule in Section 2; market activation and entry use the observed trade prefix. Cash carries forward throughout the window and is not reset at calendar boundaries. The main replay uses zero additional fees and 10-basis-point slippage; separate stresses vary fees, slippage, and delay.

Equity is cash plus the value of open inventory at the latest observed outcome-equivalent price. Daily records use the final processed event of each UTC day, with the final replay state included. With initial capital E_0, the cumulative return and drawdown paths are

$$r_t = E_t/E_0 - 1, \qquad d_t = E_t/\max\{E_0, \max_{s\le t} E_s\} - 1. \tag{9}$$

Maximum drawdown is reported as the positive magnitude $-\min_t d_t$. Closed-position counts exclude unresolved inventory, but portfolio returns and drawdowns include it. Profitable settlements do not eliminate interim mark-to-market losses.

The analysis is retrospective. Parameters and reported configurations were chosen after inspecting the same archive; the absence of model training does not make the evaluation out of sample. Local sensitivity changes one parameter at a time around B. Additional allocation and activation settings are reported with their reference configuration made explicit. All comparisons retain the same metadata-label and quantity-unconstrained fill conventions.

5 EMPIRICAL RESULTS

5.1 OVERALL PERFORMANCE

Table 3 reports the four configurations from Table 2. B grows 10,000 USDC to 35,725 USDC, a 257.25% return, with a 13.64% maximum drawdown. C has the lowest drawdown among the four configurations, 11.95%, and returns 244.18%. D has the highest terminal return, 318.32%, but its 19.02% drawdown is larger. A returns 200.76% with a 12.72% drawdown. Thus the largest terminal return and smallest drawdown occur under different allocations.

Table 3: Full-window portfolio results from 10,000 USDC. Return includes the marked value of open positions. Maximum drawdown is the positive magnitude of the largest decline from a running daily equity peak. Closed and open count market positions; all closed positions in these four runs have positive gross settlement PnL.

Configuration	Return (%)	Max. DD (%)	Equity (USDC)	Closed	Open
A	200.76	12.72	30,076	105	1
B	257.25	13.64	35,725	103	1
C	244.18	11.95	34,418	105	1
D	318.32	19.02	41,832	91	0

Figure 2 shows the complete time paths. Each configuration closes between 91 and 105 positions, all with positive settlement PnL in this sample. Nevertheless, drawdowns range from 11.95% to 19.02%. Open positions can lose marked value before a favorable settlement, and larger allocations increase the portfolio effect of these movements. A–C each retain one open position at the end; D has none. The reported returns do not treat unresolved inventory as a winning settlement.

5.2 CAPITAL DEPLOYMENT AND RISK

The comparison between A and B separates faster allocation from a different expert signal. Their activation records are identical. The 104 markets filled by both have the same direction, signal time, fill time, and fill price. B does not add new markets; it omits two small positions held by A. It assigns more notional to 81 shared positions and less to 23. Total purchased notional over the shared positions rises from 232,310 to 328,144 USDC through cash recycling, not borrowing. The result is a 56.49-percentage-point increase in return and a 0.92-point increase in maximum drawdown.

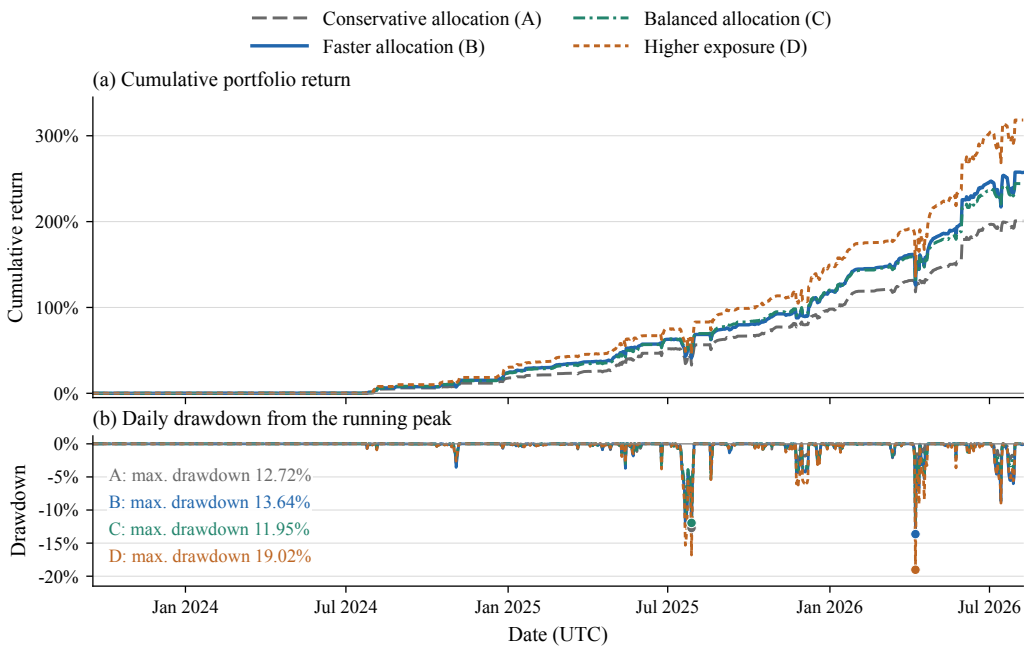


Figure 2: Full-window cumulative return and daily drawdown for A–D. Both panels use marked-to-market equity, including unresolved positions, from the same 10,000-USDC starting capital. Drawdown is measured relative to the running equity peak, as in Equation (9); Table 3 reports its largest depth. All curves use delayed, quantity-unconstrained price replay.

This comparison illustrates the role of allocation speed. A position cap is only an upper bound: a conservative deployment fraction may leave much of the available cash unused even when the same signals qualify. Conversely, a larger cap alone need not produce the largest drawdown. C's slower deployment gives it a smaller drawdown than B despite its 35% cap. D changes both the cap and the price ceiling relative to B, so their difference should not be attributed to either change in isolation. Appendix C reports calendar results, allocation diagnostics, and further settings.

5.3 Locality and parameter sensitivity

Table 4 reports all matched variations around B. Each variation changes one parameter or weighting rule while retaining the same activation records, data window, and initial capital. Wallet caps limit an address's share of selected-direction weight; history support is the weighted mean effective number of related settled markets.

The locality kernel changes which settled histories support a wallet's score, not the text encoder itself. Holding B's remaining settings fixed, replacing the thresholded quadratic kernel with non-negative cosine similarity lowers return from 257.25% to 173.15% and raises maximum drawdown from 13.64% to 26.28%. The corresponding comparison under A gives 200.76% versus 110.17% return and 12.72% versus 26.79% drawdown. These comparisons evaluate how local and diffuse semantic weighting affect performance within the complete strategy.

The matched checks around B are not uniformly favorable. Varying the maximum directional wallet share from 65% to 85% gives returns of 221.39% to 319.68%, with maximum drawdown between 13.64% and 13.65%. The permissive 85% setting therefore outperforms B in this sample. Lowering the mean related-history requirement to 1.0, however, raises drawdown to 27.59%. The locality threshold is also consequential: values of 0.45 and 0.55 yield 71.90% and 63.94% returns, with drawdowns of 30.39% and 29.53%. The favorable result at 0.50 should therefore not be read as insensitivity to the shape of semantic weighting.

Table 4: Matched sensitivity around B on the full observation window. Each variation changes one parameter or weighting rule; all other settings are fixed. Returns include marked inventory, and maximum drawdown uses daily equity.

Variation from B	Return (%)	Max. DD (%)
B (unchanged)	257.25	13.64
Selected-direction wallet cap 65%	221.39	13.64
Selected-direction wallet cap 85%	319.68	13.65
Mean effective expert history ≥ 1.0	196.57	27.59
Mean effective expert history ≥ 2.0	196.66	13.62
General position cap 15%	193.05	8.20
General position cap 35%	302.34	19.07
Kernel threshold 0.45	71.90	30.39
Kernel threshold 0.55	63.94	29.53
Order delay 30 minutes	202.69	13.83
Buy fee 1% of notional	201.95	13.89
Adverse price adjustment 25 bps	252.15	13.63
Remove additional competitive-event cap	278.53	13.64
Non-negative cosine weights	173.15	26.28

Exposure changes produce a clearer return–drawdown trade-off: a 15% general position cap gives 193.05% return and 8.20% drawdown, while a 35% cap gives 302.34% and 19.07%. Removing the additional competitive-event cap raises return to 278.53% but leaves maximum drawdown at 13.64%. That cap therefore does not lower the observed maximum drawdown in this comparison.

The execution stresses retain returns above 200%: a 30-minute delay returns 202.69%, a 1% buy fee returns 201.95%, and 25-basis-point slippage returns 252.15%, with drawdowns between 13.63% and 13.89%. These runs recompute the portfolio, so costs and delay can change fills and later capital availability. They test the stated price-replay assumptions, not order-book capacity; the fee is a hypothetical stress rather than a historical fee estimate.

Appendix B gives the corresponding position counts and equity paths. Appendix C reports further allocation and activation settings.

6 Related Literature

6.1 Prediction markets, trader skill, and imitation

Prediction markets can aggregate information even in relatively thin settings (Wolfers & Zitzewitz, 2004; Cowgill & Zitzewitz, 2015). Their forecasts are not uniformly efficient, however. Calibration depends on time to expiration (Page & Clemen, 2013), and favorite-long-shot patterns can arise from belief distortions or risk preferences (Snowberg & Wolfers, 2010). The possibility of identifying contracts likely to fail motivates risk-conditioned use of prediction-market signals (Tai et al., 2019).

The expert-aggregation literature provides the conceptual basis for selecting participants from historical performance. Budescu & Chen (2015) identify persistent positive contributors and show that excluding poor performers can improve aggregate forecasts. Atanasov et al. (2017) find gains from recency, past performance, and recalibration. Social-trading studies likewise examine whether followers can benefit from observed leaders (Lee & Ma, 2018; Dorfleitner et al., 2018; Apesteguia et al., 2020), while also documenting imitation-related risk.

Recent work studies these questions directly on decentralized prediction markets. Oliphant (2026) estimates returns to observed copying on Polymarket, and Wenth (2026) reports negative evidence from smart-money replication and multi-wallet convergence in weather markets. Other work examines persistent skill, informed minorities, and domain-specific trader classifications (Gómez-Cram et al., 2026; Yang, 2026b; Luong & Heesen, 2026). Profits can be highly concentrated and only modestly persistent (Akey et al., 2026), while large traders may have information advantages in particular domains (Cheong & Tamayo, 2026). These findings motivate both contextual skill measurement and controls against concentrated apparent consensus.

Two recent studies sharpen, but do not duplicate, this question. Diaz & You (2026) find that a small set of skilled wallets helps incorporate public information in earnings contracts; their object is market-level price discovery rather than delayed replication of wallet flow. Park (2026) shows that a public trade endpoint can give a coherent but role-incomplete wallet record. That identification warning is directly relevant to our explicitly taker-only sample and reinforces the separation between observed trading skill and a complete account of wallet wealth.

6.2 Semantic representation and execution

Sentence-embedding models make it possible to compare question meaning through fixed-dimensional vectors and cosine similarity (Reimers & Gurevych, 2019). Capponi et al. (2025) use natural-language relations among Polymarket contracts to identify leader-follower opportunities after one contract resolves. Our use of semantics is different: related past questions determine which part of a wallet's settled history is relevant to a current market, and current question text also determines an exposure class.

Directional information is not equivalent to implementable profit. Della Vedova (2026) emphasizes the role of execution, and Yang (2026a) shows that late entry can inflate apparent forecasting skill. Public-feed directions may also differ from on-chain reconstruction (Dubach, 2026). Our design therefore specifies a historical label schedule, a restricted entry horizon, and delayed price replay. We distinguish the portfolio effects of expert selection and allocation from order-book capacity, which the price-replay experiment does not identify.

7 Limitations

The candidate inventory was assembled retrospectively and may omit delisted or otherwise unavailable markets. The replay's turnover threshold determines admission within this inventory; the metadata-volume collection screen does not establish complete historical coverage. Question wording, scheduled closes, and label times use archived metadata rather than contemporaneous observations. The one-second label offset is an event-ordering convention, not a measurement of information latency or on-chain finality. Some admitted trades may occur after the real-world event is known. Wallet scores therefore describe the settlement value of observed activity, with incomplete coverage of account positions and information timing.

The price replay does not reconstruct order-book depth, spreads, queue priority, or market impact. Complementary-token prices and unrestricted fill quantities may not be attainable, especially for large requests. The reported returns therefore measure the strategy under a specified historical-price convention, not realized or capacity-certified profits (Della Vedova, 2026; Yang, 2026a; Dubach, 2026; Park, 2026).

Thresholds and configurations were selected on the same archive, so local sensitivity checks are not out-of-sample validation. The local-versus-cosine comparison does not isolate the gain over global wallet scoring or passive favorite buying. Related contracts may also share risks, and distinct wallet addresses may share an owner or coordinate their trades. Address concentration is therefore an imperfect measure of independent support. The keyword-based event classifier has not been validated against labeled event categories.

8 Conclusion

This paper combines observable market activation, local semantic weighting of settled wallet histories, current-flow consensus, and exposure limits in a rule-based prediction-market strategy. In the historical replay, the amount and timing of deployed capital materially change returns and drawdowns even when traded opportunities largely overlap. Locality and historical-support thresholds also affect the selected signals and portfolio paths. These findings show how contextual wallet selection and capital allocation jointly shape the behavior of an expert-following portfolio.

Reproducibility Statement

Appendix D specifies the data conventions, fixed configurations, code entry points, and numerical checks needed to reproduce the analysis. The anonymous supplement includes executable tests, a synthetic demo, and saved accounts for all 24 distinct empirical configurations. These support implementation and accounting checks without market-data downloads. Exact strategy replay additionally requires the full archived inputs, which are not included in the supplement.

AI use statement

Cursor and OpenAI Codex assisted with implementation and data processing, experimental design and execution, parameter exploration, and interpretation of results. They also supported literature search, figure preparation, and manuscript drafting and editing. The authors reviewed and revised the AI-assisted materials through code tests, numerical checks, and source verification, and take responsibility for the final manuscript, code, and analyses.

REFERENCES

Pat Akey, Vincent Grégoire, Nicolas Harvie, and Charles Martineau. Who wins and who loses in prediction markets? evidence from polymarket. SSRN Working Paper 6443103, Social Science Research Network, March 2026. URL https://ssrn.com/abstract=6443103.

Jose Apesteguia, Jörg Oechssler, and Simon Weidenholzer. Copy trading. *Management Science*, 66 (12):5608–5622, 2020. doi: 10.1287/mnsc.2019.3508. URL https://doi.org/10.1287/mnsc.2019.3508.

Pavel Atanasov, Phillip Rescober, Eric Stone, Samuel A. Swift, Emile Servan-Schreiber, Philip Tetlock, Lyle Ungar, and Barbara Mellers. Distilling the wisdom of crowds: Prediction markets vs. prediction polls. *Management Science*, 63(3):691–706, 2017. doi: 10.1287/mnsc.2015.2374. URL https://doi.org/10.1287/mnsc.2015.2374.

David V. Budescu and Eva Chen. Identifying expertise to extract the wisdom of crowds. *Management Science*, 61(2):267–280, 2015. doi: 10.1287/mnsc.2014.1909. URL https://doi.org/10.1287/mnsc.2014.1909.

Agostino Capponi, Alfio Gliozzo, and Brian Zhu. Agentic AI for clustering, relationship discovery, and semantic trading in prediction markets, 2025. URL https://arxiv.org/abs/2512.02436.

Wan Chu Cheong and Ane Miren Tamayo. Beyond the wisdom of the crowd: Concentrated informed trading in earnings prediction markets. SSRN Working Paper 6685139, Social Science Research Network, April 2026. URL https://ssrn.com/abstract=6685139.

Bo Cowgill and Eric Zitzewitz. Corporate prediction markets: Evidence from Google, Ford, and firm X. *The Review of Economic Studies*, 82(4):1309–1341, 2015. doi: 10.1093/restud/rdv014. URL https://doi.org/10.1093/restud/rdv014.

Joshua Della Vedova. Who profits from prediction markets? execution, not information. SSRN Working Paper 6191618, Social Science Research Network, February 2026. URL https://ssrn.com/abstract=6191618.

Zachary Diaz and Leyuan You. Skill driven information discovery: Evidence from earnings prediction markets. SSRN Working Paper 7279999, Social Science Research Network, August 2026. URL https://ssrn.com/abstract=7279999.

Gregor Dorfleitner, Lukas Fischer, Carina Lung, Philipp Willmertinger, Nico Stang, and Natalie Dietrich. To follow or not to follow – an empirical analysis of the returns of actors on social trading platforms. *The Quarterly Review of Economics and Finance*, 70:160–171, 2018. doi: 10.1016/j.qref.2018.04.009. URL https://doi.org/10.1016/j.qref.2018.04.009.

Philipp D. Dubach. The anatomy of a decentralized prediction market: Microstructure evidence from the polymarket order book, 2026. URL https://arxiv.org/abs/2604.24366.

Roberto Gómez-Cram, Yunhan Guo, Theis Ingerslev Jensen, and Howard Kung. Prediction market accuracy: Crowd wisdom or informed minority? SSRN Working Paper 6617059, Social Science Research Network, April 2026. URL https://ssrn.com/abstract=6617059.

Woonyeol Lee and Qiang Ma. Discovering expert traders on social trading services. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 22(2):224–235, 2018. doi: 10.20965/jaciii.2018.p0224. URL https://doi.org/10.20965/jaciii.2018.p0224.

Kim Long Luong and Gloria Heesen. The wisdom of the few: Skilled traders and prediction market accuracy. SSRN Working Paper 6758662, Social Science Research Network, March 2026. URL https://ssrn.com/abstract=6758662.

Patrick Oliphant. The return to imitation: Evidence from copy trading on polymarket. SSRN Working Paper 6670318, Social Science Research Network, April 2026. URL https://ssrn.com/abstract=6670318.

Lionel Page and Robert T. Clemen. Do prediction markets produce well-calibrated probability forecasts? *The Economic Journal*, 123(568):491–513, 2013. doi: 10.1111/j.1468-0297.2012.02561.x. URL https://doi.org/10.1111/j.1468-0297.2012.02561.x.

Chris Park. A price tape is not an audit trail: Identification limits in public trading records. SSRN Working Paper 7240938, Social Science Research Network, August 2026. URL https://ssrn.com/abstract=7240938.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Erik Snowberg and Justin Wolfers. Explaining the favorite–long shot bias: Is it risk-love or misperceptions? *Journal of Political Economy*, 118(4):723–746, 2010. doi: 10.1086/655844. URL https://doi.org/10.1086/655844.

Chung-Ching Tai, Hung-Wen Lin, Bin-Tzong Chie, and Chen-Yuan Tung. Predicting the failures of prediction markets: A procedure of decision making using classification models. *International Journal of Forecasting*, 35(1):297–312, 2019. doi: 10.1016/j.ijforecast.2018.04.003. URL https://doi.org/10.1016/j.ijforecast.2018.04.003.

Andreas Wenth. Four strategies, 562 trades, zero edge: A forensic autopsy of algorithmic weather betting. SSRN Working Paper 6496661, Social Science Research Network, March 2026. URL https://ssrn.com/abstract=6496661.

Justin Wolfers and Eric Zitzewitz. Prediction markets. *Journal of Economic Perspectives*, 18(2):107–126, 2004. doi: 10.1257/0895330041371321. URL https://doi.org/10.1257/0895330041371321.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. C-Pack: Packaged resources to advance general chinese embedding, 2023. URL https://arxiv.org/abs/2309.07597.

Hsiang-Chieh (Alex) Yang. Measuring trader skill in prediction markets: How late entry inflates forecasting accuracy. SSRN Working Paper 7065625, Social Science Research Network, July 2026a. URL https://ssrn.com/abstract=7065625.

Hsiang-Chieh (Alex) Yang. Skilled liquidity provision in prediction markets: Evidence from 150 million trades. SSRN Working Paper 6396698, Social Science Research Network, March 2026b. URL https://ssrn.com/abstract=6396698.

A Implementation and Evaluation Details

A.1 Candidate inventory and metadata provenance

The initial 10,000 markets were selected by descending volume in an April 24 metadata snapshot, without an outcome filter. The expanded inventory combines these with archived candidates and a Gamma keyset collection of both open and closed markets meeting a 250,000-USDC metadata-volume screen. The latter queries use `limit=100`, ascending market ID, and separate `closed=true` and `closed=false` streams; collection follows the returned cursors rather than taking a fixed number of top-volume pages. Markets created after the observation window are excluded. The final binary inventory in Table 1 excludes one unsupported multi-outcome market.

The collection screen uses retrospectively observed metadata volume. Strategy admission instead uses cumulative actual-token turnover from the archived public prints. The two thresholds therefore apply to different measurements and stages of sample construction.

For the label clock, the explicit date fields are `resolvedAt`, `umaEndDate`, `closedTime`, and `resolveDate`. The latest valid timestamp is r_m; scheduled `endDate` is not a resolution timestamp. Usable labels additionally require `closed=true` and `umaResolutionStatus=resolved`. Ordered outcome prices of at least 0.99 and at most 0.01 map to the winning and losing payouts; two prices equal to 0.5 map to equal payouts. Invalid or ambiguous records are excluded from expert history. The replay releases each usable label at $r_m + 1$ second, after processing the complete resolution-second group.

A.2 Observation clocks and account state

All timestamp groups used for historical volume admission are complete before the admission time is advanced by one second. Hourly activation bars exclude prints at their ending boundary. At equal timestamps, label-release events precede new trades. Market ordering is deterministic, using condition identifiers rather than final-volume rank. Wallet skill is recomputed as new historical labels become available; accumulated observations retain the weights assigned when they arrived. A wallet's trades in both directions remain in their respective directional sums.

The portfolio records each position's acquisition cost separately from its marked value. Acquisition cost enters the deployment target, while marked value determines equity and drawdown. Pending requests reduce the remaining deployment target but do not constitute filled inventory or earmarked cash. At execution, available cash bounds the purchase; fills debit notional and any specified fee. A resolution cancels pending requests in that market and settles held tokens at the payout assigned by the label convention. No new position can open after that event. The ending account includes unresolved inventory at its last observed price.

A.3 Question-text exposure classifier

The classifier strips leading and trailing whitespace and performs case-insensitive, word-boundary matching. A question is a competitive event if either (i) it matches `\bvs\.?\b` and does not end in a question mark, or (ii) it contains an action term and either a competition or league term from the lists below. All remaining questions are standard events.

Group	Matched terms
Action	win, winner, champion, qualify, make, reach
Competition	final, finals, championship, cup, tournament, super bowl, world series, playoff, playoffs
League	nba, wnba, nfl, nhl, mlb, mls, ncaa, uefa, fifa, ufc, pga, atp, wta

Competitive events receive the 15% submission-time equity cap in Equation (8); all other events retain the general cap.

B MATCHED HYPERPARAMETER SENSITIVITY

Table 5 holds B's source data, activation rule, historical records, and allocation settings fixed except for the named change. Wallet share means the largest address's fraction of selected-direction flow, not its fraction of all market volume. History support is the selected-direction weighted mean of effective settled-market counts. The cosine row removes the thresholded quadratic kernel but retains the same embedding vectors. The no-event-cap row retains the 25% general cap. Fee and slippage stresses are hypothetical flat costs, not estimates of historical exchange fees.

Table 5: Matched sensitivity around B. Each variation changes only the named parameter or similarity-weighting rule. Return includes marked open positions, maximum drawdown is a positive daily peak-to-trough magnitude, and losses count negative gross settlement PnL.

Variation from B	Return (%)	Max. DD (%)	Closed/open	Losses
B (unchanged)	257.25	13.64	103/1	0
Selected-direction wallet cap 65%	221.39	13.64	93/1	0
Selected-direction wallet cap 85%	319.68	13.65	128/1	0
Mean effective expert history ≥ 1.0	196.57	27.59	119/1	1
Mean effective expert history ≥ 2.0	196.66	13.62	92/1	0
General position cap 15%	193.05	8.20	105/1	0
General position cap 35%	302.34	19.07	103/1	0
Kernel threshold 0.45	71.90	30.39	115/1	2
Kernel threshold 0.55	63.94	29.53	97/1	3
Order delay 30 minutes	202.69	13.83	98/1	0
Buy fee 1% of notional	201.95	13.89	103/1	0
Adverse price adjustment 25 bps	252.15	13.63	99/1	0
Remove additional competitive-event cap	278.53	13.64	103/1	0
Non-negative cosine weights	173.15	26.28	110/0	3

Costs and delay are evaluated by rerunning the portfolio. They may change which orders fill and which later signals can be funded, rather than simply subtracting a fixed cost from a frozen trade list. Closed-position counts exclude unresolved positions. Return and drawdown include fees and marked inventory; any count of losing settlements refers to gross position PnL before separately debited fees.

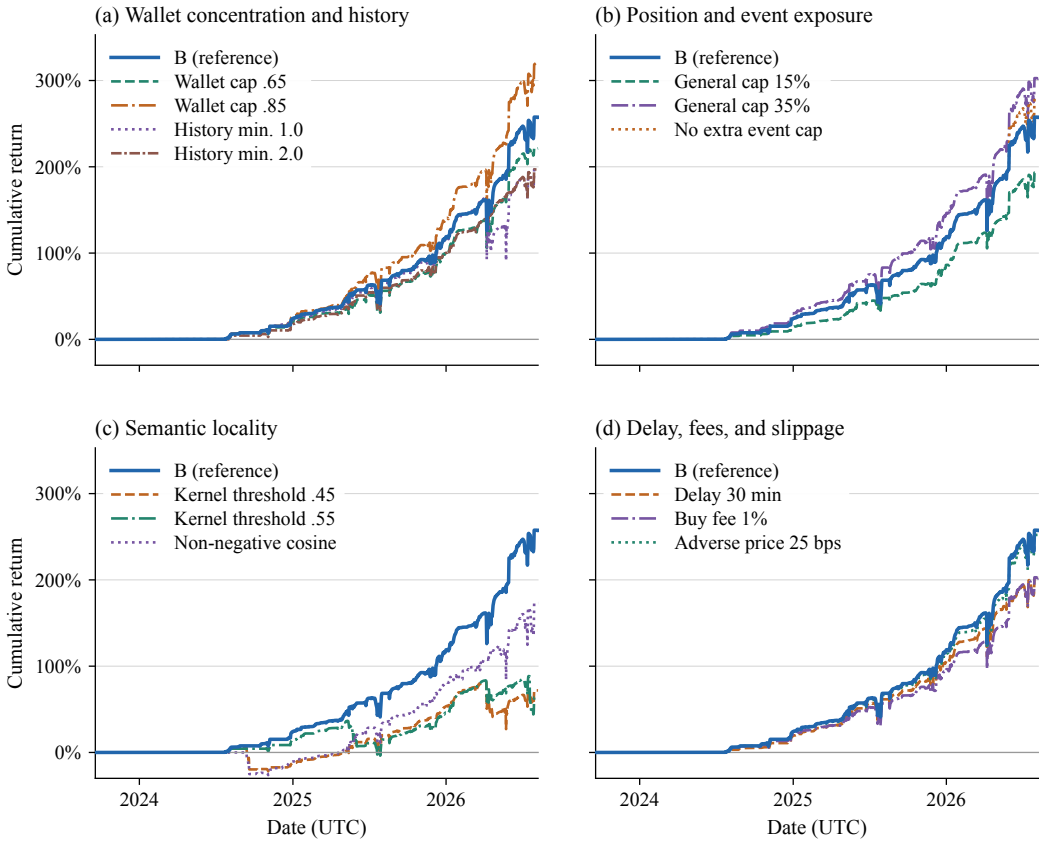


Figure 3: Full-window cumulative returns under matched changes around B. Each panel includes B and groups related parameters; settings are those in Table 5. All paths share the same initial capital and price-replay convention.

Figure 3 separates the effects of breadth, exposure, semantic locality, and execution assumptions on the full account path. A 30-minute delay returns 202.69% with a 13.83% drawdown; a 1% fee returns 201.95% with a 13.89% drawdown; and 25-basis-point slippage returns 252.15% with a 13.63% drawdown. Removing the competitive-event cap, while retaining the 25% general cap, raises return to 278.53% without materially changing maximum drawdown. The event cap therefore does not reduce the realized maximum drawdown in this particular comparison.

The locality panel shows a different pattern from the exposure panel. Moving the kernel threshold to either 0.45 or 0.55 lowers return and raises drawdown relative to B. These path changes are not simply scaled versions of the primary result: changing expert weights changes which signals qualify.

C Additional Experimental Results

C.1 Calendar performance and profit concentration

Tables 6 and 7 divide the same continuous accounts into calendar periods. Capital and expert history carry forward across years. No position opens in the partial 2023 period. B returns 24.21% in 2024, 76.36% in 2025, and 63.09% in 2026 through August 8. These linked returns compound to the full-window result; they are not independent replications. The drawdown calculation within each calendar period restarts its running peak at the previous period's ending equity.

Table 6: Calendar-period portfolio returns. Each period starts from the previous period's ending equity, not a fresh 10,000-USDC account. The first and last years are partial periods; open positions are marked to market.

Calendar period	A (%)	B (%)	C (%)	D (%)
2023 (from Sep. 18)	0.00	0.00	0.00	0.00
2024	17.63	24.21	23.58	29.99
2025	68.39	76.36	78.33	91.91
2026 (to Aug. 8)	51.84	63.09	56.17	67.69

Table 7: Within-period maximum daily drawdown. The running peak is reset to the period's starting equity; these numbers are not the full-window underwater depths observed during each year. Values are positive drawdown magnitudes.

Calendar period	A (%)	B (%)	C (%)	D (%)
2023 (from Sep. 18)	0.00	0.00	0.00	0.00
2024	2.91	3.57	2.23	2.62
2025	12.72	13.62	11.95	16.80
2026 (to Aug. 8)	5.82	13.64	11.59	19.02

Table 8 measures how much positive settlement PnL comes from the largest contributors. For B, the largest position supplies 11.02% of gross settlement profits and the ten largest supply 35.16%. All four configurations have about 35–39 effective profitable positions by the inverse-Herfindahl measure in the caption. This is an attribution of the realized path, not a counterfactual in which a large winning trade is removed: removing a trade would also change later available capital and position sizes.

Table 8: Concentration of positive gross settlement PnL. Top-k shares divide the k largest positive closed-position PnLs by total positive closed-position PnL; open positions are excluded. Effective positions equal $1/\sum_i w_i^2$, where w_i is a profitable position's share of that total. This is a descriptive attribution, not a delete-one counterfactual.

Configuration	Top 1 (%)	Top 5 (%)	Top 10 (%)	Effective positions
A	12.13	25.16	36.61	35.3
B	11.02	23.42	35.16	39.2
C	11.29	23.88	36.22	37.3
D	10.13	26.40	38.83	36.0

C.2 Allocation and print-size diagnostics

Table 9 classifies filled orders by the allocation fraction in Equation (7). For B, 66 of 104 orders reach the upper bound, 32 lie strictly between the bounds, and six reach the lower bound. The fraction determines allocation from remaining deployment capacity G, before the cash and position caps are applied. We reconstruct it from the saved signal price before slippage and the time from submission to the archived scheduled close.

Table 9: Clipping of the allocation fraction for filled orders in A–D. Lower and upper counts include equality with the corresponding bound; interior counts are strictly between them. These are fractions of remaining deployment capacity before the cash and position caps, not position sizes.

Configuration	Filled	Lower	Interior	Upper	Upper (%)
A	106	15	58	33	31.13
B	104	6	32	66	63.46
C	106	8	37	61	57.55
D	91	4	23	64	70.33

In B, 94 of 104 simulated purchases are larger in token quantity than the individual print used to assign their execution price. The median purchase-to-print quantity ratio is 40.23, and the largest purchase notional is 8,940 USDC. The denominator is the quantity of the single print supplying the price, including complementary-outcome prints. This descriptive ratio does not measure order-book depth or executable participation.

C.3 Alternative allocation and activation settings

Table 10 reports a broader set of completed configurations, listing changes relative to B. Within this table, the price-ceiling, locality-power, activation-conviction, and consensus comparisons share E2 as their reference: faster allocation with a 35% general cap and a 0.98 price ceiling. The table also includes the main allocation alternatives. Rows that change several parameters relative to B are not one-factor ablations. This is a record of retrospective comparisons, not a separately estimated efficient frontier.

Table 10: All ten exploratory configurations, with changes listed relative to B; multi-change rows are not one-factor comparisons. Return includes marked open positions, maximum drawdown is a positive daily peak-to-trough magnitude, and losses count negative gross settlement PnL.

Variation from B	Return (%)	Max. DD (%)	Closed/open	Losses
E1: B (unchanged)	257.25	13.64	103/1	0
E2: General cap 35%	302.34	19.07	103/1	0
E3: General cap 35%; $\lambda = .50$; $f_{\max} = .35$ (C)	244.18	11.95	105/1	0
E4: General cap 35%; price ceiling .97 (D)	318.32	19.02	91/0	0
E5: General cap 35%; price ceiling .99	251.07	19.17	129/1	1
E6: General cap 35%; kernel power 1.5	229.70	16.80	99/1	0
E7: General cap 35%; kernel power 2.5	284.18	16.80	101/1	0
E8: General cap 35%; activation conviction .875	150.53	43.91	116/1	3
E9: General cap 35%; activation conviction .925	71.45	35.74	97/1	1
E10: General cap 35%; consensus .80	277.25	19.07	101/1	0

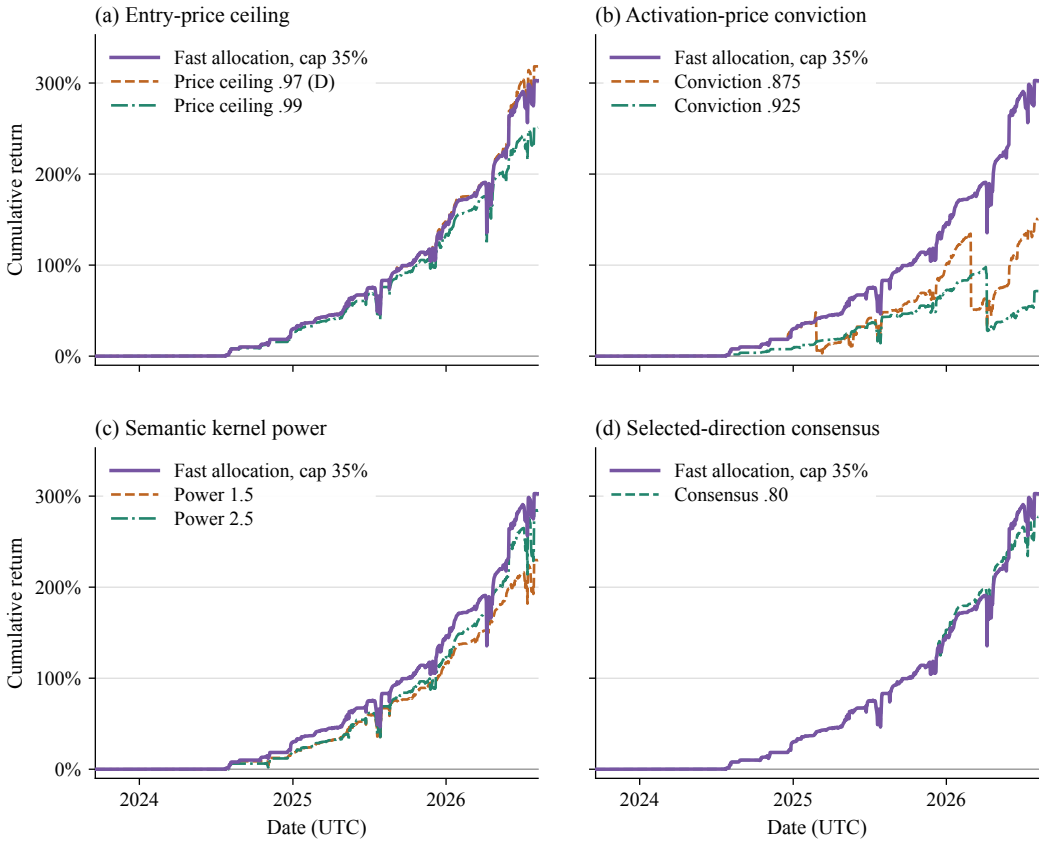


Figure 4: Cumulative returns for additional settings around the faster 35%-cap reference. Curves vary one named setting within each panel. These are the configurations in Table 10, not the matched B-centered checks in Table 5.

Figure 4 shows that activation changes can have a larger effect than allocation changes. Relative to the 0.90 activation-conviction threshold, 0.875 returns 150.53% with a 43.91% maximum drawdown, while 0.925 returns 71.45% with a 35.74% drawdown. A stricter price-conviction rule is therefore not automatically safer: it changes when flow accumulation starts and which opportunities become available. In the same 35%-cap family, price ceilings of 0.97, 0.98, and 0.99 give returns of 318.32%, 302.34%, and 251.07%, with drawdowns near 19%. Strong main-configuration results therefore do not imply stability under every nearby rule.

D REPRODUCIBILITY DETAILS

The anonymous supplement contains the source code, fixed configurations, unit tests, a synthetic demonstration, and saved empirical accounts for all 24 distinct configurations. The tests and demonstration run without market-data downloads. Saved fills, positions, activation records, and daily equity support accounting checks and regeneration of the original performance tables and figures. Full strategy replay additionally requires archived candidate metadata, fixed BGE vectors, normalized trade arrays, and wallet-history sidecars. These inputs are not included in the supplement, and contemporary API queries may not recover the same historical snapshot.

The portable input exporter is `src/paper_v2_inputs.py`, and the replay entry point is `src/paper_v2_replay.py`. Export verifies the archived source and prepared-history bindings before creating a relative-path input manifest. Replay takes explicit prepared-input, fully resolved specification, and output paths. The reference engine is the default. The optional `fast-checked` mode runs both reference and accelerated engines and requires identical portfolio records before reporting successful validation. SHA-256 manifests bind market identities, observation boundaries, metadata, vectors, normalized arrays, and implementation files. Each run records its configuration and completion status. Initial capital is 10,000 USDC in every specification.

The public inputs originate from Polymarket's Gamma and Data APIs; transaction queries use the declared window and taker-only flag. The question encoder is `BAAI/bge-large-en-v1.5`, revision `d4aa6901d3a41ba39fb536a557fa166f842b0e09`. Stored vectors avoid requiring model downloads for replay. Histories retain all eligible activity after volume admission and before the resolution boundary.

The accompanying reproduction notes map each table and figure to its result directories and fixed specifications, distinguish saved-result regeneration from full replay, and record the required external inputs. Accounting verification checks total equity against initial capital plus settled PnL, less fees, plus unrealized PnL, and reconstructs daily cash and inventory from fills and settlements. The allocation diagnostics additionally use archived scheduled close dates to reconstruct the submission-time allocation fraction.

From the unpacked anonymous supplement root, install `requirements-core.txt` in a Python environment and run the following checks. The demo creates synthetic inputs; its output is not an empirical replication.

```
PYTHONPATH=src python -m pytest -q
python verify_archive.py
python check_reference_results.py
PYTHONPATH=src python -m paper_v2_inputs demo \
  --output demo_inputs
PYTHONPATH=src python -m paper_v2_replay \
  --inputs demo_inputs/inputs --spec demo_inputs/spec.json \
  --output demo_run --engine fast-checked
```

For empirical replay, use the separately prepared full input directory and the matching fixed specification as described in `REPRODUCING.md`. The reference-result checker verifies saved-account arithmetic; the replay entry point reruns expert selection and the portfolio on the supplied event stream.